AI Research Labs

Machine Learning Foundations · Practice test 6 · 35 questions

Clustering and Support Vector Machines

How to use this test. Answer all 35 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.

6.1 K-means Clustering

Q1 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

Q2 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

Q3 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

Q4 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

Q5 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

Q6 `calculate` What is the Euclidean distance between $(1, 2)$ and $(4, 6)$?

- A 25
- B 7
- C 5
- D 3.5

Q7 `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)

Q8 `calculate` What is the Manhattan distance between $(1, 1)$ and $(4, 5)$?

- A 5
- B 25
- C 7
- D 12

Q9 `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

Q10 `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

Q11 `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; $J = 4$

Q12 `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 $(6,0)$

Q13 `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

Q14 `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

Q15 `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

Q16 `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 $w^\top x + b$

Q17 `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

Q18 `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

Q19 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 $y \in [0, 1]$

Q20 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

Q21 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

Q22 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

Q23 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

Q24 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

Q25 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

Q26 `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

Q27 `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

Q28 `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

Q29 `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

Q30 `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

Q31 `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

Q32 `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

Q33 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

Q34 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

Q35 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 **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.

Q2 **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).

Q3 **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.

Q4 **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.

Q5 **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.

Q6 **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.

Q7 **A** · lesson 6.2

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

Q8 **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.

Q9 **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.

Q10 **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.

Q11 **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.

Q12 **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.

Q13 **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.

Q14 **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.

Q15 **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.

Q16 **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.

Q17 **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.

Q18 **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.

Q19 **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.

Q20 **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.

Q21 **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.

Q22 **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.

Q23 **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.

Q24 **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.

Q25 **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.

Q26 **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 $f(x) \geq 1$.

Q27 **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.

Q28 **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.

Q29 **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.

Q30 **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.

Q31 **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.

Q32 **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.

Q33 **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.

Q34 **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.

Q35 **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.