AI Research Labs

246 questions · 49 lessons

Question *bank*

Every quiz question in the course, written as exam-style scenarios with misconception-based options. Answers and explanations are at the back.

1 Machine Learning Foundations (25)

2 Statistical Learning and Model Evaluation (41)

3 Linear Regression (45)

4 Logistic Regression and Classification (55)

5 Regularisation and Model Selection (45)

6 Clustering and Support Vector Machines (35)



Free course with quizzes, progress tracking and Scoopy, an AI tutor:
airesearchlabs.si/learn/ml

Machine Learning Foundations · Edition November 2026 · © 2026 AI Research Labs. Free for personal study.

How to use this test. Answer all 246 questions without notes, then mark yourself with the answer key at the back. Each question has exactly one best answer. Score 75% or more and you've mastered the material; below that, reread the lessons for the questions you missed.

1.1 What is Machine Learning?

Q1 `apply` An online shop wants to flag fraudulent orders. Fraud patterns change every few weeks and show up as combinations of hundreds of weak signals. It has two years of orders labelled fraud / not fraud. Why is machine learning a better fit than hand-written rules?

A Machine learning guarantees no false alarms once the training set is large enough

B Machine learning doesn't need examples of past fraud, so it adapts to new patterns automatically

C The rule is too complex and changes too fast to maintain by hand, and labelled examples exist to learn it from

D Hand-written rules can only use a handful of input variables, so they can never see all of the weak signals at once

Q2 `concept` A system stores every training example and, for a new input, returns the answer of the identical stored input. Why is this not useful machine learning?

A Looking up stored answers is slower than evaluating a trained model

B It has no answer for inputs it hasn't seen, so it doesn't generalise

C It only fails when the training data contain outliers

D Memorising is fine as long as the training set is large enough

Q3 `concept` In $\hat{y} = f(X)$, which description is correct?

A X are the inputs, f is the rule learned from data, $\hat{y}$ is the predicted target

B $\hat{y}$ is the true observed value, and f is a rule that the programmer writes by hand in advance

C X is the target, f is the dataset, $\hat{y}$ is the true observed value

D f is the loss function and $\hat{y}$ is the average training error

Q4 `apply` Two models predict which patients will be readmitted. Model A has higher overall accuracy. Model B catches more of the patients who are readmitted but raises more false alarms. A missed readmission costs far more than an unnecessary follow-up call. Which model is more useful?

A Model B, because the right choice depends on what each kind of mistake costs

B Neither can be judged, because accuracy and cost can't be compared

C Model A, because false alarms always cost more than they appear to

D Model A, because higher overall accuracy always leads to better decisions

Q5 `concept` How do statistics and machine learning relate?

A They share goals; statistics leans toward inference, ML toward prediction at scale, and they are converging

B They are separate fields that happen to use the same datasets

C Machine learning has replaced statistics, so tools like p-values are now obsolete

D Statistics is mainly about prediction at scale, while ML only describes data that has already been collected in the past

1.2 Types of Machine Learning

Q6 apply A streaming service has viewing histories for 2 million users and no labels. It wants to discover groups of users with similar tastes. What kind of task is this?

- A Time-series forecasting of future views
- B Supervised regression on total watch time
- C Supervised classification, with the group as the label
- D Unsupervised clustering

Q7 apply A hospital predicts each patient's length of stay, in whole days, from their admission record. What kind of task is this?

- A Supervised classification, because days are whole numbers
- B Supervised regression
- C Time-series forecasting, because it involves days
- D Unsupervised clustering, because stays vary so much

Q8 concept A loan dataset has columns age, income, postcode and "defaulted (yes/no)". For a default-prediction model, which are the features and which is the label?

- A Features: defaulted. Labels: age, income and postcode
- B Features: age, income, postcode. Label: defaulted
- C All four columns are features; there is no label
- D Label: income. Features: age, postcode and defaulted

Q9 apply You have 200 observations and expect a roughly straight-line relationship between X and Y. Why might linear regression beat K-nearest neighbours here?

- A Non-parametric models always have higher bias than parametric ones, whatever the data
- B A parametric model needs less data, and its assumed shape matches the problem
- C K-nearest neighbours can't be used for predicting numbers
- D Linear regression makes no assumptions about the shape of the data

Q10 diagnose A team codes product colour as red = 1, green = 2, blue = 3 and fits linear regression to predict colour. What's the main problem?

- A Colour has no labels here, so the team should be using clustering instead
- B Nothing, as long as predictions are rounded to the nearest code
- C Colour is a category; the codes add a fake order and spacing, so classify instead
- D Linear regression can't be fitted to a target that takes only three distinct values

1.3 The ML Modelling Process

Q11 concept When should the data be split into training, validation and test sets?

- A After feature selection, so the test set has the same features
- B Before any statistic, such as a mean or a feature ranking, is computed from it
- C After a first model is trained, to see how much data it needs
- D After standardising the full dataset, so that all three sets share exactly the same scale

Q12 concept Which of these is a hyperparameter rather than a parameter fitted from data?

A The intercept of a fitted regression line

B The slope coefficient estimated by least squares

C The centroid positions found by K-means

D The maximum depth allowed for a decision tree

Q13 concept Team A compares logistic regression with an SVM. Team B compares decision trees of depth 3 and depth 15. Which statement is correct?

A A chooses complexity, because SVMs are more complex than logistic regression

B A chooses an algorithm; B chooses complexity within one algorithm

C Both teams are choosing how complex the model is allowed to be

D Both teams are choosing between two different learning algorithms

Q14 diagnose After the final test, error analysis suggests a useful new feature. The team adds it, retrains, evaluates on the same test set again and reports the new, higher score. What's wrong?

A Nothing: the ML process is iterative, so looping back to the test set is expected

B New features can never be added to a model once it has been tested at all

C The test set has steered a development decision, so the new score is optimistic

D The test labels have to be recomputed whenever a new feature is added to the model

Q15 concept Why is the ML workflow described as iterative rather than a one-way pipeline?

A Every deployed model has to be retrained from scratch once a day to stay accurate

B Evaluation and error analysis often send you back to change the data, features or model

C The test set is reused at each loop to measure how much the model has improved

D Models are fitted by optimisers that run through the data for many iterations

1.4 Fundamental Limits of Machine Learning

Q16 concept Which statement correctly describes the No Free Lunch theorem?

A Deep learning is the one family of methods that beats every other on average

B Simpler algorithms always generalise better than complex ones on real problems

C Over all possible problems no algorithm wins; on a real one, the best-suited assumptions win

D Every algorithm reaches the same accuracy on any particular dataset you give it, whatever its structure

Q17 apply A colleague says: "Gradient boosting won on our last three projects, so we'll skip comparing models this time." What does No Free Lunch suggest?

A Pick the simplest model instead, because No Free Lunch favours simple models

B Past wins don't show it suits this problem; compare candidates on this data

C Agreed: three wins in a row show it's the best algorithm for this kind of work

D The choice doesn't matter, since No Free Lunch says all models perform equally

Q18 concept What makes it possible for a model to predict well at inputs it never saw during training?

- A Removing all of the noise from the training labels before fitting
- B A test set that is larger than the training set it is compared with
- C A model large enough to memorise every single training point exactly
- D Assumptions about structure, such as smoothness, plus representative data

Q19 diagnose A skin-lesion classifier trained on images from one clinic's camera performs poorly at a second clinic that uses different cameras. What is the most likely explanation?

- A The model has too few parameters to learn what skin lesions look like
- B The test set used during development was too small to be reliable
- C The new images come from a different distribution than the training images
- D The model is non-parametric, so it can't transfer to new clinics

Q20 concept Breiman's "Rashomon effect" refers to...

- A Models fitted to the same data always agreeing closely on their coefficients
- B The most accurate model also turning out to be the easiest one to interpret
- C Adding more features to a model always improving its predictions on new data
- D Many different models fitting about equally well, so coefficients can be unstable

1.5 Empirical Risk and Population Risk

Q21 calculate Predictions $\hat{y} = (2, 4, 6)$; true values $y = (3, 4, 9)$. What is the empirical risk under squared loss?

- A $10/3 \approx 3.33$
- B 10
- C 1.83
- D $4/3 \approx 1.33$

Q22 calculate Same data: $\hat{y} = (2, 4, 6)$, $y = (3, 4, 9)$. What is the empirical risk under absolute loss?

- A $4/3 \approx 1.33$
- B 4
- C 3
- D $10/3 \approx 3.33$

Q23 concept Why can't population risk $R(f)$ be computed directly?

- A It is undefined when the loss is squared error
- B It requires more computation than is practical
- C It needs the test set, which must stay hidden until the very end of the project
- D It is an expectation over the unknown data-generating distribution

Q24 `interpret` A model has MSE 0.5 on its training data and MSE 4.0 on fresh data from the same source. What does this indicate?

- A Both numbers estimate population risk equally well, so average them
- B Training error understates true risk; the large gap suggests overfitting
- C The loss function is wrong, because the two numbers differ
- D The model underfits, because the error on fresh data is high

Q25 `concept` A model reaches zero training loss. What can you conclude about its population risk?

- A It must be large, since zero training loss always means overfitting
- B It is exactly equal to the noise variance
- C Nothing certain: it could still be large if the model fitted noise
- D It must also be zero, since training data come from the population

2.1 Loss Functions and Risk

Q26 `calculate` A prediction is 8 and the true value is 11. What are the squared loss and the absolute loss?

- A 9 and 3
- B 6 and 3
- C 9 and −3
- D 3 and 9

Q27 `calculate` The true label is $y = 0$ and the model predicts $p = 0.9$ for class 1. What is the cross-entropy loss?

- A $-\ln(0.1) \approx 2.303$
- B $-\ln(0.9) \approx 0.105$
- C $-\ln(0.5) \approx 0.693$
- D 0.9

Q28 `apply` You're predicting house prices. A handful of mansions sell for 20 times the typical price, and you don't want them to dominate the fit. Which loss is the better choice?

- A Absolute loss, because it grows only linearly with the size of an error
- B Squared loss, because it handles large values better than absolute loss
- C Cross-entropy, because it penalises confident mistakes the most heavily
- D 0–1 loss, because it only counts whether each prediction is exactly right

Q29 `interpret` For a case with true label 1, model A predicts $p = 0.8$ and model B predicts $p = 0.2$. How do their cross-entropy losses compare?

- A B's loss (1.609) is about 7 times A's (0.223)
- B B's loss is 4 times A's, since $0.8/0.2 = 4$
- C A's loss is larger, since it is more confident
- D They're equal, because only misclassified cases count

Q30 concept A logistic regression is fitted by minimising cross-entropy, then judged by its expected cost at a chosen threshold. Is that inconsistent?

- A Yes: cost-based evaluation only makes sense for models that were trained with hinge loss
- B No: cross-entropy and expected cost are just two names for the same quantity
- C Yes: a model should always be evaluated with the same loss it was trained on
- D No: the loss must be easy to optimise, while the metric should reflect the real task

2.2 Empirical Risk Minimisation (ERM)

Q31 concept What does empirical risk minimisation (ERM) minimise?

- A The test-set error, compared across every candidate model in turn
- B The number of parameters needed to reach a target training accuracy
- C The expected loss over the whole population, within a chosen family of models
- D The average loss on the training sample, within a chosen family of models

Q32 concept Ordinary least squares is ERM with...

- A squared loss over the family of linear functions
- B absolute loss over the family of linear functions
- C cross-entropy loss over the family of linear functions
- D squared loss over the family of all possible functions

Q33 concept In $\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{R}_n(f)$, what is $\hat{f}$?

- A The function in $\mathcal{F}$ with the fewest parameters
- B The average of all the functions in the family $\mathcal{F}$
- C The lowest value of empirical risk reached by any function in $\mathcal{F}$
- D The function in $\mathcal{F}$ that achieves the lowest empirical risk

Q34 apply With $n = 10$ training points, you run ERM with squared loss over all polynomials of degree up to 9. What do you get?

- A A flat line at the mean of y, because the extra terms cancel each other out
- B The model with the lowest possible population risk among all the polynomials up to degree 9
- C A straight line, because ERM always prefers the simplest member of the family
- D A curve through every training point, with zero training loss but likely poor predictions

Q35 concept How does ridge regression fit into the ERM picture?

- A ERM with cross-entropy loss, applied to a continuous target
- B ERM applied to a smaller random subset of the training data, chosen afresh each time
- C ERM with absolute loss instead of squared loss, and no penalty
- D ERM with squared loss plus a penalty on the size of the coefficients

2.3 Underfitting and Overfitting

Q36 interpret Training error is 2% and validation error is 28%. What's the most likely diagnosis?

- A A good fit: the training error is impressively low
- B Overfitting: it fits quirks of the training data
- C Leakage: validation data has reached the model
- D Underfitting: the validation error is too high

Q37 interpret Training error is 35% and validation error is 36%. What's the diagnosis, and what might help?

- A Leakage; re-split the data and train again
- B Overfitting; add regularisation to shrink the gap between them
- C A good fit, because the two errors are so close
- D Underfitting; try a more flexible model or better features

Q38 concept Why can't training error be used to choose between a degree-1 and a degree-10 polynomial?

- A Training error is too noisy to compare across models
- B A degree-10 polynomial always has a higher training error
- C Training error can't be computed for polynomial models
- D Training error always favours the more flexible model

Q39 diagnose A model overfits. Which of these changes would make it WORSE?

- A Dropping weak, noisy features from the model
- B Collecting more training data
- C Adding a regularisation penalty
- D Adding higher-order polynomial terms

Q40 interpret As model complexity increases from very simple to very flexible, what typically happens?

- A Training error keeps falling; test error falls, then rises
- B Test error keeps falling; training error falls, then rises
- C Training error rises; test error keeps falling
- D Training and test error both keep falling together

2.4 Generalisation and the Generalisation Gap

Q41 concept Which change tends to shrink the generalisation gap?

- A Removing the regularisation penalty
- B Giving the model more free parameters
- C Collecting more representative training data
- D Adding random noise to the training labels

Q42 diagnose Validation accuracy is 96%, but the final test accuracy is 79%. What should the team do?

A Merge the test set into training and report the new training accuracy
B Investigate why, improve using train and validation data, then test on fresh data
C Report 96% as the expected accuracy, since this particular test set may just be unlucky
D Keep tuning the model until the score on this test set reaches 96%

Q43 interpret Training error is 30% and validation error is 31%. The gap is tiny. Is this a good model?

A No: a good model should have a negative gap, with lower validation error
B Yes: a small gap shows that the model generalises well
C Yes, as long as the training set was reasonably large
D Not necessarily: both errors are high, suggesting underfitting

Q44 interpret On a learning curve, as the training-set size grows, what typically happens?

A Training error falls while validation error rises, so the gap widens
B Both errors stay roughly constant once there are more than 100 training examples
C Training and validation error both fall steadily towards zero
D Training error rises a little, validation error falls, and the gap narrows

Q45 diagnose A model validated well at launch, but its accuracy drops steadily over the following months. What is the most likely cause?

