Machine Learning — Question Set (B)

40 questions

Topics: Introduction • ANN • ARM • Clustering • Trees • Regression • Assessment • SVM • Preprocessing • Probabilistic • Advanced

Q1. Introduction

Missing-level • Single

Which statement best describes the “generalization gap” of a model?

A. Difference between training loss and validation/test loss
B. Difference between precision and recall
C. Difference between accuracy and F1-score
D. Difference between feature scaling methods

Q2. Introduction

Missing-level • Single

A model has high training error and high validation error. What is the most likely issue?

A. Underfitting (high bias)
B. Overfitting (high variance)
C. Data leakage from test to train
D. Too much regularization is impossible

Q3. Introduction

Missing-level • Single

Which of the following is a hyperparameter (not learned directly from training loss minimization)?

A. Learning rate in gradient descent

B. Weight vector w in linear regression

C. Bias term b in logistic regression

D. Predicted label ŷ for each sample

Q4. Introduction

Missing-level • Single

Which choice is the best reason to keep a separate "test set"?

A. To estimate final performance on unseen data after model selection

B. To tune hyperparameters (e.g., λ, C)

C. To make training faster

D. To increase the number of training samples

Q5. Introduction

Missing-level • Multiple

Which tasks are typically unsupervised? (Choose ALL that apply)

A. Customer segmentation based on purchase vectors
B. Predicting house price from square footage
C. Discovering topics in a collection of documents
D. Classifying emails as spam or not spam

Q6. ANN / Neural networks

Missing-level • Single

Why are all weights in a multi-layer neural network NOT initialized to zero?

A. Zero initialization keeps neurons symmetric so they learn the same features
B. Zero initialization makes the loss function non-differentiable
C. Zero initialization prevents using mini-batch training
D. Zero initialization forces gradients to explode

Q7. ANN / Neural networks

Missing-level • Single

In backpropagation, which quantity is propagated from later layers to earlier layers?

A. Gradient of the loss with respect to activations (error signal)

B. The raw input features x

C. The learning rate η

D. The one-hot labels y

Q8. ANN / Neural networks

Easy • Single

Which activation function outputs values strictly in the range (0, 1)?

A. Sigmoid
B. ReLU
C. Tanh
D. Identity (linear)

Q9. ANN / Neural networks

Easy • Single

Which technique reduces overfitting by randomly “dropping” units during training?

A. Dropout
B. Batch normalization
C. Momentum
D. Early stopping always increases overfitting

Q10. ARM / Association Rule Mining

Missing-level • Single

If we increase the minimum confidence (minconf) while keeping minsup fixed, what usually happens?

A. Fewer association rules pass the threshold

B. More frequent itemsets are generated

C. Support of all itemsets increases

D. Apriori needs fewer database scans regardless of max itemset size

Q11. ARM / Association Rule Mining

Missing-level • Single

Which item is NOT required to run the Apriori algorithm?

A. Class labels for each transaction
B. A transaction database
C. A minimum support threshold
D. A method to count itemset supports

Q12. ARM / Association Rule Mining

Missing-level • Single

Suppose support(X)=0.20 and support(X ∪ Y)=0.05. What is confidence(X → Y)?

A. 0.25

B. 0.05

C. 0.15

D. 4.00

Q13. ARM / Association Rule Mining

Easy • Single

Lift($X \rightarrow Y$) is defined as:

A. confidence($X \rightarrow Y$) / support(Y)

B. support($X \cup Y$) + support(X)

C. support(X) / support(Y)

D. confidence($X \rightarrow Y$) − support(Y)

Q14. ARM / Association Rule Mining

Easy • Single

Apriori anti-monotone property implies: if an itemset is infrequent, then all its supersets are:

A. Infrequent
B. Frequent
C. Candidates in the next iteration
D. Guaranteed to have higher confidence

Q15. ARM / Association Rule Mining

Hard • Single

A dataset has 10 transactions. support({A,B})=4/10, support({C})=3/10, and support({A,B,C})=3/10. With minsup=30% and minconf=80%, which rule is valid?

A. $\{A,B\} \longrightarrow \{C\}$

B. $\{C\} \longrightarrow \{A,B\}$

C. $\{A\} \longrightarrow \{B,C\}$

D. No rule is valid

Q16. Clustering

Easy • Single

K-means clustering chooses centroids to minimize:

A. Sum of squared distances from points to their assigned centroid
B. Number of misclassified labels
C. Total absolute correlation between features
D. Entropy of the class distribution

Q17. Clustering

Easy • Single

Before running k-means with Euclidean distance on features with very different scales, you should usually:

A. Standardize/normalize features

B. Convert labels to one-hot vectors

C. Increase K until training loss is zero

D. Remove all outliers so every cluster is spherical

Q18. Clustering

Medium • Single

1D points: {1, 2, 9, 10}. Initial centroids: c1=1, c2=10. After one k-means update (assign then recompute), what are the new centroids?

A. c1=1.5, c2=9.5

B. c1=2.0, c2=10.0

C. c1=3.0, c2=9.0

D. c1=1.0, c2=10.0

Q19. Clustering

Hard • Multiple

Which statements about k-means are TRUE? (Choose ALL that apply)

A. The objective value (WCSS) never increases after an assignment+update iteration

B. K-means always finds the global optimum regardless of initialization

C. Increasing K can only decrease (or keep) the optimal WCSS on the training data

D. K-means naturally handles non-convex clusters (e.g., two-moons) without issues

Q20. Decision Trees / Tree-based

Missing-level • Single

Information gain for a split is based on the reduction of:

A. Impurity (e.g., entropy or Gini) from parent to children

B. Learning rate from epoch to epoch

C. Number of features used in the model

D. Regularization strength in the loss function

Q21. Decision Trees / Tree-based

Missing-level • Single

In a random forest, why is a random subset of features often considered at each split?

