AI Research Labs

Machine Learning Foundations · Practice test 5 · 45 questions

Regularisation and Model Selection

How to use this test. Answer all 45 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.

5.1 Why Regularisation is Needed

Q1 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

Q2 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

Q3 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

Q4 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

Q5 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

Q6 `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

Q7 `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

Q8 `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

Q9 `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

Q10 `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

Q11 `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

Q12 calculate How many models does forward stepwise selection fit with $p = 8$, including the empty model?

A 37
B 256
C 9
D 36

Q13 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 $R^2 > 0$

Q14 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

Q15 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)

Q16 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

Q17 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 $(2,\ -0.67)$

Q18 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

Q19 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

Q20 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)

Q21 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 $(4,\ 0,\ 1.2)$

Q22 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 $(3,\ -1.8,\ 0.2)$

Q23 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

Q24 `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

Q25 `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)

Q26 `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 $\lambda = 0$

Q27 `concept` Ridge regression with $\lambda = 0$ is...

A ordinary least squares

B undefined

C the mean-only (intercept) model

D LASSO regression

Q28 `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

Q29 `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

Q30 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

Q31 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

Q32 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

Q33 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

Q34 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

Q35 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

Q36 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

Q37 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

Q38 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

Q39 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

Q40 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

Q41 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

Q42 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

Q43 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

Q44 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

Q45 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

Answer key

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

Q1 **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.

Q2 **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.

Q3 **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.

Q4 **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 $\lambda \to \infty$.

Q5 **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.

Q6 **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.

Q7 **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.

Q8 **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.

Q9 **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.

Q10 **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.

Q11 **A** · lesson 5.3

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

Q12 **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.

Q13 **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.

Q14 **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.

Q15 **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).

Q16 **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 $\lambda = 9$.

Q17 **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 $\lambda/2$.

Q18 **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 $p > n$.

Q19 **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.

Q20 **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.

Q21 **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 $\lambda = 1$.

Q22 **C** · lesson 5.5

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

Q23 **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.

Q24 **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.

Q25 **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.

Q26 **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.

Q27 **A** · lesson 5.6

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

Q28 **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.

Q29 **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.

Q30 **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.

Q31 **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.

Q32 **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.

Q33 **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.

Q34 **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.

Q35 **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.

Q36 **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.

Q37 **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.

Q38 **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.

Q39 **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.

Q40 **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.

Q41 **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.

Q42 **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.

Q43 **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.

Q44 **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.

Q45 **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.