A The model's training error was too low at launch time
B The regularisation strength λ was set too high
C The incoming data have drifted away from the training distribution
D The model was trained and tuned with too few cross-validation folds at launch

2.5 The Bias–Variance Trade-off

Q46 concept Which component of expected squared prediction error can no model reduce?

A The squared bias
B The training error
C The irreducible noise σ^2
D The variance of the fitted $\hat{f}$

Q47 concept As model complexity increases, what generally happens to bias and variance?

A Both bias and variance rise
B Bias falls and variance rises
C Bias rises and variance falls
D Both bias and variance fall

Q48 `interpret` A linear model is fitted to 50 different training sets drawn from the same source. The fitted lines are almost identical, but all miss the true curve the same way. How would you describe it?

- A High bias, low variance
- B Low bias, low variance
- C High bias, high variance
- D Low bias, high variance

Q49 `apply` You double the training data for a very flexible model. What do you expect?

- A Variance falls; bias and noise stay roughly the same
- B All three components fall by about half
- C Noise falls; bias and variance stay roughly the same
- D Bias falls; variance and noise stay roughly the same

Q50 `concept` In the bias–variance decomposition, what does "variance" refer to?

- A How spread out the residuals are on the training set after fitting
- B How spread out the target values y are in the data
- C How much the features vary between observations
- D How much the fitted prediction changes across training sets

2.6 Training, Validation and Test Sets

Q51 `apply` You need to choose the polynomial degree for a regression. Which data should you use to make that choice?

- A All of the data combined, to use every observation
- B The validation set, or cross-validation on the training data
- C The training set, since that's where the model is fitted
- D The test set, since it gives the most honest score

Q52 `calculate` You have 1,000 observations and use a 60:20:20 train/validation/test split. How many observations are in the test set?

- A 400
- B 200
- C 100
- D 600

Q53 `diagnose` A team tries 40 model configurations, checks each on the test set, and reports the best test score as its expected performance. What's wrong?

- A Nothing, because none of the models was trained on the test data
- B They tried too few configurations to be confident of finding a genuinely good model
- C Picking the best of 40 on the test set makes the reported score optimistic
- D The test set should be larger than the training set for this to work

Q54 concept Only 2% of transactions are fraudulent. Why use a stratified split?

A To remove duplicate transactions before splitting
B To make the training set as balanced as the test set
C To keep the fraud rate similar in every subset
D To put all fraud cases in the training set

Q55 apply You've chosen a model and its hyperparameters using the validation set. What is a common next step before reporting performance?

A Refit it on training + test, then report the validation score
B Refit the chosen model on training + validation, then test once
C Re-run the selection on the test set to confirm the choice
D Tune the hyperparameters again on the test set to finalise

2.7 Cross-validation

Q56 calculate The validation errors from 5-fold cross-validation are 4, 6, 5, 7 and 3. What is the CV estimate?

A 7
B 4.5
C 5
D 25

Q57 calculate You compare 3 candidate models with 10-fold cross-validation. How many model fits does that take, not counting the final refit?

A 30
B 13
C 10
D 3

Q58 diagnose Why is ordinary random K-fold CV inappropriate for forecasting next month's demand from monthly data?

A K-fold CV needs at least 1,000 observations to produce stable estimates
B Monthly demand is a count, and K-fold CV only works for continuous, unbounded targets
C Random folds let the model train on later months and validate on earlier ones
D Each month would end up in the validation fold more than once overall

Q59 interpret Model A has mean CV RMSE 2.22 with fold errors between 2.1 and 2.4. Model B has mean 2.20 with fold errors between 1.5 and 3.2. Which conclusion is most reasonable?

A B is clearly better, because it has the lower mean CV error
B They're similar on average, and A's stability may make it preferable
C The CV results are invalid, because the fold errors differ
D A is overfitting, because its fold errors are so close together

Q60 diagnose A team standardises features using the mean and SD of the whole training set, then runs 5-fold CV on it. Is there a leak?

A No: only feature selection, never scaling, can leak inside cross-validation
B Yes: held-out folds shaped the scaling; standardise inside each fold
C Yes: they should have standardised using the test set instead
D No: standardising doesn't learn anything from the data

2.8 Data Leakage and Good ML Practices

Q61 diagnose A team fills missing blood-test values with the median of the whole dataset, then splits it 80/20. Which leak is this?

A Target leakage
B Group (related-record) leakage
C Preprocessing before splitting
D Time leakage

Q62 diagnose To predict whether a customer will cancel next month, a model uses "number of calls to the cancellation hotline this month". Which leak is this?

A Test-set reuse for selection
B Target leakage
C Feature selection before CV
D Preprocessing before splitting

Q63 calculate Training statistics for house size are $\mu = 1{,}400$ and $\sigma = 400$ sq ft. What is the standardised value z of a 1,800 sq ft house in the test set?

A $+0.8$
B -0.4
C $+1.0$
D $+0.6$

Q64 diagnose A team standardises using training data only, picks the best of 12 models by 5-fold CV on the training set, and evaluates the winner once on the test set. Verdict?

A Leak: 12 models is too many to compare
B Leak: CV and testing used the same model
C No leak: this is the correct workflow
D Leak: preprocessing came before CV

Q65 diagnose A dermatology dataset has three photos per patient. The photos are split at random into training and test sets. What's the fix?

A Use more CV folds so that each photo is tested exactly once
B Split by patient, so each patient's photos stay together
C Remove two of each patient's photos before training
D Standardise each photo before splitting at random

Q66 diagnose From 20,000 candidate genes, a team keeps the 50 most correlated with the label using all rows, then reports 5-fold CV accuracy. What's wrong?

A Validation labels helped choose the genes; select inside each fold
B Five folds are too few; with ten folds there would be no problem
C It only leaks if fewer than 50 genes are truly related to the label
D Nothing: CV protects against this because each row is held out once

3.1 Regression Fundamentals

Q67 apply Which of these is a regression problem?

A Predicting whether an email is spam or not
B Predicting a patient's blood type from tests
C Grouping customers into similar segments
D Predicting tomorrow's electricity demand in MWh

Q68 concept In the model $Y = f(X) + \varepsilon$, what does ε represent?

A The uncertainty in the estimated slope coefficient
B The residuals the fitted model makes on the training set
C Random noise: everything the predictors don't explain
D The intercept, the value of Y when every X is 0

Q69 calculate A house sold for £250k. The model's fitted price was £235k. What is the residual, and what does its sign mean?

A +£15k: the model under-predicted
B +£15k: the model over-predicted
C −£15k: the model over-predicted
D −£15k: the model under-predicted

Q70 concept How are residuals related to the noise term ε?

A Residuals are always zero in a correct model
B Residuals are estimates of the unobservable errors
C Residuals are exactly equal to the true errors
D The errors are observed directly; residuals are only guessed

Q71 calculate A fitted model is $\hat{y} = 30 + 2.6x$, where x is floor area in m² and $\hat{y}$ is price in £k. What's the predicted price of a 100 m² house?

A £260k
B £32.6k
C £290k
D £2,630k

3.2 Simple and Multiple Linear Regression

Q72 `concept` Is $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$ a linear regression model?

- A Only when β_2 is close to zero
- B No: the X^2 term makes it logistic
- C Yes: it is linear in the coefficients
- D No: its curve isn't a straight line

Q73 `concept` Which of these models is NOT a linear regression model?

- A $Y = \beta_0 + \beta_1 X + \beta_2 X^3 + \varepsilon$
- B $Y = \beta_0 + \beta_1 X_1 X_2 + \varepsilon$
- C $Y = \beta_0 e^{\beta_1 X} + \varepsilon$
- D $Y = \beta_0 + \beta_1 \log X + \varepsilon$

Q74 `interpret` In $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$, what is the effect of a one-unit increase in X_1?

- A $\beta_1 \beta_3$
- B $\beta_1 + \beta_3 X_2$
- C β_1
- D $\beta_1 + \beta_3$

Q75 `concept` A multiple regression with two predictors fits what shape in (X_1, X_2, Y) space?

- A A straight line
- B A plane
- C A sphere
- D An S-shaped curve

Q76 `apply` A scatter plot shows a clearly curved relationship between X and Y. What's the simplest fix that stays within linear regression?

- A Add X^2 as an extra column and refit with OLS
- B Switch to logistic regression for the curve
- C Run K-means first to find the curve's shape
- D Convert Y to 0/1 and fit the data again

3.3 Ordinary Least Squares (OLS)

Q77 `calculate` A regression has $n = 50$ observations and $p = 4$ predictors, plus an intercept. What are the dimensions of the design matrix X?

- A 50×5
- B 5×5
- C 50×4
- D 4×50

Q78 concept When does $\hat{\beta} = (X^\top X)^{-1} X^\top y$ fail to exist?

A When $X^\top X$ is singular

B When the response has units

C When some residuals are negative

D When the model has an intercept

Q79 concept What does ordinary least squares minimise?

A The sum of the residuals

B The sum of absolute residuals

C The number of predictors used

D The sum of squared residuals

Q80 calculate For the data $(x, y) = (1, 1), (2, 3), (3, 2)$, with an intercept, what is $X^\top X$?

A $\begin{pmatrix} 3 & 6 \\ 6 & 12 \end{pmatrix}$

B $\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}$

C $\begin{pmatrix} 6 & 3 \\ 14 & 6 \end{pmatrix}$

D $\begin{pmatrix} 3 & 6 \\ 6 & 14 \end{pmatrix}$

Q81 diagnose A model includes both "height in cm" and "height in inches" as predictors. What happens when you fit it with OLS?

A Both columns get exactly the same coefficient

B It fits fine, just more slowly than usual

C $X^\top X$ is singular, so there's no unique solution

D OLS averages the two columns into one automatically

3.4 Calculating Regression Coefficients by Hand

Q82 calculate For the points $(x, y) = (0, 1), (1, 3), (2, 5), (3, 7)$, what are $\hat{\beta}_0$ and $\hat{\beta}_1$?

A $\hat{\beta}_0 = 0$, $\hat{\beta}_1 = 2.5$

B $\hat{\beta}_0 = 1$, $\hat{\beta}_1 = 2$

C $\hat{\beta}_0 = 1.5$, $\hat{\beta}_1 = 2$

D $\hat{\beta}_0 = 2$, $\hat{\beta}_1 = 1$

Q83 calculate For the points $(1, 2), (2, 1), (3, 4)$, what is the slope $\hat{\beta}_1$?

A 1.5

B 2/3

C 0.5

D 1

Q84 concept Why does the slope formula fail when every x_i has the same value?

A S_{xy} becomes negative, so the slope is invalid
B $S_{xx} = 0$, so the formula divides by zero
C The residuals no longer add up to zero
D The intercept becomes undefined first

Q85 concept In a model with an intercept, the OLS residuals always...

A have variance 1
B sum to zero
C are all positive
D equal the fitted values

Q86 interpret The fitted line for $x = 1, \ldots, 5$ and $y = 2, 4, 5, 4, 5$ is $\hat{y} = 2.2 + 0.6x$. Which point must it pass through?

A $(5, 5)$, the last observation
B $(3, 4)$, the point of means
C $(0, 0)$, the origin
D $(1, 2)$, the first observation

3.5 Interpreting Coefficients

Q87 calculate For the fitted model $\hat{y} = 5 + 2x_1 - 4x_2$, what is the prediction at $(x_1, x_2) = (3, 1)$?

A 15
B 7
C 11
D 3

Q88 interpret Predictor A is in metres with $\hat{\beta} = 40$; predictor B is in millimetres with $\hat{\beta} = 0.05$. Does A matter more?

A Yes, provided both coefficients are statistically significant
B Yes, because 40 is far bigger than 0.05
C No, because B is measured more precisely
D Can't tell from raw coefficients; standardise first

Q89 concept Why can a predictor's coefficient change sign when another predictor is added to the model?

A OLS estimates are random, so their signs often flip by chance
B Each coefficient is an effect holding the others fixed
C The intercept absorbs part of every coefficient
D Adding predictors always reverses existing signs

Q90 interpret An observational study finds a positive coefficient for "hours of TV per day" when predicting body weight. What can you conclude?

A Cutting TV by an hour will reduce weight by $\hat{\beta}$

B More TV is associated with higher predicted weight

C Watching more TV causes people to gain weight

D TV has no real effect, as studies are biased

Q91 interpret A model gives price (£k) $= -50 + 3\times$ area (m²). How should you read the intercept of −50?

A The model must be wrong, because a house price can never be a negative number

B Every house sells for £50k less than the model predicts

C Every house loses £50k of value before its area is counted

D It's the prediction at zero area, outside the data, so not meaningful

3.6 Residuals and Regression Assumptions

Q92 interpret A plot of residuals against fitted values shows a clear U-shape. What does it suggest?

A The model fits the data perfectly

B The errors are correlated over time

C The linear form is missing curvature

D The error variance is not constant

Q93 interpret The spread of the residuals triples as the fitted values increase. What is the issue, and what does it affect?

A Heteroskedasticity; SEs and p-values become unreliable

B Multicollinearity; coefficients flip their signs

C Underfitting; the R^2 drops all the way down to zero as well

D Leakage; the test score becomes inflated

Q94 interpret Residuals from a model of monthly sales, plotted in time order, show long runs of positive values followed by runs of negative ones. What's the likely issue?

A Heteroskedasticity: the spread keeps growing

B Perfect multicollinearity among the Xs

C Outliers: a few of the months are extreme

D Autocorrelation: the errors aren't independent

Q95 diagnose A colleague says: "My residual plot looks fine, so the model must be correct." What's the best response?

A Residual plots only matter when the sample size is small, so it's fine here

B Agreed: a clean residual plot confirms the model is correct

C A clean plot is consistent with the assumptions but proves nothing

D Only a Q–Q plot can show whether the model is correct

Q96 concept Does fitting OLS (computing $\hat{\beta}$) require normally distributed errors?

A Yes: without normal errors $\hat{\beta}$ can't be computed

B Yes: R^2 is only defined when the errors are normal

C No: the assumptions matter for SEs, p-values and CIs

D No: OLS results never rely on any assumptions at all

3.7 R^2, Adjusted R^2, RMSE and Other Metrics

Q97 calculate A model has RSS = 30 and TSS = 120. What is R^2?

A 0.25

B 0.75

C 0.80

D 4

Q98 calculate RSS = 30, TSS = 120, $n = 25$ and $p = 4$. What is the adjusted R^2?

A 0.80

B 0.70

C 0.75

D 0.65

Q99 calculate On three test cases the residuals are $2, -1, 3$. What are the MAE and the RMSE?

A MAE 2, RMSE 2.16

B MAE 2, RMSE 4.67

C MAE 2, RMSE 1.33

D MAE 1.33, RMSE 2.16

Q100 interpret A model's out-of-sample R^2 is −0.2. What does that mean?

