

Which of the following is a supervised learning task?

- A. Grouping customers into market segments using only purchase histories (no labels).
- B. Predicting house prices from features using historical labeled sale prices.
- C. Finding principal components of a dataset using PCA.
- D. Discovering association rules in market-basket data.

You split data into training, validation, and test sets. The validation set is typically used to...

- A. Estimate final generalization performance after all design choices are frozen.
- B. Fit model parameters (weights) directly.
- C. Tune hyperparameters / select among candidate models.
- D. Increase the training set size via augmentation.

A model gets 99% accuracy on the training set but only 60% on the test set. This is most indicative of...

- A. Underfitting.
- B. Overfitting.
- C. Perfect generalization.
- D. A guaranteed data leak.

In many models (e.g., ridge regression), increasing the regularization strength  $\lambda$  tends to...

- A. Decrease bias and increase variance.
- B. Increase bias and decrease variance.
- C. Decrease both bias and variance.
- D. Increase both bias and variance.

Which of the following is a regression problem?

- A. Detecting whether an email is spam (yes/no).
- B. Predicting tomorrow's temperature (a real value).
- C. Classifying images into 10 digit classes.
- D. Assigning customers to "high/medium/low" risk categories.

In a standard feed-forward neural network, a neuron output is typically computed as...

- A. The average of all inputs.
- B. A weighted sum of inputs plus a bias, passed through an activation function.
- C. The product of all inputs.
- D. The maximum input value.

Backpropagation is used to compute...

- A. The gradient of the loss with respect to each weight.
- B. The confusion matrix for evaluation.
- C. The optimal number of hidden layers automatically.
- D. The optimal number of clusters in k-means.

Dropout is primarily used to...

- A. Make training fully deterministic.
- B. Reduce overfitting by randomly zeroing a subset of activations during training.
- C. Increase the number of trainable parameters.
- D. Replace gradient descent with a closed-form solution.

Which activation function is most commonly chosen to mitigate vanishing gradients in deep networks?

- A. Sigmoid
- B. Tanh
- C. ReLU
- D. Step function

In 200 transactions, itemset  $\{A,B\}$  appears in 50 transactions. What is  $\text{support}(\{A,B\})$ ?

- A. 0.10
- B. 0.25
- C. 0.40
- D. 50

The confidence of an association rule  $X \rightarrow Y$  equals...

- A.  $\text{support}(X) / \text{support}(Y)$
- B.  $\text{support}(X \cup Y) / \text{support}(X)$
- C.  $\text{support}(Y) / \text{support}(X)$
- D.  $\text{support}(X \cup Y)$

If  $\text{lift}(X \rightarrow Y) > 1$ , then...

- A. X and Y are negatively correlated.
- B. X and Y are independent.
- C. The occurrence of X increases the probability of Y.
- D. The rule has zero confidence.

The anti-monotone property used by Apriori states that...

- A. If an itemset is frequent, all its supersets are frequent.
- B. If an itemset is infrequent, all its supersets are infrequent.
- C. If an itemset is frequent, all its subsets are infrequent.
- D. Support always increases with itemset size.

A major computational cost in the Apriori algorithm is...

- A. Computing lift after rules are generated.
- B. Generating candidate itemsets and repeatedly scanning the database to count supports.
- C. Sorting items alphabetically.
- D. Choosing the best distance metric.

Consider 6 transactions:

T1 {A,B,C}

T2 {A,B}

T3 {A,B,C}

T4 {A,C}

T5 {B,C}

T6 {A,B,C}

With minsup = 50% and minconf = 75%, itemset {A,B,C} is frequent.

Considering ONLY rules derived from {A,B,C} with antecedent size 2, which set meets minconf?

A. { $A \rightarrow B$ ,  $B \rightarrow A$ }

B. { $AB \rightarrow C$ ,  $AC \rightarrow B$ ,  $BC \rightarrow A$ }

C. { $A \rightarrow C$ ,  $C \rightarrow A$ }

D. No rule derived from {A,B,C} satisfies minconf

In k-means clustering, you must specify \_\_\_\_ in advance.

- A. The number of clusters  $k$
- B. Class labels for each point
- C. A learning rate
- D. Kernel parameters

Standard k-means most commonly uses which distance measure?

- A. Hamming distance
- B. Euclidean distance (squared)
- C. Jaccard similarity
- D. KL divergence

Which problem is best suited for clustering?

- A. Predicting a patient's blood pressure (a real value).
- B. Grouping news articles into topics when no labels are available.
- C. Classifying handwritten digits (0–9) with labeled images.
- D. Predicting whether a transaction is fraudulent (yes/no).

1D points: {1, 2, 6, 7}. Run k-means with  $k=2$  and initial centroids  $c1=1$  and  $c2=7$ . After ONE assignment + centroid update step, what are the new centroids?

- A.  $c1=1, c2=7$
- B.  $c1=1.5, c2=6.5$
- C.  $c1=2, c2=6$
- D.  $c1=3.0, c2=5.0$

A decision tree is most commonly a(n)...

- A. Supervised learning algorithm
- B. Unsupervised learning algorithm
- C. Reinforcement learning algorithm
- D. Self-supervised only algorithm

The ID3 decision tree algorithm selects a split attribute by maximizing...

