AI Research Labs

Machine Learning Foundations · Practice test 1 · 25 questions

Machine Learning Foundations

How to use this test. Answer all 25 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

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.