A 20% of its predictions on new data are wrong

B It predicts worse than the training mean on new data

C The predictors are negatively correlated with Y

D There must be an error, since R^2 can't be negative

Q101 interpret A column of pure random noise is added to an OLS model. What typically happens to R^2 and adjusted R^2?

A R^2 falls slightly; adjusted R^2 tends to rise

B Neither changes, since it's unrelated to Y

C Both fall, since the column is pure noise

D R^2 rises slightly; adjusted R^2 tends to fall

3.8 Standard Errors, Confidence Intervals and p-values

Q102 `calculate` A coefficient is $\hat{\beta}_1 = 1.5$ with SE $= 0.5$. What are the t-statistic and an approximate 95% CI (use $t \approx 2$)?

- A $t = 3$, CI $(0.5,\ 2.5)$
- B $t = 3$, CI $(1,\ 2)$
- C $t = 0.75$, CI $(0.5,\ 2.5)$
- D $t = 0.33$, CI $(1,\ 2)$

Q103 `calculate` For $x = 1, 2, 3, 4$ and $y = 1, 3, 2, 4$: $\hat{\beta}_1 = 0.8$, RSS = 1.8 and $S_{xx} = 5$. What is $\mathrm{SE}(\hat{\beta}_1)$?

- A 0.18
- B 0.60
- C 0.424
- D 0.36

Q104 `interpret` What does a p-value of 0.03 for a coefficient mean?

- A If the true coefficient were 0, results this extreme would occur 3% of the time
- B The coefficient is large enough to matter in practice
- C There is a 3% chance that the true coefficient is exactly zero
- D There is a 97% chance that the fitted regression model is the correct one for the data

Q105 `interpret` A coefficient has $\hat{\beta} = 0.9$ with a 95% CI of $(-0.2,\ 2.0)$. What can you conclude?

- A The predictor has no effect, since the interval includes zero
- B Not significant at 5%; the true effect could be negative, zero or quite large
- C The effect is significant, since the estimate is positive and larger than the interval's lower end
- D The true coefficient is 0.9, with 95% certainty

Q106 `calculate` A regression has $n = 40$ observations and $p = 3$ predictors plus an intercept. How many degrees of freedom do the coefficient t-tests use?

- A 37
- B 40
- C 36
- D 39

3.9 Multicollinearity and VIF

Q107 `calculate` Regressed on the other predictors, predictor X_j has $R_j^2 = 0.95$. What is its VIF?

- A 95
- B 1.05
- C 0.95
- D 20

Q108 `calculate` A predictor's VIF is 16. By roughly how much is its coefficient's standard error inflated?

- A About 1.6 times
- B About 8 times
- C About 16 times
- D About 4 times

Q109 `interpret` Two sensors measure almost the same pollutant ($r = 0.98$). In a regression with both, one has a large p-value, though on its own it predicts the response well. Why?

- A The sensor is physically unrelated to the response being predicted
- B p-values are unreliable whenever the sample size is very large, as it is here
- C Given the other sensor, it adds little information, so its SE is inflated
- D The model has overfitted, so every p-value is meaningless

Q110 `concept` Why can a model with severe multicollinearity still predict well on new data from the same source?

- A The combined effect is well determined; only its split is unstable
- B Multicollinearity only affects the intercept, not the slopes
- C VIF only changes R^2, never the fitted coefficients
- D Correlated predictors always make predictions more accurate on similar data

Q111 `apply` Which of these is NOT a sensible way to deal with strongly correlated predictors?

- A Add a second copy of one of the correlated predictors
- B Fit ridge regression instead of plain OLS
- C Collect data where the predictors vary independently
- D Combine the correlated predictors into a single one

4.1 Regression versus Classification

Q112 `concept` Why is ordinary linear regression a poor choice for a yes/no outcome coded 0/1?

- A It always underfits whenever Y is binary
- B It cannot be fitted to 0/1 response values
- C It can only use one predictor at a time
- D Its predictions can fall below 0 or above 1

Q113 `diagnose` To predict which of three unordered product categories a customer buys, a team codes them 1, 2, 3 and fits linear regression. What's the core problem?

- A The predictions will be too accurate on the training data
- B Linear regression needs at least ten distinct target values
- C The categories should be coded 0, 1, 2 instead of 1, 2, 3
- D The coding imposes a meaningless order and equal spacing

Q114 concept Logistic regression is a linear model of what?

A The probability of $Y = 1$

B The odds of $Y = 1$

C The class label Y itself

D The log-odds of $Y = 1$

Q115 apply A logistic regression outputs $\hat{p} = 0.62$ for a customer. How does that become a class prediction?

A Round each coefficient to 0 or 1, then recompute $\hat{p}$

B The model outputs a class directly; 0.62 is its confidence

C Pick the class whose coefficient has the larger value

D Compare it with a threshold, e.g. predict 1 if $\hat{p} > 0.5$

Q116 apply You built a churn classifier. Which set of measures should you report?

A The VIF of each predictor and the intercept

B Sensitivity, specificity and AUC

C R^2 and the RMSE of the 0/1 predictions

D Adjusted R^2 and the coefficient p-values

4.2 Odds, Log-odds and the Sigmoid Function

Q117 calculate Convert the probability $p = 0.2$ to odds and to log-odds.

A Odds 0.2, log-odds −1.61

B Odds 0.8, log-odds −0.22

C Odds 0.25, log-odds −1.39

D Odds 4, log-odds 1.39

Q118 calculate The log-odds are 1.1. Using $e^{1.1} \approx 3$, what is the probability?

A 0.75

B 0.33

C 0.25

D 0.90

Q119 calculate A score is $z = -0.6$ (use $e^{0.6} \approx 1.82$). What is $p = \sigma(z)$?

A 0.354

B 0.549

C 0.646

D 0.400

Q120 interpret Odds of 3 for an event correspond to what probability?

A 0.33

B 0.67

C 0.30

D 0.75

Q121 `calculate` $p = 0.9$ has log-odds of about 2.2. What are the log-odds of $p = 0.1$?

A About 0.1
B About 0.22
C About −0.9
D About −2.2

4.3 The Logistic Regression Model

Q122 `calculate` A model has log-odds $= -3 + 0.5x_1 + 1.2x_2$. What is $P(Y = 1)$ at $(x_1, x_2) = (4, 1)$ (use $e^{-0.2} \approx 0.82$)?

A 0.55
B 0.20
C 0.82
D 0.45

Q123 `interpret` In that model, $e^{1.2} = 3.32$. How do you interpret it?

A Each one-unit rise in x_2 multiplies the odds of $Y = 1$ by 3.32
B Each one-unit rise in x_2 makes $Y = 1$ exactly 3.32 times as likely
C Each one-unit rise in x_2 raises the probability of $Y = 1$ by 3.32
D Each one-unit rise in x_2 adds 3.32 to the log-odds of $Y = 1$

Q124 `concept` If $\beta_1 = 0.7$, what does a one-unit increase in x_1 do?

A Adds 0.7 to the log-odds
B Multiplies the log-odds by 0.7
C Adds 0.7 to the probability
D Multiplies the probability by 0.7

Q125 `interpret` Two customers have churn probabilities 0.10 and 0.50. Each makes one more complaint, which has an odds ratio of 2.23. What happens?

A Both probabilities rise by the same amount, since the odds ratio is fixed
B The 0.10 customer's probability rises more, since it has more room to grow
C Both probabilities are multiplied by 2.23, giving 0.22 and 1.12
D Both odds rise ×2.23, but the 0.50 customer's probability rises more

Q126 `concept` Where is logistic regression's decision boundary at threshold 0.5?

A Where the predicted probability reaches exactly 1 everywhere
B Along the S-shaped sigmoid curve of predicted probabilities
C Where $\beta_0 + \beta^\top x = 0$, a straight line or hyperplane
D Along a circle centred on the mean of the training data

4.4 Maximum Likelihood Estimation

Q127 concept Why can't logistic regression be fitted with a closed-form formula like OLS's $(X^\top X)^{-1}X^\top y$?

- A Its loss is not convex, so there are many local optima
- B β sits inside the sigmoid, so the equations are nonlinear
- C It has too many parameters to invert $X^\top X$ in practice
- D Logistic regression doesn't have a loss function to minimise

Q128 calculate Labels $y = (1, 1, 0)$; predicted probabilities $p = (0.7, 0.4, 0.2)$. What is the log-likelihood?

- A About −0.83
- B About −2.10
- C About −1.50
- D About −0.51

Q129 concept Why do we maximise the log-likelihood rather than the likelihood itself?

- A Taking logs prevents the model from overfitting the data
- B The plain likelihood cannot be computed for logistic models
- C The log gives a different, more accurate set of coefficients
- D Logs turn products into sums, and the maximiser is unchanged

Q130 calculate A coefficient is $\hat{\beta} = -0.9$ with SE $= 0.45$. Is it significant at 5% (Wald test)?

- A Yes: $z = -2$ and $|z| > 1.96$
- B Yes: $|\hat{\beta}|$ is bigger than its standard error
- C No: $z = -0.5$, which is too small
- D No: a negative coefficient can't be significant

Q131 calculate A logistic model with no predictors is fitted to labels $(1, 0, 0, 1, 1)$. What is the maximum-likelihood estimate of p?

- A 0.4
- B 0.6
- C 1.0
- D 0.5

4.5 Cross-entropy / Log-loss

Q132 calculate For a case with true label 1, what is the cross-entropy loss when the model predicts 0.8, and when it predicts 0.2?

- A 1.609 and 0.223
- B 0.223 and 1.609
- C 0.8 and 0.2
- D 0.2 and 0.8

Q133 calculate For a case with true label 0, the model predicts $p = 0.1$ for class 1. What's the loss?

A 0.9
B 0.1
C 0.105
D 2.303

Q134 calculate Two cases: $y = (1, 0)$ with predictions $p = (0.9, 0.2)$. What is the mean cross-entropy?

A 1.609
B 0.164
C 0.150
D 0.328

Q135 concept Which loss does standard logistic regression minimise?

A Cross-entropy (negative log-likelihood)
B 0–1 loss, i.e. the misclassification rate
C Hinge loss, as in support vector machines
D Squared error on the 0/1 labels

Q136 interpret Two classifiers have identical accuracy, but A has a much lower cross-entropy than B. What does that tell you?

A A's probabilities are better: confident when right, cautious when wrong
B Nothing, because only accuracy matters once a threshold is fixed
C A is overfitting, because a lower loss means it memorised the data
D A must have used a better threshold than B to classify the cases

4.6 The Confusion Matrix

Q137 concept A fraud model flags a legitimate transaction as fraud. Which cell of the confusion matrix is that?

A True negative
B False negative
C True positive
D False positive

Q138 concept A screening test reports "no disease" for a patient who has the disease. Which cell is that?

A False positive
B False negative
C True positive
D True negative

Q139 calculate At threshold 0.5, four cases have (score, actual): (0.8, +), (0.6, −), (0.3, +), (0.2, −). What are TP, FP, FN and TN?

A 2, 1, 0, 1
B 1, 0, 1, 2
C 2, 0, 0, 2
D 1, 1, 1, 1

Q140 apply For an email spam filter, which error is usually worse?

A A false negative: one spam email reaches the inbox
B They always cost the same, so neither matters more
C A false positive: a real email is hidden in spam
D A true negative: the filter didn't learn anything

Q141 concept In hypothesis-testing language, a false positive is a...

A Type III error
B Type II error
C Sampling error
D Type I error

4.7 Accuracy, Precision, Recall and Specificity

Q142 calculate TP = 30, FN = 20, FP = 10, TN = 140. What are the sensitivity and specificity?

A 0.60 and 0.933
B 0.75 and 0.875
C 0.85 and 0.933
D 0.60 and 0.875

Q143 calculate Same table (TP = 30, FN = 20, FP = 10, TN = 140). What is the precision?

A 0.933
B 0.75
C 0.85
D 0.60

Q144 concept Which formula gives specificity?

A $TP/(TP+FP)$
B $TN/(TN+FP)$
C $TP/(TP+FN)$
D $FP/(FP+TN)$

Q145 interpret A test has sensitivity 0.95 and specificity 0.60. In plain words?

A It misses many sick people but rarely raises a false alarm
B It gives the correct answer for about 95% of all people tested
C It catches almost every sick person but flags 40% of healthy ones
D 60% of the people it flags as positive are actually sick

Q146 calculate A disease affects 1% of people. A test has 90% sensitivity and 90% specificity. Of the people who test positive, roughly what share actually have the disease?

A About 90%
B About 50%
C About 1%
D About 8%

4.8 Sensitivity and Specificity Trade-offs

Q147 interpret You lower the classification threshold from 0.5 to 0.2. What generally happens?

A Specificity rises and sensitivity falls
B Both sensitivity and specificity rise
C Sensitivity rises and specificity falls
D Neither changes; only accuracy does

Q148 concept With the same fitted model, can you raise sensitivity and specificity together by moving the threshold?

A Only if the model's accuracy is above 90%
B Yes, but only when the classes are balanced
C No: moving it trades one for the other
D Yes: there is always a threshold that improves both

Q149 interpret What happens at a threshold of 0, where every case is flagged positive?

A Sensitivity 0, specificity 1
B Sensitivity 1, specificity 0
C Accuracy becomes 1
D Both equal 0.5

Q150 diagnose A team lowers the threshold until sensitivity reaches 100%, then reports this as a big success. What's the catch?

A Sensitivity can never genuinely reach 100% on real-world data
B It means the model has overfitted the training data
C Lowering the threshold also lowers the model's AUC
D Flagging nearly everyone floods them with false positives

Q151 apply What would improve both sensitivity and specificity at the same time?

A Lowering the threshold well below the 0.5 default
B Raising the threshold well above the 0.5 default
C Better features, so the two classes' scores overlap less
D Oversampling the positive class in the test set before scoring

4.9 Threshold Selection and Classification Costs

Q152 `calculate` A false positive costs 1 and a false negative costs 9. At threshold 0.5 there are 5 FP and 10 FN; at 0.2 there are 25 FP and 3 FN. Which threshold has the lower cost?

- A 0.2, with cost 52 vs 95
- B They cost the same
- C 0.5, with cost 15 vs 28
- D 0.5, with cost 95 vs 52

Q153 `calculate` A false positive costs £500 and a false negative £1,500. With well-calibrated probabilities, which threshold minimises expected cost?

- A 0.50
- B 0.25
- C 0.33
- D 0.75

Q154 `apply` A missed diagnosis is 20 times as costly as a false alarm. Roughly where should the threshold be?

- A At 0.5, the default
- B Well above 0.5, around 0.95
- C Well below 0.5, around 0.05
- D Around 0.2, a typical value

Q155 `diagnose` A team picks the threshold that minimises cost on the test set and reports that cost as its expected cost in production. What's wrong?

- A They should always use 0.5 so the result is comparable to others
- B Costs can only be measured on the training set, never the test set
- C Nothing: costs, unlike accuracy, can't be overfitted at all
- D The threshold was tuned on the test set, so the cost is optimistic

Q156 `concept` If a false negative costs more than a false positive, which way should the threshold move from 0.5?