- A. Gini index
- B. Information gain
- C. Margin
- D. Likelihood

Pruning a decision tree is mainly used to...

- A. Increase model complexity
- B. Reduce overfitting and improve generalization
- C. Guarantee zero training error
- D. Make the tree deeper

In a random forest, each individual tree is typically trained...

- A. On the full dataset using all features, making trees identical.
- B. On a bootstrap sample and using random feature subsets when splitting.
- C. Only on misclassified samples from the previous tree.
- D. Using gradient descent on a convex loss.

Compared with a single deep decision tree, combining many de-correlated trees (random forest) mainly reduces...

- A. Bias
- B. Variance
- C. Both bias and variance always
- D. Neither; it only speeds up training

Training set S has 10 examples (6 positive, 4 negative). Two candidate binary attributes split S as:

Attribute A: (4+,0-) and (2+,4-)

Attribute B: (3+,1-) and (3+,3-)

Using entropy-based information gain, which attribute yields the larger information gain?

- A. Attribute A
- B. Attribute B
- C. They are equal
- D. Cannot be determined

Ordinary Least Squares (OLS) learns parameters by minimizing...

- A. Sum of squared residuals
- B. Sum of residuals
- C. Hinge loss
- D. Number of misclassifications

In linear regression  $y = w_0 + w_1x_1 + \dots + w_nx_n$ , if  $w_j > 0$  then increasing  $x_j$  (holding others fixed) tends to...

- A. Decrease predicted  $y$
- B. Increase predicted  $y$
- C. Leave predicted  $y$  unchanged
- D. Make prediction binary

Which statement about L1 vs L2 regularization is correct?

- A. L1 regularization tends to set some coefficients exactly to zero (feature selection).
- B. L2 regularization always yields sparse solutions.
- C. L1 cannot reduce overfitting.
- D. L2 works only for classification.

A common assumption in classical linear regression is that the errors...

- A. Have zero mean and constant variance (homoscedasticity).
- B. Are always positive.
- C. Make all input features independent.
- D. Force coefficients to sum to 1.

Consider linear regression with Gaussian noise and a zero-mean Gaussian prior on weights  $w$ .

Maximizing the posterior (MAP) is equivalent to minimizing...

- A.  $SSE + \lambda \|w\|_1$
- B.  $SSE + \lambda \|w\|_2^2$
- C. Cross-entropy +  $\lambda \|w\|_2^2$
- D. Hinge loss +  $\lambda \|w\|_2^2$

For a highly imbalanced binary classification problem (e.g., 99% negative, 1% positive), which metric is usually more informative than accuracy?

- A. Accuracy
- B. Precision/Recall (e.g., F1-score or PR-AUC)
- C. Mean Squared Error (MSE)
- D.  $R^2$

In k-fold cross-validation (e.g.,  $k=5$ ), the model is trained and evaluated...

- A. Once, using a single holdout split.
- B.  $k$  times, each time using a different fold as the validation set.
- C.  $2k$  times, once per class.
- D. Only on the test set.

A binary classifier has  $TP=30$ ,  $FP=10$ ,  $FN=20$ ,  $TN=40$ . What is the F1-score for the positive class?

- A. 0.60
- B. 0.67
- C. 0.75
- D. 0.80

In a hard-margin linear SVM, support vectors are the training points that...

- A. Lie farthest from the separating hyperplane.
- B. Lie on the margin boundaries and define the optimal hyperplane.
- C. Are misclassified by the final model.
- D. Have the largest feature norms.

For a soft-margin SVM with objective  $\|w\|^2 + C \cdot \sum \xi_i$ , increasing  $C$  usually...

- A. Widens the margin and allows more violations.
- B. Penalizes violations more, often reducing training error but risking overfitting.
- C. Eliminates the need for kernels.
- D. Forces  $w = 0$ .

Min–max normalization of a feature  $x$  to the range  $[0,1]$  transforms  $x$  into...

- A.  $(x - \text{mean})/\text{std}$
- B.  $(x - \text{min})/(\text{max} - \text{min})$
- C.  $x / \|x\|$
- D.  $\log(x)$

Choose the most reasonable order in a text classification pipeline.

- A. Train model → Collect data → Tokenize → Remove stop-words
- B. Collect data → Clean text → Tokenize → Remove stop-words → Vectorize (TF-IDF) → Train model
- C. Remove stop-words → Train model → Vectorize → Collect data
- D. Vectorize → Collect data → Train model → Clean text

Bayes' rule states that...

- A.  $P(A|B) = P(A)P(B)$
- B.  $P(A|B) = P(B|A)P(A) / P(B)$
- C.  $P(A|B) = P(B) / P(A)$
- D.  $P(A|B) = 1 - P(B|A)$

The “naive” assumption in Naive Bayes is that...

- A. All classes are equally likely.
- B. Features are conditionally independent given the class label.
- C. The decision boundary must be linear.
- D. Training requires gradient descent.

Which pairing is correct?

- A. Logistic regression – generative model
- B. Gaussian Naive Bayes – discriminative model
- C. Gaussian Naive Bayes – generative model; Logistic regression – discriminative model
- D. k-means – supervised learning