A. To reduce correlation between trees and improve ensemble variance reduction
B. To guarantee perfect accuracy on training data
C. To make each tree deterministic
D. To ensure all trees use the same splits

Q22. Decision Trees / Tree-based

Missing-level • Single

Which approach is commonly used to reduce overfitting in a decision tree?

A. Pruning (pre-pruning or post-pruning)
B. Using a larger max depth with no stopping criteria
C. Removing the training set and training on the test set
D. Always choosing splits with the smallest information gain

Q23. Decision Trees / Tree-based

Easy • Single

Gain ratio is often preferred over information gain because it:

A. Reduces bias toward attributes with many distinct values
B. Eliminates the need to compute entropy
C. Works only for regression trees
D. Always produces shallower trees

Q24. Decision Trees / Tree-based

Easy • Single

Compared to k-NN and SVM, a key advantage of decision trees is that they:

A. Are easy to interpret as a set of if-then rules

B. Always outperform ensembles on test data

C. Require feature scaling to work correctly

D. Cannot handle categorical features

Q25. Decision Trees / Tree-based

Hard • Single

A dataset has entropy H(parent)=1.0. Split by feature A gives weighted child entropy 0.60. Split by feature B gives weighted child entropy 0.72. Which split has higher information gain?

A. Feature A

B. Feature B

C. Both equal

D. Cannot be determined without knowing the number of samples

Q26. Linear Regression / Regression

Missing-level • Single

Which regularization is most likely to produce sparse coefficients (some exactly zero)?

A. L1 (Lasso) regularization

B. L2 (Ridge) regularization

C. No regularization

D. Early stopping guarantees sparsity

Q27. Linear Regression / Regression

Missing-level • Single

In linear regression, the coefficient of determination R^2 is commonly interpreted as:

A. The fraction of variance in y explained by the model
B. The slope of the best-fit line
C. The average absolute error
D. The probability the model is correct

Q28. Linear Regression / Regression

Easy • Single

Ordinary Least Squares (OLS) fits parameters by minimizing:

A. Sum of squared residuals
B. Sum of absolute residuals
C. Classification error rate
D. Hinge loss

Q29. Linear Regression / Regression

Easy • Single

A key difference between linear regression and logistic regression is that logistic regression:

A. Models $P(y=1|x)$ using a sigmoid (or similar) link and is used for classification
B. Always yields a closed-form solution
C. Requires categorical features only
D. Minimizes squared error on real-valued targets

Q30. Linear Regression / Regression

Hard • Multiple

For ridge regression, increasing the regularization strength λ typically: (Choose ALL that apply)

A. Increases bias

B. Decreases variance

C. Makes coefficients larger in magnitude

D. Moves coefficients toward zero

Q31. Model assessment

Missing-level • Single

For an imbalanced classification dataset, which cross-validation variant is most appropriate?

A. Stratified k-fold (preserve class proportions in each fold)
B. Random k-fold without constraints
C. Leave-one-out only
D. Train-test split without shuffling

Q32. Model assessment

Missing-level • Single

Precision for the positive class is defined as:

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

Q33. Model assessment

Hard • Single

In a disease screening task, false negatives are extremely costly (missing a sick patient). Which metric should you prioritize when selecting a threshold?

A. Recall (sensitivity)
B. Precision
C. Specificity only
D. Overall accuracy

Q34. SVM

Missing-level • Single

In soft-margin SVM, the hyperparameter C mainly controls:

A. Trade-off between margin width and classification errors (slack penalties)
B. Number of hidden layers
C. Number of clusters used in training
D. Learning rate schedule

Q35. SVM

Missing-level • Single

In a hard-margin linear SVM, which training points become support vectors?

A. Points that lie on the margin boundaries

B. All points in the majority class

C. Only misclassified points

D. All points farthest from the hyperplane

Q36. Preprocessing

Missing-level • Single

Why should you compute scaling parameters (mean/variance) using only the training set?

A. To avoid data leakage from validation/test into training
B. Because scaling on all data is mathematically incorrect
C. Because models cannot handle scaled validation data
D. Because scaling changes labels

Q37. Preprocessing

Missing-level • Single

A common simple strategy for missing values in a numeric feature is:

A. Impute with the mean/median of the training feature
B. Replace missing values with random class labels
C. Drop all columns that contain any missing value
D. Always set missing values to 0 (no exceptions)

Q38. Probabilistic

Missing-level • Single

Compared to Maximum Likelihood Estimation (MLE), Maximum A Posteriori (MAP) estimation:

A. Incorporates a prior distribution over parameters
B. Ignores observed data
C. Always produces the same estimate as MLE
D. Requires no assumptions about parameters

Q39. Probabilistic

Missing-level • Single

The “naive” assumption in Naive Bayes classifier is that features are:

A. Conditionally independent given the class label
B. Always independent (unconditionally)
C. Always normally distributed
D. Always binary

Q40. Advanced (Misc)

Missing-level • Single

In the EM algorithm for latent-variable models, the E-step primarily computes:

A. Expected latent-variable responsibilities given current parameters

B. A new learning rate schedule

C. The exact global optimum in one step

D. A decision boundary that maximizes the margin