- A It should stay at 0.5
- B Down, so more cases are flagged
- C Up, so fewer cases are flagged
- D It depends only on accuracy

4.10 ROC Curve and AUC

Q157 `concept` What does an ROC curve plot?

- A Precision vs recall, over all possible thresholds
- B True positive rate vs false positive rate, over all thresholds
- C Accuracy vs the threshold value, from 0 up to 1
- D Specificity vs precision, measured at the default threshold of 0.5

Q158 calculate Positives are scored 0.8, 0.5, 0.4; negatives are scored 0.6, 0.3. What is the AUC, by counting pairs?

A 5/6 ≈ 0.83
B 3/6 = 0.50
C 4/6 ≈ 0.67
D 2/5 = 0.40

Q159 calculate There are 4 positives and 6 negatives. At some threshold, 3 positives and 2 negatives are flagged. Where is this point on the ROC curve?

A TPR 0.75, FPR 0.67
B TPR 0.33, FPR 0.75
C TPR 0.75, FPR 0.33
D TPR 0.60, FPR 0.40

Q160 interpret Model A has AUC 0.92 and model B has AUC 0.55. Which statement is best supported?

A B is useless at every threshold and must be scrapped
B A is correct for about 92% of the cases it scores
C A ranks positives above negatives much more reliably
D A's best threshold for classifying is exactly 0.5

Q161 concept Why is AUC called "threshold-independent", and what can't it tell you?

A It always uses 0.5, so it can't handle unequal costs
B It summarises all thresholds, so it can't say which one to use
C It ignores the scores, so it can't compare two models
D It uses only the best threshold, so it hides the trade-off

4.11 Class Imbalance

Q162 diagnose A fraud model reaches 99.5% accuracy on data where 0.5% of transactions are fraud. Why isn't this impressive?

A Accuracy above 99% is a sure sign of overfitting
B Predicting "not fraud" for everything also scores 99.5%
C Accuracy can't be computed properly when classes are imbalanced
D Fraud models need 100% accuracy to be deployable

Q163 diagnose "We oversampled the minority class, then split the data 80/20 for evaluation." What can go wrong?

A The test set ends up too small to give a reliable score
B The model becomes unable to predict the minority class on new data
C Oversampling removes the majority class from the data
D Copies of the same records land in both sets, inflating the score

Q164 concept In stratified oversampling, the training set is balanced. What should the validation set look like?

A It should reuse the training cases
B It should keep the original class ratio
C It should be balanced 50/50 as well
D It should contain only the rare class

Q165 interpret 95% of cases are negative. A model has 96% accuracy and 0.20 sensitivity. How good is it?

A Excellent, because 96% accuracy is far above the 50% of chance
B Barely better than the 95% baseline, and it misses 80% of positives
C Good, because a sensitivity of 0.20 is quite normal for rare classes like this
D Poor overall, because accuracy below 99% is not acceptable

Q166 concept What does giving the rare class a larger weight in the loss do?

A Makes its errors costlier, so the model attends to it more
B Moves the classification threshold to exactly 0.5
C Removes most of the majority-class cases from training
D Copies the rare cases into the validation set

5.1 Why Regularisation is Needed

Q167 concept Which two objectives does regularisation balance?

A Fitting the training data well, and keeping the model simple
B Low bias, and as little irreducible noise in the data as possible
C High precision, and high recall at one threshold
D Training speed, and the memory the model needs

Q168 apply Which of these datasets most needs regularisation?

A 10,000 rows with 3 uncorrelated predictors
B 1,000 rows with a single numeric predictor
C 50,000 rows with 10 well-measured predictors
D 80 patients with 500 gene measurements each

Q169 concept Which of these is NOT a benefit of a more parsimonious model?

A Easier to interpret and explain
B A guaranteed lower training error
C More robust when new data arrive
D Cheaper to collect inputs for

Q170 concept What does a regularised regression reduce to when $\lambda = 0$?

A LASSO regression
B An undefined model
C Ordinary least squares
D The mean-only (intercept) model

Q171 concept Why does plain OLS break down when there are more predictors than observations?

- A $X^\top X$ is singular, so infinitely many fits are perfect
- B OLS needs every predictor to be standardised first
- C The intercept can't be estimated when $p > n$
- D The residuals become too large to be minimised

5.2 Model Complexity and Feature Selection

Q172 concept Which change makes a model MORE complex?

- A Increasing λ in ridge
- B Increasing K in K-means
- C Increasing K in K-NN
- D Decreasing C in an SVM

Q173 concept Why can't training RSS be used to decide how many predictors to keep?

- A It can only be computed for models with a single predictor at a time
- B It never increases when a predictor is added, so it picks the biggest model
- C It always favours the smallest model, because fewer terms mean less error
- D It depends on the units of Y, so different models can't be compared

Q174 diagnose "We dropped every variable with $p > 0.05$, refitted, and all the remaining p-values are below 0.05." What's the main problem?

- A They should have kept every variable in the final model, significant or not
- B The selection is ignored, so the new p-values are over-confident
- C p-values can't be recalculated after a model is refitted
- D The threshold should have been 0.01 instead of 0.05

Q175 diagnose A team picks the 30 features most correlated with the target using all the data, then uses CV to choose a model. Is that sound?

- A Yes: CV protects the estimate as long as the folds are chosen at random
- B No: they should have used the test set to choose the 30 features instead
- C No: selection used the validation folds' targets; select inside each fold
- D Yes: using correlation, rather than p-values, avoids any form of leakage

Q176 interpret As more variables are added, in-sample error keeps falling but out-of-sample error is lowest at about 8 variables. What does that show?

- A Out-of-sample error is unreliable with many variables
- B The in-sample error must have been computed wrongly
- C Beyond about 8 variables the model starts to underfit
- D Beyond about 8 variables the model starts to overfit

5.3 Best Subset and Stepwise Selection

Q177 calculate How many models does best subset selection fit with $p = 8$ candidate predictors, including the empty model?

A 256
B 8
C 64
D 37

Q178 calculate How many models does forward stepwise selection fit with $p = 8$, including the empty model?

A 37
B 256
C 9
D 36

Q179 concept Best subset gives the best model of each size, $M_0, \ldots, M_p$. How should you choose among them?

A Validation or CV error, or a criterion that penalises size
B The largest model, since it uses the most information available
C The lowest training RSS, since it measures the fit
D The model with the fewest predictors that has $R^2 > 0$

Q180 apply Forward stepwise picks B first, then adds A. But the best two-variable model is actually {A, C}. How is that possible?

A It isn't possible: stepwise always finds the best pair
B The RSS for {A, C} must have been computed incorrectly
C Forward stepwise only works if the predictors are standardised
D Greedy steps can't undo B, so the best pair can be missed

Q181 concept Which penalty corresponds to subset selection?

A λ times the sum of absolute coefficients
B λ times the number of non-zero coefficients
C λ times the sum of squared coefficients
D λ times the largest absolute coefficient

5.4 Ridge Regression (L2)

Q182 calculate Ridge with one centred predictor: $\sum xy = 6$ and $\sum x^2 = 3$. What is the slope at $\lambda = 3$?

A 1
B 2
C 6
D 0.5

Q183 calculate $X^\top X = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ and $X^\top y = (6, -2)^\top$. What is the ridge solution at $\lambda = 2$?

A $(6,\ -2)$
B $(3,\ -1)$
C $(1.5,\ -0.5)$
D $(2,\ -0.67)$

Q184 concept Which statement about ridge regression is true?

A It has a closed-form solution
B Its bias falls as λ increases
C It can set coefficients exactly to zero
D It needs fewer predictors than rows

Q185 concept Why does ridge always have a unique solution when $\lambda > 0$?

A It removes the intercept, which is the source of singularity
B It always fits by gradient descent, which never fails
C Adding λ lifts every eigenvalue of $X^\top X$ above zero
D It quietly drops one of every pair of correlated predictors

Q186 interpret What happens to a ridge model as $\lambda \to \infty$?

A It turns into OLS, the unpenalised least-squares fit
B Slopes become exactly 0 at some finite value of λ
C Slopes shrink towards 0 and it predicts $\bar{y}$ for everyone
D The intercept is shrunk to 0, so it predicts 0 for everyone

5.5 LASSO Regression (L1)

Q187 calculate Orthonormal design with OLS coefficients $b = (4, -0.8, 1.2)$. What does LASSO give with threshold $\lambda/2 = 1$?

A $(3,\ -0.8,\ 0.2)$
B $(2,\ -0.4,\ 0.6)$
C $(3,\ 0,\ 0.2)$
D $(4,\ 0,\ 1.2)$

Q188 calculate Same OLS coefficients $b = (4, -0.8, 1.2)$, orthonormal design. What does ridge give with $\lambda = 1$?

A $(3,\ 0,\ 0.2)$
B $(4,\ -0.8,\ 1.2)$
C $(2,\ -0.4,\ 0.6)$
D $(3,\ -1.8,\ 0.2)$

Q189 apply You have 500 predictors, believe only about 10 matter, and need an interpretable model. Ridge or LASSO?

- A LASSO, because its exact zeros give a sparse model
- B Ridge, because it keeps all 500 predictors stable
- C Ridge, because LASSO can't handle 500 predictors
- D LASSO, because it has a closed-form solution

Q190 interpret As λ increases in LASSO, what typically happens?

- A The variance of the fitted model increases
- B The bias of the fit decreases
- C The number of non-zero coefficients falls
- D The training error falls

Q191 apply Two predictors are almost exact copies of each other. What does LASSO tend to do with them?

- A Set both to zero, because they're redundant together
- B Keep one and set the other to zero, somewhat arbitrarily
- C Give both of them exactly the same non-zero coefficient
- D Fail to converge, because $X^\top X$ is singular

5.6 Understanding λ (Lambda)

Q192 concept For which value of λ does a regularised regression achieve its lowest training error?

- A $\lambda = 1$
- B $\lambda \to \infty$
- C The value chosen by cross-validation
- D $\lambda = 0$

Q193 concept Ridge regression with $\lambda = 0$ is...

- A ordinary least squares
- B undefined
- C the mean-only (intercept) model
- D LASSO regression

Q194 interpret Training MSE goes from 0.8 to 1.1 and test MSE from 3.0 to 1.9 as λ rises from 0 to 5. How do you read this?

- A OLS was overfitting; the penalty cut variance more than it added bias
- B $\lambda = 5$ is underfitting, since the training error increased
- C Regularisation made the model worse, since training error went up
- D The test set must have leaked, since test error fell so much

Q195 concept Compared with OLS, the ridge solution for some $\lambda > 0$ has...

A higher RSS, but a lower penalised objective

B lower RSS, but a higher penalised objective

C exactly the same RSS as OLS on the data

D lower RSS and a lower penalised objective

Q196 interpret In a LASSO path, the coefficient for variable A hits zero at $\lambda = 0.05$ and the one for B at $\lambda = 3$. What does this suggest?

A B matters more: it survives much stronger penalties

B A matters more, because it was selected first

C They matter equally, since both reach zero

D Neither matters, because both are eventually dropped

5.7 Ridge versus LASSO

Q197 concept What is the main difference between ridge and LASSO solutions?

A LASSO can set coefficients exactly to zero; ridge only shrinks them

B Ridge can set coefficients exactly to zero; LASSO only shrinks them

C Ridge needs standardised predictors, while LASSO works on raw units

D LASSO has a closed-form solution, while ridge must be found iteratively

Q198 concept Geometrically, why does LASSO give exact zeros?

A Its round constraint region touches the RSS contours along an axis

B Its diamond constraint has corners on the axes, where contours tend to touch

C It searches over every subset of predictors first, and only then shrinks the coefficients

D Its RSS contours are circles centred on the origin of the coefficients

Q199 apply You have 30 highly correlated predictors that all plausibly contribute a little. Which would you favour?

A LASSO, because it keeps the most important one

B Best subset selection, because p is small

C Ridge, because it shares the credit across them

D Plain OLS, because correlation doesn't matter

Q200 concept Why is LASSO called a middle ground between best subset selection and ridge?

A It averages the coefficients from best subset and ridge into a single model

B It has ridge's closed form, but only uses half of the predictors at a time

C It searches every subset like best subset, but adds ridge's penalty to each

D It selects variables like subset, but is a cheap shrinkage problem like ridge

Q201 concept What is elastic net?

A LASSO applied to logistic regression

B A penalty that mixes the L1 and L2 penalties

C Ridge with a separate λ per predictor

D Best subset selection with a ridge penalty added on top

5.8 Mathematical Objectives and Worked Calculations

Q202 calculate For $\beta = (2, -1, 0, 3)$, what are $\|\beta\|_0$, $\|\beta\|_1$ and $\|\beta\|_2^2$?

- A 3, 4, 14
- B 4, 6, 14
- C 3, 6, 14
- D 3, 6, 6

Q203 calculate A fit has RSS = 20 and $\hat{\beta} = (1, -2, 0)$; take $\lambda = 3$. What are the ridge, LASSO and L0 objectives?

- A 23, 23, 23
- B 35, 26, 29
- C 29, 35, 26
- D 35, 29, 26

Q204 calculate A logistic fit has $-\ell = 8.0$ and $\beta = (2, -1, 0.5)$; $\lambda = 2$. What is the LASSO-penalised objective?

- A 11.5
- B 15.0
- C 18.5
- D 9.0

Q205 interpret Fit A: RSS 10, $\beta = (4, -3)$. Fit B: RSS 13, $\beta = (2, -1)$. Under ridge with $\lambda = 0.1$, which has the lower objective?

- A A, with 12.5 against 13.5
- B They tie at 13
- C B, with 13.5 against 12.5
- D B, with 18 against 35

Q206 calculate A fit has RSS = 12 and $\hat{\beta} = (3, -2, 0.5, 0)$; $\lambda = 2$. What is the subset (L0) objective?

- A 23
- B 18
- C 20
- D 38.5

5.9 Selecting λ Using Cross-validation

Q207 concept How should λ be chosen?

- A As large as possible, to be safe against overfitting
- B By the lowest training error
- C By the lowest error on the test set
- D By cross-validation on the training data

Q208 diagnose "We standardised using the whole dataset, then cross-validated λ." What's the issue?

A They should have used min–max scaling instead
B Ridge and LASSO don't need standardising
C Standardising makes λ impossible to tune
D Leakage: held-out folds influenced the scaling

Q209 concept What does the one-standard-error rule choose?

A The smallest λ within one SE of the minimum
B The largest λ within one SE of the minimum CV error
C The λ that gives the minimum CV error exactly
D The λ whose CV error has the smallest standard error across folds

Q210 interpret OLS, ridge and LASSO reach test RMSE of 0.478, 0.477 and 0.477 on a dataset with 6,000 rows and 11 predictors. Was regularisation pointless?

A No: ridge is clearly better than OLS, since 0.477 is less than 0.478
B Yes: the differences show regularisation was applied incorrectly here
C Yes: if accuracy doesn't improve, plain OLS is always the better choice
D Not necessarily: it can still stabilise coefficients and drop redundant ones

Q211 concept Why must predictors be standardised before ridge or LASSO?

A Standardising removes multicollinearity between predictors
B The OLS solution can't be computed without standardising
C The penalty depends on coefficient size, which depends on units
D Standardising guarantees that the training error will decrease afterwards

6.1 K-means Clustering

Q212 concept Which description fits K-means?

A Unsupervised; maximises the margin between groups
B Supervised; maximises the likelihood of the labels
C Supervised; minimises the hinge loss on labelled data
D Unsupervised; minimises the within-cluster sum of squares

Q213 concept In the update step, where does each centroid move?

A To the mean of the points assigned to it
B To a randomly chosen point in its cluster
C To the median of the points assigned to it
D To the point furthest from other centroids

Q214 concept What does K mean in K-means, and in K-nearest neighbours?

A Both: the number of clusters the data is split into
B K-means: number of clusters. K-NN: number of neighbours
C Both: the number of neighbours used for each point
D K-means: number of features. K-NN: number of classes

Q215 diagnose A hospital has labelled outcomes and uses K-means with $K = 2$ to predict whether patients will be readmitted. What's wrong?

A Labels exist, so it should use a supervised classifier
B K should be 3, to allow for an extra "unsure" group of patients
C K-means needs at least 10 clusters to be reliable
D Nothing: two clusters match the two outcomes

Q216 concept What does K-means produce?

A A predicted label and a probability for each point
B A ranked list of the most important features
C A decision boundary with a margin between classes
D A cluster for each point, plus the K centroids

6.2 Distance Calculations and Cluster Assignment

Q217 calculate What is the Euclidean distance between $(1, 2)$ and $(4, 6)$?

A 25
B 7
C 5
D 3.5

Q218 calculate Centroids are $\mu_1 = (0, 0)$ and $\mu_2 = (6, 0)$. Which cluster does the point $(2, 2)$ join?

A Cluster 1 (squared distance 8 vs 20)
B Cluster 2 (distance 4 vs 2.8)
C Neither: it is equidistant
D Cluster 2 (squared distance 8 vs 20)

Q219 calculate What is the Manhattan distance between $(1, 1)$ and $(4, 5)$?

A 5
B 25
C 7
D 12

Q220 diagnose A retailer clusters customers on annual income (£20k–£200k) and store visits (1–40), unscaled. The clusters split purely by income. Why?

A K was set too small to separate the different visit-frequency patterns
B Store visits are irrelevant to how customers behave
C K-means ignores any feature with fewer than 50 values
D Income's huge range dominates the distance; standardise first

Q221 concept Why can K-means compare squared distances instead of distances when assigning points?

A Squaring preserves the order, so the nearest centroid is the same
B Squared distances reduce the effect of outliers
C Squared distance is a more accurate measure than ordinary distance
D K-means requires every distance to be a whole number

6.3 Updating Centroids and Convergence

Q222 calculate 1-D points 1, 3, 8, 10 with $K = 2$ and initial centroids 1 and 10. What are the final centroids and J?

A 1 and 10; $J = 8$
B 2 and 9; $J = 2$
C 4 and 9; $J = 10$
D 2 and 9; $J = 4$

Q223 calculate 2-D points $(0, 0), (0, 2), (6, 0), (6, 2)$, initial centroids $(0, 0)$ and $(6, 0)$. What are the final centroids?

A $(3, 0)$ and $(3, 2)$
B $(0, 1)$ and $(6, 1)$
C $(0, 2)$ and $(6, 2)$
D $(0, 0)$ and $(6, 0)$

Q224 interpret WCSS for $K = 1$ to 6 is 100, 45, 22, 19, 17, 16. Which K would you choose?

A 6, the lowest WCSS
B 1, the simplest
C 2, the biggest drop
D 3, the elbow

Q225 concept Why is K-means usually run several times from different random starts?

A Restarts standardise the features automatically
B It only finds a local optimum, which depends on the start
C The algorithm doesn't converge at all without several restarts
D Each run is used to try a different value of K

Q226 concept Why can't you choose K by finding the K with the lowest J?

A J depends on the random start, so it's meaningless
B J always rises as K grows, so it would pick $K = 1$
C J always falls as K grows, reaching 0 at $K = n$
D J can't be computed for more than ten clusters

6.4 SVM Fundamentals

Q227 concept How does a linear SVM classify a new point x?

- A By the nearest centroid
- B By the sign of $w^\top x + b$
- C By a vote of its neighbours
- D By the sigmoid of $w^\top x + b$

Q228 apply When the classes can be separated, which line does a hard-margin SVM choose?

- A The one with the widest margin to the nearest points
- B The one passing through the two class means
- C The one closest to the larger of the two classes
- D The first line the optimiser finds that separates the classes

Q229 concept What does a standard SVM output, compared with logistic regression?

- A A score and a class, but no probability by default
- B A probability and a class, just like logistic regression
- C A probability, but no class until calibrated
- D Only the support vectors, not a class

Q230 concept Which labels does the standard SVM formulation use?

- A $y \in \{1, 2\}$
- B $y \in \{0, 1\}$
- C $y \in \{-1, +1\}$
- D $y \in [0, 1]$

Q231 concept Which training points affect a linear SVM's boundary?

- A Only the training points that end up misclassified by the SVM
- B Only the point closest to each class's centre
- C Every training point, weighted equally in the fit
- D Only the support vectors: points on or inside the margin

6.5 Margins and Support Vectors

Q232 calculate An SVM has $w = (3, 4)$ and $b = -5$. What is the margin width?

- A 2
- B 0.4
- C 5
- D 0.2

Q233 calculate Same SVM ($w = (3, 4)$, $b = -5$). What is the distance of the point $(2, 1)$ from the boundary, and its class?

- A Distance 1, class −1
- B Distance 0.4, class +1
- C Distance 1, class +1
- D Distance 5, class +1

Q234 calculate The positive support vector is $(4, 0)$ and the negative one is $(2, 0)$, with $w = (a, 0)$. Find a, b and the margin width.

A $a = 0.5$, $b = -1.5$, width 4
B $a = 2$, $b = -6$, width 1
C $a = 1$, $b = -3$, width 2
D $a = 1$, $b = 3$, width 2

Q235 concept What happens if you delete a training point that is not a support vector?

A The boundary shifts towards the deleted point
B Nothing: the boundary stays exactly the same
C The margin doubles in width to compensate
D All the remaining points become support vectors

Q236 concept Why does minimising $\frac{1}{2}\|w\|^2$ maximise the margin?

A It reduces the number of misclassified training points
B The width is $\|w\|^2$, so it gets smaller as well
C It forces the bias term b to be exactly zero
D The width is $2/\|w\|$, so a smaller $\|w\|$ widens it

6.6 Hard-margin and Soft-margin SVM

Q237 calculate What is the hinge loss for a point with $y = +1$ and $f(x) = 0.3$?

A 1.3
B 0.3
C 0
D 0.7

Q238 calculate What is the hinge loss for a point with $y = -1$ and $f(x) = 0.5$?

A −0.5
B 0
C 1.5
D 0.5

Q239 calculate A soft-margin SVM has $\frac{1}{2}\|w\|^2 = 1$ and slack values summing to 1.5. What is the objective with $C = 2$?

A 5
B 2.5
C 4
D 3.5

Q240 interpret A point's slack (hinge loss) is $\xi = 0.4$. Where is it?

A Exactly on the decision boundary
B On the wrong side of the decision boundary
C On the correct side, but inside the margin
D Outside the margin on the correct side

Q241 concept Why is a soft margin needed in practice?

- A Real data overlap, so hard-margin constraints may be impossible
- B Hard-margin SVMs can't produce support vectors at all
- C Hard-margin SVMs can only handle two input features
- D A soft margin always gives a narrower and therefore more accurate fit

6.7 The C Parameter and Kernels

Q242 interpret What generally happens when you increase C in a soft-margin SVM?

- A The kernel changes from linear to RBF automatically
- B No change, since C only affects the speed of training
- C A narrower margin, fewer violations, and more risk of overfitting
- D A wider margin, more violations, and more risk of underfitting

Q243 concept The SVM's C plays a role most like which quantity?

- A λ in ridge regression
- B The threshold in logistic regression
- C K in K-means clustering
- D $1/\lambda$ in ridge regression

Q244 concept What does the kernel trick avoid?

- A Using support vectors to define the boundary
- B Standardising the features before training
- C Choosing a value for C by cross-validation on the training data
- D Computing the high-dimensional feature map $\phi(x)$ explicitly

Q245 interpret An RBF-kernel SVM uses a very large γ. What's most likely?

- A A boundary identical to a linear SVM's
- B A smooth boundary that underfits the data
- C No support vectors at all in the solution
- D A wiggly, very local boundary that overfits

Q246 apply Two classes form concentric rings. Which method can separate them?

- A An SVM with an RBF kernel
- B A linear SVM with a large C
- C Logistic regression on x_1, x_2
- D K-means with $K = 2$

Answer key

Each answer comes with the reasoning, including why the most tempting wrong option is wrong.

Q1 **C** · lesson 1.1

ML pays off when the rule is too intricate or changes too fast to write and maintain, and examples with known answers are available. It never guarantees zero errors. Rules *can* use many variables; the problem is writing and updating them. And this supervised setup only works because past fraud is labelled.

Q2 **B** · lesson 1.1

The point of learning is predicting **new** cases. Pure memorisation is perfect on the training data and silent everywhere else, however big the training set. Speed and outliers aren't the issue.

Q3 **A** · lesson 1.1

Features X go in, the fitted model f maps them to a prediction, and the hat marks $\hat{y}$ as an estimate, not the observed y. In ML the rule f is learned, not hand-written.

Q4 **A** · lesson 1.1

Predictions exist to support decisions. When misses are much costlier than false alarms, the model that misses fewer cases can be more valuable even with lower accuracy. Accuracy treats both errors as equal, which is exactly what's wrong here.

Q5 **A** · lesson 1.1

Both learn from data. Statistics has emphasised inference (is this effect real?), ML predictive accuracy on new data. Modern practice uses both: you'll compute p-values and test-set errors in this course.

Q6 **D** · lesson 1.2

There is no target column and the goal is to find structure, so it's unsupervised clustering. It can't be classification, because the groups aren't known in advance: there are no labels to learn from.

Q7 **B** · lesson 1.2

The target is a quantity you can average, so it's regression; being a whole number doesn't make it a category. It's supervised because past stays are known. It's not forecasting a series: each patient is a separate prediction.

Q8 **B** · lesson 1.2

The label is the answer the model must predict (default); the features are the inputs used to predict it. Using "defaulted" as a feature to predict something else would turn the question upside down.

Q9 **B** · lesson 1.2

With limited data and a shape that really is close to linear, a parametric model's assumption is an asset: few parameters, stable estimates. K-NN can do regression, but needs more data. Non-parametric models usually have *lower* bias and higher variance, and linear regression assumes linearity.

Q10 **C** · lesson 1.2

The numbers 1, 2, 3 are just names, but regression treats blue as "twice as far" from red as green is. That's meaningless, so use a classifier. Rounding hides the issue without fixing it, and the colours *are* labels, so it isn't clustering.

Q11 **B** · lesson 1.3

Anything learned from the full data (means for scaling, which features look useful) lets held-out information leak into the model. Split first; then learn scaling and feature choices from the training part only and apply them to the rest.

Q12 **D** · lesson 1.3

Hyperparameters are settings you choose (depth, degree, K, λ, C), usually on validation data. Slopes, intercepts and centroids are all computed from the training data by the fitting procedure.

Q13 **B** · lesson 1.3

The algorithm is the learning technique; complexity is how flexible that technique is allowed to be. Two depths of the same tree differ only in flexibility, and depth 15 can carve out far finer regions, so it's much more likely to overfit.

Q14 **C** · lesson 1.3

Iterating is normal, but the test set is meant to stay out of the loop. Once a test result drives a change, that test set is effectively validation data. Improve using validation data or CV, then report on a fresh test set.

Q15 **B** · lesson 1.3

Looking at errors often reveals problems upstream: the task is mis-defined, a feature is missing, the model is too simple. What must not loop is the test set; it's used once at the end.

Q16 **C** · lesson 1.4

NFL is an average over every conceivable problem, most of which look like noise. Real problems have structure, so methods do differ in practice. The lesson is to compare candidates on relevant data, not that the choice doesn't matter.

Q17 **B** · lesson 1.4

Success depends on how well an algorithm's assumptions match the problem. Earlier projects may have had different structure. NFL doesn't say all models are equal on *your* data, nor that simple ones always win.

Q18 **D** · lesson 1.4

Training data alone says nothing about unseen inputs. Learning works because the true relationship has structure the model's assumptions can exploit, and the data come from the same population you'll predict for. Memorising does the opposite.

Q19 **C** · lesson 1.4

Models assume future data look like training data. Different cameras change colour, lighting and resolution, breaking that assumption (distribution shift). No algorithm choice fixes it alone; you need representative data from the new setting.

Q20 **D** · lesson 1.4

Several different models can achieve nearly the same fit. That's why coefficients can change a lot between samples or specifications even when predictions don't, which you'll meet again with multicollinearity (3.9).

Q21 **A** · lesson 1.5

Errors are $1, 0, 3$; squares $1, 0, 9$; their mean is $10/3$. $4/3$ is the absolute-loss risk, 10 is the sum rather than the mean, and 1.83 is its square root (the RMSE).

Q22 **A** · lesson 1.5

Absolute errors are $1, 0, 3$, mean $4/3$. 4 forgets to divide by n; 3 is the largest single error; $10/3$ is the squared-loss risk.

Q23 **D** · lesson 1.5

$R(f) = \mathbb{E}[L(Y, f(X))]$ averages over the whole population, which we never observe. A test set gives an *estimate* of it, not the value itself.

Q24 **B** · lesson 1.5

0.5 is empirical risk on data the model was fitted to; 4.0 estimates population risk. A big gap means the model fitted quirks of its training sample. Underfitting would show high error on *both*.

Q25 **C** · lesson 1.5

Training loss is only evidence about population risk. A flexible model can drive it to zero by memorising noise, but zero training loss doesn't *prove* overfitting either (the data could be noise-free). Only held-out data can tell you.

Q26 **A** · lesson 2.1

The error is $11 - 8 = 3$. Squared loss $3^2 = 9$; absolute loss $|3| = 3$. Losses are never negative, and "6" doubles the error instead of squaring it.

Q27 **A** · lesson 2.1

For $y = 0$ the loss is $-\ln(1 - p) = -\ln 0.1 = 2.303$: the model gave only 0.1 to what actually happened. 0.105 is what you'd get if the true label were 1.

Q28 **A** · lesson 2.1

Squared loss squares big errors, so a few mansions can drag the whole fit towards them. Absolute loss weights an error of 20 as 20, not 400. 0–1 and cross-entropy are for classification, not prices.

Q29 **A** · lesson 2.1

Loss is $-\ln p$ for the true class: $-\ln 0.8 = 0.223$, $-\ln 0.2 = 1.609$. The log makes the penalty grow sharply as the probability given to the truth falls; it isn't a simple ratio of probabilities. Every case contributes, not just misclassified ones.

Q30 **D** · lesson 2.1

Fitting needs a smooth, well-behaved loss; evaluation should measure what matters in practice (costs, AUC, sensitivity). They're different quantities, and using different ones for the two jobs is normal.

Q31 **D** · lesson 2.2

ERM can only use what it has: the training sample. It picks the member of the family $\mathcal{F}$ with the lowest average training loss. Population risk is the goal but can't be computed, and the test set must not be used for fitting.

Q32 **A** · lesson 2.2

OLS minimises the residual sum of squares among straight lines (or hyperplanes). Over *all* functions, squared loss could be driven to zero by interpolating the data, which is not what OLS does.

Q33 **D** · lesson 2.2

"arg min" returns the *argument* that minimises (the function itself), not the minimum value. $\min \hat{R}_n$ would be the number; $\arg\min$ is the model.

Q34 **D** · lesson 2.2

A degree-9 polynomial has 10 coefficients, enough to pass exactly through 10 points. ERM happily picks it: training loss is zero. It nearly always overfits, which is why ERM must be paired with held-out evaluation or regularisation.

Q35 **D** · lesson 2.2

Ridge adds $\lambda \sum \beta_j^2$ to the RSS, so it's regularised ERM: still minimising training loss, but with a price on complexity. Absolute loss without a penalty would be least absolute deviations, a different method.

Q36 **B** · lesson 2.3

A big gap between very low training error and much higher validation error is the overfitting signature. Underfitting would show high error on *both*. Leakage tends to make validation look too *good*, not bad.

Q37 **D** · lesson 2.3

Both errors are high and similar, so the model can't capture the pattern even on data it was fitted to. Regularisation would make it simpler still. A small gap is only good news when both errors are low.

Q38 **D** · lesson 2.3

A degree-10 polynomial can always fit the training data at least as well as a straight line (it includes the line as a special case), so training error systematically rewards complexity. Compare on validation data or by cross-validation.

Q39 **D** · lesson 2.3

Overfitting means too much flexibility for the data available. More polynomial terms add flexibility. The other three reduce variance: more data, a penalty on complexity, or fewer features.

Q40 **A** · lesson 2.3

This is the U-shaped curve. Training error only goes down as flexibility grows. Error on new data improves while the model captures real structure, then worsens as it starts fitting noise. The bottom of the U is the sweet spot.

Q41 **C** · lesson 2.4

With more representative data, training performance reflects real performance more closely and the model has less room to memorise. More parameters and less regularisation widen the gap; noisier labels give more noise to fit.

Q42 **B** · lesson 2.4

Possible causes include overfitting to the validation set, distribution shift or a preprocessing mismatch. Tuning against the same test set would leak it, and reporting 96% ignores the evidence. A fresh, untouched test set gives the honest final number.

Q43 **D** · lesson 2.4

A small gap only says training performance is representative of new data. Here both are poor, so the model is too simple. Good generalisation means low error on new data, not just a small gap.

Q44 **D** · lesson 2.4

With more data it's harder to memorise everything (training error creeps up) and the fit captures real structure better (validation error falls). Neither reaches zero: noise sets a floor.

Q45 **C** · lesson 2.4

Models assume future data look like training data. When customer behaviour, prices or sensors change, that assumption fails gradually: distribution shift. Monitoring and retraining on recent data address it; CV settings at launch don't.

Q46 **C** · lesson 2.5

Noise is randomness in Y that the predictors can't explain. Bias falls with more flexible models and variance with more data or regularisation, but σ^2 is a floor under every model.

Q47 **B** · lesson 2.5

Flexible models can follow the true shape more closely (less bias) but react more to the particular sample they were trained on (more variance). That tension is the trade-off.

Q48 **A** · lesson 2.5

Stable across samples means low variance; consistently wrong in the same direction means high bias. It's like darts tightly grouped but away from the bullseye.

Q49 **A** · lesson 2.5

More data makes the fitted function change less between samples, which reduces variance. It doesn't change what the model family can represent (bias) or the randomness in Y (noise).

Q50 **D** · lesson 2.5

It is the variance of $\hat{f}(x_0)$ over repeated training samples, a property of the fitting procedure. The spread of the data itself is a different quantity, and confusing the two is a common slip.

Q51 **B** · lesson 2.6

Degree is a hyperparameter, so choosing it is a model-selection decision. Training error always prefers higher degrees, and using the test set would make the final score optimistic.

Q52 **B** · lesson 2.6

20% of 1,000 is 200. Training gets 600 and validation 200; 400 would be validation and test combined.

Q53 **C** · lesson 2.6

No model saw test rows, but the *choice* did. Among 40 tries, the winner partly won by luck on those particular rows, so its score overstates what to expect on new data. Select on validation, then test once.

Q54 **C** · lesson 2.6

With a rare class, a random split can leave a validation or test set with very few positives, making its scores unreliable. Stratifying preserves the class proportions in each subset; it doesn't rebalance them.

Q55 **B** · lesson 2.6

Once all decisions are made, using more data for the final fit usually helps, and the untouched test set gives one honest estimate. Any selection or tuning on the test set turns it into validation data.

Q56 **C** · lesson 2.7

The CV estimate is the mean of the fold errors: $25/5 = 5$. 25 is the sum; 7 is the worst fold.

Q57 **A** · lesson 2.7

Each candidate is fitted once per fold: $3 \times 10 = 30$. That cost is the price of a more stable comparison than a single split.

Q58 **C** · lesson 2.7

That's time leakage: the model learns from the future it's meant to predict. Use forward-chaining folds that always validate on the period after the training window. Every observation still appears in exactly one validation fold in ordinary K-fold.

Q59 **B** · lesson 2.7

A 0.02 difference in means is small next to B's huge spread, which signals instability (high variance). Consistent fold errors are a good sign, not overfitting. Look at the spread as well as the mean.

Q60 **B** · lesson 2.7

The mean and SD are statistics learned from data, including the rows that later act as validation folds. Each round should compute them from its own training folds. Using the test set would be a worse leak.

Q61 **C** · lesson 2.8

The median was computed using the test rows, so test information entered the pipeline. Split first, compute the median on training data only, and use that value for the test set too.

Q62 **B** · lesson 2.8

Calling the cancellation line is a consequence of already deciding to cancel. It makes the model look superb in testing but arrives too late to be useful. Use only information available when the prediction is made.

Q63 **C** · lesson 2.8

$z = (1800 - 1400)/400 = +1.0$. Always use the training statistics for test data. -0.4 is what you'd get by wrongly using test-set statistics ($\mu = 2000$, $\sigma = 500$); $+0.6$ uses the full-data statistics.

Q64 **C** · lesson 2.8

Scaling is learned from training data, selection uses only training data, and the test set is used once at the end. Comparing many models by CV is fine; the problem is only selecting on the test set. (Strictly, scaling should also be refitted inside each CV fold.)

Q65 **B** · lesson 2.8

Photos of the same patient share features, so the model can recognise the person rather than the condition, inflating the test score. A group split keeps every patient on one side. Deleting photos wastes data and isn't needed.

Q66 **A** · lesson 2.8

With 20,000 genes, many correlate with the label by chance. Choosing them using every row's label means the validation folds influenced the features, so CV accuracy is inflated. CV only protects steps that happen inside the folds.

Q67 **D** · lesson 3.1

Demand in MWh is a quantity, so it's regression. Spam and blood type are categories (classification), and grouping without labels is clustering.

Q68 **C** · lesson 3.1

ε is the unobservable noise in the true relationship. Residuals $y_i - \hat{y}_i$ are our *estimates* of it from a fitted model, which is a common mix-up.

Q69 **A** · lesson 3.1

Residual = observed − fitted = 250 − 235 = +15. A positive residual means the actual value is above the prediction, so the model predicted too low.

Q70 **B** · lesson 3.1

We never see the true errors, because we don't know the true f. Residuals use the fitted model instead, so they approximate the errors. Even a correct model has non-zero residuals because of the noise.

Q71 **C** · lesson 3.1

$30 + 2.6 \times 100 = 290$. £260k forgets the intercept; £2,630k multiplies the intercept too; £32.6k drops the factor of 100.

Q72 **C** · lesson 3.2

"Linear" refers to the unknown coefficients. X^2 is just another column computed from the data, so OLS fits it exactly as before, even though the fitted curve bends.

Q73 **C** · lesson 3.2

In $\beta_0 e^{\beta_1 X}$ the coefficient β_1 sits inside an exponential, so the model isn't linear in its parameters. The others only transform the inputs ($\log X$, $X_1 X_2$, X^3), which is allowed.

Q74 **B** · lesson 3.2

With an interaction, the slope for X_1 depends on the value of X_2: differentiate (or compare X_1 and $X_1 + 1$) to get $\beta_1 + \beta_3 X_2$. Plain β_1 is the effect only when $X_2 = 0$.

Q75 **B** · lesson 3.2

$\beta_0 + \beta_1 X_1 + \beta_2 X_2$ is a flat plane over the (X_1, X_2) floor. With more predictors it's a hyperplane, which can't be drawn but works the same way.

Q76 **A** · lesson 3.2

Adding a squared term lets OLS fit a parabola while staying linear in the coefficients. Logistic regression is for binary outcomes, and K-means doesn't fit relationships at all.

Q77 **A** · lesson 3.3

One row per observation and one column per coefficient: four predictors plus the column of ones for the intercept. 5×5 is the size of $X^\top X$.

Q78 **A** · lesson 3.3

The formula needs the inverse of $X^\top X$. It doesn't exist with perfect multicollinearity or with more coefficients than observations. Units, intercepts and residual signs are irrelevant.

Q79 **D** · lesson 3.3

OLS minimises $\sum (y_i - \hat{y}_i)^2$. The plain sum of residuals is always zero with an intercept, so it can't pick a line; absolute residuals define a different method.

Q80 **D** · lesson 3.3

Entries are $n = 3$, $\sum x = 6$ and $\sum x^2 = 1 + 4 + 9 = 14$. (With $X^\top y = (6, 13)^\top$ the solution is $\hat{\beta} = (1, 0.5)$.) 12 would be $(\sum x)^2/3$, not $\sum x^2$.

Q81 **C** · lesson 3.3

One column is an exact multiple of the other (perfect multicollinearity), so infinitely many coefficient pairs give the same fit. Software either fails or silently drops one column.

Q82 **B** · lesson 3.4

$\bar{x} = 1.5$, $\bar{y} = 4$, $S_{xy} = 10$, $S_{xx} = 5$, so $\hat{\beta}_1 = 2$ and $\hat{\beta}_0 = 4 - 2(1.5) = 1$. The points lie exactly on $y = 1 + 2x$.

Q83 **D** · lesson 3.4

$\bar{x} = 2$, $\bar{y} = 7/3$. $S_{xy} = (-1)(-1/3) + 0 + (1)(5/3) = 2$ and $S_{xx} = 2$, so the slope is $2/2 = 1$. A common slip is to divide by n in only one of the two sums; both S_{xy} and S_{xx} are plain sums.

Q84 **B** · lesson 3.4

With no variation in x there's no information about how y changes with x, and the denominator $\sum (x_i - \bar{x})^2$ is zero. With little variation, the slope is computable but very unstable.

Q85 **B** · lesson 3.4

This follows from the normal equation for β_0. It's a handy check on hand calculations: in dataset B the residuals $(-0.8, 0.6, 1.0, -0.6, -0.2)$ sum to 0.

Q86 **B** · lesson 3.4

An OLS line with an intercept always passes through $(\bar{x}, \bar{y}) = (3, 4)$: check $2.2 + 0.6 \times 3 = 4$. It need not pass through any data point or the origin.

Q87 **B** · lesson 3.5

$5 + 2(3) - 4(1) = 7$. 15 ignores the minus sign; 11 forgets the x_2 term; 3 drops the intercept.

Q88 **D** · lesson 3.5

Coefficient size depends on units: 0.05 per mm is 50 per metre, larger than A's 40. Standardise predictors (per-SD effects) before comparing, and consider their uncertainty too.

Q89 **B** · lesson 3.5

A coefficient answers "what happens to Y when this predictor changes and the other included ones don't?" Change which predictors are included and you change the question, so the answer (even its sign) can change.

Q90 **B** · lesson 3.5

Observational regression shows association, holding the included predictors fixed. Confounders (activity, diet) could drive both. Causal claims, like "cutting TV reduces weight", need experiments or strong assumptions.

Q91 **D** · lesson 3.5

The intercept is the predicted value when all predictors are zero. A 0 m² house doesn't exist, so −50 is just where the line crosses the axis. The model can still predict well in the range of real houses.

Q92 **C** · lesson 3.6

A systematic curve means the model consistently over- then under-predicts: structure is left in the residuals. Adding X^2 or other terms may help. Non-constant variance shows as a fan, not a U.

Q93 **A** · lesson 3.6

A fan shape means the noise variance isn't constant. The usual standard-error formula assumes it is, so p-values and confidence intervals can't be trusted. Transforming Y (e.g. log) or weighted least squares can help.

Q94 **D** · lesson 3.6

Runs of same-sign residuals over time mean one month's error predicts the next. OLS assumes independent errors, so its standard errors are off. Time-series methods are designed for this.

Q95 **C** · lesson 3.6

Diagnostics can reveal problems; they can't certify a model. Missing variables or poor out-of-sample performance can hide behind a tidy plot. Check held-out error as well.

Q96 **C** · lesson 3.6

Minimising RSS is pure algebra and works on any data. The assumptions (mean zero, constant variance, independence, roughly normal) are what justify the uncertainty measures built on top of $\hat{\beta}$.

Q97 **B** · lesson 3.7

$R^2 = 1 - 30/120 = 0.75$. 0.25 is RSS/TSS itself (the share of error left over); 4 is TSS/RSS.

Q98 **B** · lesson 3.7

$1 - \frac{30/20}{120/24} = 1 - \frac{1.5}{5} = 0.70$. Residual df is $n - p - 1 = 20$; total df is $n - 1 = 24$. 0.75 is the unadjusted R^2.

Q99 **A** · lesson 3.7

MAE $= (2 + 1 + 3)/3 = 2$. MSE $= (4 + 1 + 9)/3 = 4.67$, so RMSE $= \sqrt{4.67} = 2.16$. 1.33 is the signed average error, and 4.67 is the MSE, not its root.

Q100 **B** · lesson 3.7

Out-of-sample R^2 compares the model with predicting $\bar{y}_{\text{train}}$ for everyone. Negative means the model's squared error is larger than that baseline's: worse than useless on new data. Only in-sample R^2 (with an intercept) is bounded below by 0.

Q101 **D** · lesson 3.7

In-sample R^2 can never decrease when a predictor is added: OLS finds some small chance fit. Adjusted R^2 charges for the extra parameter, so a useless predictor usually lowers it.

Q102 **A** · lesson 3.8

$t = 1.5/0.5 = 3$. CI $= 1.5 \pm 2 \times 0.5 = (0.5, 2.5)$, which excludes 0, so it's significant at 5%. $(1, 2)$ uses one SE instead of two.

Q103 **C** · lesson 3.8

$\hat{\sigma}^2 = \mathrm{RSS}/(n-p-1) = 1.8/2 = 0.9$, so $\mathrm{SE} = \sqrt{0.9/5} = \sqrt{0.18} = 0.424$. 0.18 forgets the square root; 0.60 and 0.36 use RSS instead of $\hat{\sigma}^2$.

Q104 **A** · lesson 3.8

A p-value is calculated *assuming* H_0 (and the model's assumptions). It is not the probability that H_0 is true, and it says nothing about effect size.

Q105 **B** · lesson 3.8

The interval includes 0, so we fail to reject H_0. But it's wide: the data are consistent with a small negative effect or a sizeable positive one. "Not significant" isn't "no effect".

Q106 **C** · lesson 3.8

Residual df $= n - p - 1 = 40 - 3 - 1 = 36$: one is lost for each estimated coefficient, including the intercept. 37 forgets the intercept; 39 is $n - 1$.

Q107 **D** · lesson 3.9

$\mathrm{VIF} = 1/(1 - 0.95) = 20$: its coefficient's variance is 20 times what it would be with uncorrelated predictors. 1.05 is $1/0.95$.

Q108 **D** · lesson 3.9

VIF inflates the **variance**; the standard error is its square root, so it's inflated by $\sqrt{16} = 4$.

Q109 **C** · lesson 3.9

With near-duplicate predictors the model can't tell which deserves the credit, so each coefficient is poorly determined. Insignificance here means "not needed alongside the other", not "irrelevant".

Q110 **A** · lesson 3.9

Like two people carrying one sofa, the total contribution is clear but how it's divided between them isn't. Predictions depend on the total, interpretation on the split.

Q111 **A** · lesson 3.9

Duplicating information pushes $X^\top X$ even closer to singular. Combining, regularising (ridge shares the credit) or getting more varied data all reduce the problem.

Q112 **D** · lesson 4.1

A straight line is unbounded, so its "probabilities" can be negative or exceed 1, and the constant-variance noise assumption fails for 0/1 data. It *can* be fitted with many predictors; the outputs just don't make sense.

Q113 **D** · lesson 4.1

Regression would treat category 3 as "further" from 1 than 2 is, which means nothing for unordered categories. Recoding to 0, 1, 2 has the same flaw. Use a classifier.

Q114 **D** · lesson 4.1

$\ln \frac{P}{1-P} = \beta_0 + \sum \beta_j X_j$. The probability itself is an S-shaped (sigmoid) function of the predictors, not a linear one.

Q115 **D** · lesson 4.1

Logistic regression gives a probability; a decision rule turns it into a class. The threshold is a choice (0.5 by default, lower when misses are costly).

Q116 **B** · lesson 4.1

Classifiers are judged by how they handle each class (sensitivity, specificity, precision) and how well they rank (AUC). R^2 and RMSE are regression measures, and VIFs describe predictors, not performance.

Q117 **C** · lesson 4.2

Odds $= 0.2/0.8 = 0.25$, and $\ln 0.25 = -1.386$. Odds 4 is the odds *against*; $-1.61 = \ln 0.2$ skips the odds step.

Q118 **A** · lesson 4.2

Log-odds of 1.1 means odds of $e^{1.1} \approx 3$, i.e. three to one in favour, so $p = 3/(1+3) = 0.75$. 0.33 wrongly treats the odds as $1/p$, and 0.25 is the probability of the other class.

Q119 **A** · lesson 4.2

$p = 1/(1+e^{0.6}) = 1/2.82 = 0.354$. A negative score always means a probability below 0.5. 0.646 is the answer for $z = +0.6$, and 0.549 is the odds $e^{-0.6}$ rather than the probability.

Q120 **D** · lesson 4.2

$p = \text{odds}/(1+\text{odds}) = 3/4$. "3 to 1 in favour" means 3 successes for every failure. 0.33 confuses odds with $1/p$.

Q121 **D** · lesson 4.2

Log-odds are symmetric around $p = 0.5$: $\ln(0.1/0.9) = -\ln 9 = -2.2$. Swapping p and $1-p$ just flips the sign.

Q122 **A** · lesson 4.3

$z = -3 + 2 + 1.2 = 0.2$, so $p = 1/(1+e^{-0.2}) = 1/1.82 = 0.55$; predicted class 1 at threshold 0.5. 0.20 is z itself, not a probability.

Q123 **A** · lesson 4.3

Exponentiated coefficients are odds ratios: they multiply the odds, holding x_1 fixed. The log-odds rise by 1.2 (not 3.32), and "3.32 times as likely" confuses odds with probability.

Q124 **A** · lesson 4.3

Coefficients act additively on the log-odds scale (and multiply the odds by $e^{0.7} \approx 2$). The change in probability depends on where you start.

Q125 **D** · lesson 4.3

Odds go 0.111 → 0.248 (p 0.10 → 0.20) and 1 → 2.23 (p 0.50 → 0.69). The same odds ratio gives +0.10 versus +0.19 in probability, because the sigmoid is steepest near 0.5.

Q126 **C** · lesson 4.3

$p > 0.5$ exactly when $z > 0$, so the boundary is $z = 0$: linear in x. That's why logistic regression is a linear classifier even though its probability curve is S-shaped.

Q127 **B** · lesson 4.4

Setting the gradient of the log-likelihood to zero gives equations with β inside $1/(1+e^{-x^\top \beta})$, which can't be solved algebraically. The problem is convex (one optimum), so numerical methods find it reliably.

Q128 **C** · lesson 4.4

Probabilities of what happened: 0.7, 0.4 and $1 - 0.2 = 0.8$. $\ln 0.7 + \ln 0.4 + \ln 0.8 = -0.357 - 0.916 - 0.223 = -1.50$. −2.10 uses $\ln 0.2$ for the third case instead of $\ln 0.8$.

Q129 **D** · lesson 4.4

The log is strictly increasing, so the same β maximises both; sums of logs are just easier to differentiate and avoid tiny products. It changes neither the answer nor overfitting.

Q130 **A** · lesson 4.4

$z = \hat{\beta}/\text{SE} = -0.9/0.45 = -2$. Significance depends on $|z|$, so negative coefficients count too. -0.5 divides the wrong way round, and being one SE from zero ($|z| > 1$) is not enough.

Q131 **B** · lesson 4.4

With a single shared probability, the MLE is the proportion of ones: 3/5. Setting the derivative of $3\ln p + 2\ln(1-p)$ to zero gives the same answer. 0.4 is the proportion of zeros.

Q132 **B** · lesson 4.5

Loss is $-\ln p$ for the true class: $-\ln 0.8 = 0.223$, $-\ln 0.2 = 1.609$. Low probability on the truth is punished heavily; the loss isn't the probability itself.

Q133 **C** · lesson 4.5

For $y = 0$ the loss is $-\ln(1-p) = -\ln 0.9 = 0.105$: the model gave 0.9 to the truth, so the loss is small. 2.303 would be right if the true label were 1.

Q134 **B** · lesson 4.5

Losses are $-\ln 0.9 = 0.105$ and $-\ln 0.8 = 0.223$. Their mean is $0.328/2 = 0.164$; 0.328 is the sum. 1.609 wrongly uses $-\ln 0.2$ for the second case.

Q135 **A** · lesson 4.5

Maximising the log-likelihood is the same as minimising cross-entropy. Hinge loss belongs to SVMs, and 0–1 loss is too jumpy to optimise directly.

Q136 **A** · lesson 4.5

Accuracy only checks which side of the threshold each prediction lands on. Cross-entropy scores the probabilities themselves, which matters for ranking, costs and calibration. It doesn't depend on any threshold.

Q137 **D** · lesson 4.6

Predicted positive (fraud), actually negative (legitimate): a false alarm. A false negative would be real fraud that got through.

Q138 **B** · lesson 4.6

Predicted negative, actually positive: a miss. In screening, misses are usually the costlier error.

Q139 **D** · lesson 4.6

Flagged (score > 0.5): 0.8 (+, TP) and 0.6 (−, FP). Not flagged: 0.3 (+, FN) and 0.2 (−, TN). So one of each.

Q140 **C** · lesson 4.6

Missing an important real email typically costs more than seeing one extra spam message, so spam filters favour fewer false positives. Which error is worse always depends on the use case.

Q141 **D** · lesson 4.6

A Type I error is rejecting a true null hypothesis (raising an alarm when nothing is there): a false positive. A Type II error is missing a real effect: a false negative.

Q142 **A** · lesson 4.7

Sensitivity $= 30/(30+20) = 0.60$; specificity $= 140/(140+10) = 0.933$. 0.75 is the precision $30/40$, and 0.85 the accuracy $170/200$.

Q143 **B** · lesson 4.7

Precision $= TP/(TP+FP) = 30/40 = 0.75$: of the 40 cases flagged, 30 were truly positive. 0.60 is the sensitivity (it divides by the actual positives instead).

Q144 **B** · lesson 4.7

Specificity is the share of actual negatives correctly cleared. $TP/(TP+FN)$ is sensitivity, $TP/(TP+FP)$ precision, and $FP/(FP+TN)$ the false positive rate ($1-$ specificity).

Q145 **C** · lesson 4.7

Sensitivity 0.95: few misses among the sick. Specificity 0.60: 40% of healthy people get a false positive. Neither number is overall accuracy or precision.

Q146 **D** · lesson 4.7

Per 1,000 people: 10 are sick and 9 test positive; 990 are healthy and 99 test positive. Precision $= 9/108 \approx 8\%$. With a rare condition, false positives swamp true ones even for a good test.

Q147 **C** · lesson 4.8

A lower threshold flags more cases: fewer positives are missed (sensitivity up) but more negatives are flagged (specificity down).

Q148 **C** · lesson 4.8

The threshold slides along a fixed trade-off (the ROC curve). Improving both needs a better model whose score distributions for the two classes overlap less.

Q149 **B** · lesson 4.8

Every positive is caught (no misses) and every negative is a false alarm. Accuracy then equals the share of positives, which is rarely useful.

Q150 **D** · lesson 4.8

Perfect sensitivity is easy if you flag everything. The cost appears as collapsed specificity and precision. Always report both sides of the trade-off. AUC doesn't depend on the threshold.

Q151 **C** · lesson 4.8

Moving the threshold trades one for the other. Only a better model, one whose scores separate the classes more cleanly, shifts the whole trade-off upward. Altering the test set changes measurements, not the model.

Q152 **A** · lesson 4.9

At 0.5: $5 \times 1 + 10 \times 9 = 95$. At 0.2: $25 \times 1 + 3 \times 9 = 52$. 15 and 28 just count errors without weighting them by cost.

Q153 **B** · lesson 4.9

Alert when $p > c_{FP}/(c_{FP} + c_{FN}) = 500/2000 = 0.25$. Misses cost more, so the threshold moves below 0.5. 0.75 swaps the two costs; 0.33 is $500/1500$.

Q154 **C** · lesson 4.9

$c_{FP}/(c_{FP} + c_{FN}) = 1/21 \approx 0.05$. Flag a case even when the probability is low, and monitor the resulting volume of false alarms.

Q155 **D** · lesson 4.9

The threshold is a hyperparameter. Choosing it on the test set makes the test cost a best-case number. Tune on validation data or by CV, then measure once on the test set.

Q156 **B** · lesson 4.9

Costly misses mean you'd rather raise more false alarms than miss positives, so you flag at lower probabilities. The theory gives $p > c_{FP}/(c_{FP} + c_{FN})$, which is below 0.5 here.

Q157 **B** · lesson 4.10

Each point is one threshold: sensitivity on the vertical axis, $1-$ specificity on the horizontal. The top-left corner is perfect, the diagonal is random guessing.

Q158 **C** · lesson 4.10

6 positive–negative pairs. 0.8 beats both negatives (2); 0.5 beats 0.3 only (1); 0.4 beats 0.3 only (1). That's 4 out of 6.

Q159 **C** · lesson 4.10

TPR $= 3/4 = 0.75$; FPR $= 2/6 = 0.33$. 0.60 and 0.40 divide by the 5 flagged cases instead of by the actual class sizes; 0.67 is the specificity.

Q160 **C** · lesson 4.10

AUC is the probability a random positive outscores a random negative: a ranking measure. It isn't accuracy, it doesn't pick a threshold, and 0.55 is weak but slightly better than chance.

Q161 **B** · lesson 4.10

AUC is the area under the whole ROC curve, so it doesn't depend on any one cut-off. To choose an operating threshold you still need costs or a target sensitivity.

Q162 **B** · lesson 4.11

The majority-class baseline already scores 99.5% while catching zero fraud. Report sensitivity, precision and a confusion matrix, and compare against that baseline.

Q163 **D** · lesson 4.11

Duplicated minority cases end up on both sides of the split, so the model is tested on records it has seen. Oversample only the training part, after splitting (inside each CV fold).

Q164 **B** · lesson 4.11

Validation should reflect the real world the model will face; otherwise precision and accuracy are meaningless. Only the training data is rebalanced, to help the model learn the rare class.

Q165 **B** · lesson 4.11

Predicting "negative" for everyone already gets 95%. The model adds one point of accuracy while catching only a fifth of the positives. Chance here is 95%, not 50%.

Q166 **A** · lesson 4.11

Class weights change how much each mistake counts during fitting, an alternative to oversampling that doesn't duplicate records. It doesn't delete data or touch the validation set.

Q167 **A** · lesson 5.1

The regularised objective is training error plus λ times a complexity penalty. λ sets the exchange rate between fit and simplicity. Noise can't be reduced by any model.

Q168 **D** · lesson 5.1

With $p > n$, OLS has no unique solution and can fit the training data perfectly, noise and all. The other datasets have many observations per predictor, so plain OLS is stable.

Q169 **B** · lesson 5.1

Simpler models have equal or *higher* training error; their benefit shows up on new data. Interpretability, lower data-collection cost and robustness are genuine advantages.

Q170 **C** · lesson 5.1

With no penalty, ridge and LASSO both minimise plain RSS, which is OLS. The mean-only model is the opposite extreme, as $\lambda \to \infty$.

Q171 **A** · lesson 5.1

With $p + 1 > n$ there are more unknowns than equations: many coefficient vectors give zero training error, and $(X^\top X)^{-1}$ doesn't exist. Regularisation picks one sensible solution.

Q172 **B** · lesson 5.2

More clusters let K-means fit the data more finely (lower J, more complexity). Larger K in K-NN averages over more neighbours (smoother), larger λ shrinks more, and smaller C widens the SVM margin: all simpler.

Q173 **B** · lesson 5.2

OLS can always set a new coefficient to zero, so adding a predictor can't hurt the training fit. Use validation or CV error, or a size-penalised criterion such as adjusted R^2 or AIC/BIC.

Q174 **B** · lesson 5.2

The surviving variables were chosen because they looked significant on this data; reporting their p-values as if no choice was made overstates certainty. It also targets in-sample significance, not prediction. Use CV-based selection or LASSO.

Q175 **C** · lesson 5.2

Feature selection is a learning step. Doing it on all the data lets rows that later act as validation influence which features are used, inflating CV scores. Put it inside each training fold.

Q176 **D** · lesson 5.2

In-sample error always falls with more variables. Once added variables mostly fit noise, out-of-sample error rises: the classic U-shape. The best model sits at its minimum.

Q177 **A** · lesson 5.3

Every subset of 8 predictors: $2^8 = 256$. 37 is the forward-stepwise count; 64 is 8^2.

Q178 **A** · lesson 5.3

Start with the empty model, then try 8 additions, then 7, and so on down to 1: $1 + \frac{8\times 9}{2} = 37$ fits. 36 forgets the empty model; 9 counts only the final sequence $M_0, \ldots, M_8$, not the candidates tried at each step.

Q179 **A** · lesson 5.3

Training RSS always picks M_p, the largest model. You need an estimate of performance on new data (CV) or a criterion such as adjusted R^2, AIC or BIC.

Q180 **D** · lesson 5.3

Stepwise commits to the best single variable and never removes it. If A and C work well together but neither is the best alone, the greedy path misses them. That's the price of fitting far fewer models.

Q181 **B** · lesson 5.3

Subset selection penalises $\|\beta\|_0$, how many coefficients are non-zero. Squares give ridge (L2) and absolute values give LASSO (L1).

Q182 **A** · lesson 5.4

With one centred predictor, ridge adds λ to the denominator: slope $= \sum xy / (\sum x^2 + \lambda) = 6/(3+3) = 1$. 2 is the unpenalised OLS slope ($\lambda = 0$), and 0.5 is what you get at $\lambda = 9$.

Q183 **C** · lesson 5.4

$X^\top X + 2I = 4I$, so $\hat\beta = (6, -2)/4 = (1.5, -0.5)$. $(3, -1)$ is the OLS solution; $(2, -0.67)$ divides by 3, adding $\lambda/2$.

Q184 **A** · lesson 5.4

$\hat\beta = (X^\top X + \lambda I)^{-1} X^\top y$. Ridge shrinks but never zeroes coefficients, its bias *rises* with λ, and it works even when $p > n$.

Q185 **C** · lesson 5.4

$X^\top X$ has non-negative eigenvalues; zero ones make it singular. $X^\top X + \lambda I$ adds λ to each, so all are positive and the matrix is invertible, even with perfect collinearity.

Q186 **C** · lesson 5.4

An infinite penalty forces all slopes towards zero. The intercept isn't penalised, so predictions approach the mean. Ridge never reaches exact zeros at finite λ; that's LASSO.

Q187 **C** · lesson 5.5

Soft thresholding subtracts 1 from each magnitude and sets anything below 1 to zero: $4 \to 3$, $|-0.8| < 1 \to 0$, $1.2 \to 0.2$. $(2, -0.4, 0.6)$ is ridge with $\lambda = 1$.

Q188 **C** · lesson 5.5

Ridge divides every coefficient by $1 + \lambda = 2$. Nothing becomes exactly zero, unlike LASSO's $(3, 0, 0.2)$.

Q189 **A** · lesson 5.5

LASSO drops irrelevant predictors (exact zeros), leaving a small readable model. Ridge would keep all 500 with small coefficients. LASSO handles many predictors but has no closed form.

Q190 **C** · lesson 5.5

A bigger penalty pushes more coefficients to exactly zero. Training error rises, variance falls and bias rises, the same direction as ridge.

Q191 **B** · lesson 5.5

With near-duplicates, LASSO usually picks one and drops the other, and which one can flip with small data changes. Ridge would instead share the credit evenly. So a dropped variable isn't necessarily irrelevant.

Q192 **D** · lesson 5.6

$\lambda = 0$ is OLS, which by definition minimises RSS on the training data. Any penalty trades some training fit for simplicity; CV picks the best *test* error, not training error.

Q193 **A** · lesson 5.6

No penalty means plain least squares. The mean-only model is the $\lambda \to \infty$ limit.

Q194 **A** · lesson 5.6

Training error rising slightly with λ is expected. The big drop in test error is the point of regularisation: less variance, better generalisation. Underfitting would show high error on both.

Q195 **A** · lesson 5.6

Ridge minimises RSS + penalty, so it beats OLS on that objective. OLS minimises plain RSS, so it beats ridge on training RSS. Read questions carefully: "training error" means RSS.

Q196 **A** · lesson 5.6

The longer a coefficient survives as λ grows, the more it contributes to the fit. Every coefficient eventually reaches zero, so that alone means nothing.

Q197 **A** · lesson 5.7

That's LASSO's built-in variable selection. Ridge has the closed form, and both need standardised predictors because both penalties depend on coefficient sizes.

Q198 **B** · lesson 5.7

The L1 constraint $\sum |\beta_j| \leq s$ is a diamond with corners on the axes; the expanding RSS ellipse usually hits a corner first, where a coefficient is zero. Ridge's disc is round, with no corners.

Q199 **C** · lesson 5.7

Ridge keeps all of them with stable, evenly spread coefficients. LASSO would drop most of them arbitrarily; OLS would give huge, unstable coefficients; best subset over 30 predictors means a billion models.

Q200 **D** · lesson 5.7

Like subset selection it produces sparse models; like ridge it is a continuous optimisation that's fast to solve. It gets most of the interpretability without the 2^p search.

Q201 **B** · lesson 5.7

Elastic net combines $\sum |\beta_j|$ and $\sum \beta_j^2$. With all weight on L1 it is LASSO, all on L2 it is ridge; in between it selects variables while handling correlated groups more stably.

Q202 **C** · lesson 5.8

Non-zeros: 3. Sum of absolute values: $2 + 1 + 0 + 3 = 6$. Sum of squares: $4 + 1 + 0 + 9 = 14$. "4" counts the zero; "4" for L1 forgets the absolute value of −1.

Q203 **D** · lesson 5.8

Ridge adds λ times the sum of squares: $20 + 3(1 + 4 + 0) = 35$. LASSO adds the sum of absolute values: $20 + 3(1 + 2 + 0) = 29$. L0 counts non-zero coefficients: $20 + 3 \times 2 = 26$. Swapping the ridge and LASSO norms is the usual slip.

Q204 **B** · lesson 5.8

$8 + 2(2 + 1 + 0.5) = 15$. 18.5 is the ridge objective $8 + 2(4 + 1 + 0.25)$; 11.5 forgets to multiply by λ.

Q205 **A** · lesson 5.8

A: $10 + 0.1(16 + 9) = 12.5$. B: $13 + 0.1(4 + 1) = 13.5$. With a small penalty the better-fitting model wins; 18 vs 35 is the result for $\lambda = 1$, where B wins.

Q206 **B** · lesson 5.8

Three coefficients are non-zero, so $12 + 2 \times 3 = 18$. 20 counts all four; 23 and 38.5 are the LASSO and ridge objectives.

Q207 **D** · lesson 5.9

λ is a hyperparameter, so tune it by CV and test once at the end. The lowest training error is always at $\lambda = 0$, and the test set must not drive choices.

Q208 **D** · lesson 5.9

Means and SDs are learned from data. Computed on everything, they include the rows that act as validation folds. Compute them inside each training fold.

Q209 **B** · lesson 5.9

CV error is noisy, so models within one SE of the best are statistically indistinguishable. Taking the largest such λ gives the simplest of them.

Q210 **D** · lesson 5.9

With many rows per predictor, OLS already has low variance, so accuracy barely changes. But shrinking a near-duplicate predictor (or dropping it with LASSO) gives steadier coefficients and a cheaper model. A 0.001 difference is noise.

Q211 **C** · lesson 5.9

A predictor in millimetres needs a coefficient 1,000 times smaller than one in metres, so it would be penalised far less. Standardising puts every predictor on the same footing. It doesn't remove correlation.

Q212 **D** · lesson 6.1

K-means uses no labels and minimises $J = \sum_k \sum_{i \in C_k} \|x_i - \mu_k\|^2$. Hinge loss and margins belong to SVMs; likelihood to logistic regression.

Q213 **A** · lesson 6.1

The mean minimises the sum of squared distances to a set of points, which is exactly what J measures. Using the median gives a different method (k-medians).

Q214 **B** · lesson 6.1

Same letter, different methods. K-means is unsupervised and splits data into K groups; K-NN is supervised and predicts from the K closest labelled points.

Q215 **A** · lesson 6.1

Clusters group similar patients; nothing forces them to line up with readmission. When the outcome is known, train a model to predict it. Clusters can still be useful as extra features.

Q216 **D** · lesson 6.1

The output is a partition of the data and the centre of each group. There's no target, so there are no predicted labels or probabilities in the supervised sense.

Q217 **C** · lesson 6.2

The coordinate differences are 3 and 4, so the Euclidean distance is $\sqrt{3^2 + 4^2} = \sqrt{25} = 5$. 7 is the Manhattan distance $3 + 4$, and 25 is the squared distance K-means compares when assigning points.

Q218 **A** · lesson 6.2

To μ_1: $4 + 4 = 8$. To μ_2: $16 + 4 = 20$. It joins the nearer centroid, cluster 1.

Q219 **C** · lesson 6.2

Manhattan distance adds the absolute differences along each axis: $|4 - 1| + |5 - 1| = 3 + 4 = 7$. 5 is the straight-line (Euclidean) distance, and 25 is its square.

Q220 **D** · lesson 6.2

Distance adds squared differences, so a £1,000 income gap outweighs a 30-visit gap. Standardising puts both features on comparable scales.

Q221 **A** · lesson 6.2

For non-negative numbers, $a < b \iff a^2 < b^2$, so the nearest centroid doesn't change and the square root can be skipped. Squaring actually *increases* the influence of far points.

Q222 **D** · lesson 6.3

1, 3 join the first centroid and 8, 10 the second. Means are 2 and 9; reassigning changes nothing. $J = (1 + 1) + (1 + 1) = 4$. $J = 2$ forgets to square or to include all points.

Q223 **B** · lesson 6.3

The left pair joins the first centroid, the right pair the second. Their means are $(0, 1)$ and $(6, 1)$, and the next assignment is unchanged, so it has converged.

Q224 **D** · lesson 6.3

Big drops up to $K = 3$, small gains afterwards. WCSS always falls as K grows, so its minimum can't choose K. The biggest single drop (100 to 45) doesn't mark where gains level off.

Q225 **B** · lesson 6.3

Each run is guaranteed to converge, but possibly to a poor solution. Keeping the run with the lowest J (or using k-means++ starts) reduces that risk.

Q226 **C** · lesson 6.3

With one cluster per point, every distance is zero. It's the same trap as choosing complexity by training error. Use the elbow, silhouette scores or usefulness.

Q227 **B** · lesson 6.4

Positive score means $+1$, negative means -1. A sigmoid would turn the score into a probability, which is logistic regression's approach.

Q228 **A** · lesson 6.4

Many lines separate the data; the SVM picks the one in the middle of the widest empty "street". A wide margin tolerates more noise in new data.

Q229 **A** · lesson 6.4

SVMs give a decision score whose sign is the class. Turning scores into probabilities needs an extra calibration step; logistic regression outputs probabilities directly.

Q230 **C** · lesson 6.4

With ± 1 labels, $y_i f(x_i) > 0$ means "correctly classified" for both classes, which makes the margin constraints and hinge loss neat. Logistic regression uses 0/1.

Q231 **D** · lesson 6.4

Points comfortably beyond the margin have zero hinge loss and don't influence the solution. Logistic regression, by contrast, is influenced by every point.

Q232 **B** · lesson 6.5

$\|w\| = \sqrt{9 + 16} = 5$, and width $= 2/\|w\| = 0.4$. 0.2 is $1/\|w\|$, the distance from the boundary to one margin line only.

Q233 **C** · lesson 6.5

$f = 6 + 4 - 5 = 5$, positive, so class $+1$. Distance $= |f|/\|w\| = 5/5 = 1$. 5 is the raw score, not a distance.

Q234 **C** · lesson 6.5

$4a + b = 1$ and $2a + b = -1$ give $a = 1$, $b = -3$. Width $= 2/1 = 2$, the distance between the support vectors. $a = 2$ would place them at ± 3, not ± 1.

Q235 **B** · lesson 6.5

Only support vectors determine the hard-margin solution. Points outside the margin have no influence, which also makes SVMs robust to points far from the boundary.

Q236 **D** · lesson 6.5

Margin width and $\|w\|$ are inversely related. Squaring and halving just make the optimisation smooth. The constraints, not this term, keep points on the right side.

Q237 **D** · lesson 6.6

$\max(0, 1 - 0.3) = 0.7$. The point is on the correct side but inside the margin, so it pays a partial penalty. It would be 0 only if $f(x) \geq 1$.

Q238 **C** · lesson 6.6

$yf(x) = -0.5$, so the loss is $\max(0, 1 + 0.5) = 1.5$: the point is misclassified. Losses above 1 always mean the wrong side of the boundary.

Q239 **C** · lesson 6.6

$1 + 2 \times 1.5 = 4$. 2.5 is the value with $C = 1$. A larger C makes violations count more.

Q240 **C** · lesson 6.6

$0 < \xi < 1$ means $0 < yf(x) < 1$: correctly classified but closer than the margin line. $\xi = 0$ is outside the margin, $\xi = 1$ on the boundary, $\xi > 1$ misclassified.

Q241 **A** · lesson 6.6

One overlapping point makes the hard-margin constraints unsatisfiable, and one outlier can force a very narrow margin. Slack variables allow violations at a price set by C.

Q242 **C** · lesson 6.7

C multiplies the training violations, so a larger C tolerates fewer of them and accepts a narrower margin: less bias, more variance. A small C does the opposite.

Q243 **D** · lesson 6.7

C weights the training error while λ weights the penalty, so a large C means weak regularisation, like a small λ. Getting the direction backwards is a common slip.

Q244 **D** · lesson 6.7

The SVM only needs inner products, and $K(x, x') = \phi(x)^\top \phi(x')$ gives them directly, even when ϕ is huge or infinite-dimensional. C and scaling still need attention.

Q245 **D** · lesson 6.7

Large γ makes each training point's influence reach only a tiny distance, so the boundary bends around individual points. Small γ gives smooth, possibly underfitting boundaries.

Q246 **A** · lesson 6.7

The boundary must be a closed curve. Linear methods (including logistic regression on the raw coordinates) draw straight lines, and K-means splits by distance to two centres. The RBF kernel can draw a circle.