Machine Learning

Medium → Hard Multiple-Choice Questions

Machine Learning — Medium/Hard MCQs

Q1 / 40 • Lecture 1: Introduction to Machine Learning • Difficulty: Medium

You have 10,000 labeled examples but the labels are noisy (≈15% incorrect). Which change is most likely to improve generalization without collecting more data?

A. Increase model capacity (deeper network) until training error is ~0%

B. Use stronger regularization and/or early stopping, and monitor validation loss

C. Optimize only training accuracy; validation is unreliable with label noise

D. Remove the validation split and train on all data to reduce variance

Q2 / 40 • Lecture 1: Introduction to Machine Learning • Difficulty: Hard

Which statement best describes why empirical risk minimization (ERM) alone can fail, and what structural risk minimization (SRM) changes?

A. ERM fails only for non-convex losses; SRM makes the loss convex

B. ERM can overfit with limited data; SRM trades off fit vs capacity via a complexity term / model class hierarchy

C. ERM fails only for high-dimensional data; SRM reduces dimension with PCA

D. ERM is biased; SRM removes bias by using unbiased estimators

Machine Learning — Medium/Hard MCQs

Q3 / 40 • Lecture 1: Introduction to Machine Learning • Difficulty: Hard

Suppose Model A has lower training error than Model B, but higher validation error. Which is the most likely explanation?

A. Model A has higher bias than Model B

B. Model A has higher variance (overfitting) than Model B

C. Model B has higher variance than Model A

D. Validation error is always larger than training error for any model

Q4 / 40 • Lecture 2: Linear Regression • Difficulty: Medium

In ordinary least squares (OLS), if features are highly collinear, what is the most direct consequence for the estimated coefficients?

A. Coefficients are unbiased and have low variance

B. Coefficients remain unbiased (under standard assumptions) but can have large variance / instability

C. Coefficients become biased but variance decreases

D. Predictions must be exact on the training set

Q5 / 40 • Lecture 2: Linear Regression • Difficulty: Hard

Ridge regression solves: $\min_w ||y - Xw||^2 + \lambda||w||^2$. Which Bayesian interpretation matches ridge?

A. MAP with a Laplace prior on w

B. MAP with a zero-mean Gaussian prior on w

C. Posterior mean with a uniform prior on w

D. MLE with a Gaussian likelihood of y but no prior

Q6 / 40 • Lecture 2: Linear Regression • Difficulty: Hard

Consider gradient descent on OLS with learning rate η. Which preprocessing most directly improves conditioning and typically allows larger stable η?

A. Shuffling data once before training

B. Centering and scaling features to comparable variance (standardization)

C. Adding polynomial features

D. Removing the intercept term

Q7 / 40 • Lecture 2: Linear Regression • Difficulty: Medium

When the noise variance σ^2 is unknown, which model selection approach is most appropriate for choosing λ in ridge regression?

A. Training error minimization

B. K-fold cross-validation

C. Always set $\lambda = 0$

D. Use the number of features as λ

Machine Learning — Medium/Hard MCQs

Q8 / 40 • Lectures 3–4: Clustering • Difficulty: Medium

Which statement is TRUE about k-means?

A. It minimizes the sum of Euclidean distances (not squared) to centroids

B. It minimizes within-cluster sum of squared distances to centroids

C. It is invariant to feature scaling

D. It guarantees the global optimum of its objective

Q9 / 40 • Lectures 3–4: Clustering • Difficulty: Hard

k-means can be viewed as a limiting case of which probabilistic model under which constraint?

A. Gaussian mixture model with equal spherical covariances and hard assignments as variance → 0

B. Naive Bayes with Bernoulli features and soft assignments

C. Hidden Markov model with equal transition probabilities

D. Poisson mixture model with shared rate parameter

Q10 / 40 • Lectures 3–4: Clustering • Difficulty: Medium

Silhouette score for a point compares:

A. Its distance to the farthest cluster center vs nearest center

B. Its mean intra-cluster distance vs mean nearest-cluster distance

C. The ratio of cluster sizes

D. The difference between training and test likelihoods

Q11 / 40 • Lectures 3–4: Clustering • Difficulty: Hard

In DBSCAN, what is the most likely outcome of increasing ε (eps) while keeping minPts fixed?

A. More clusters, more noise points

B. Fewer clusters; clusters may merge; fewer noise points

C. No change, because DBSCAN is scale invariant

D. Always exactly k clusters, like k-means

Q12 / 40 • Lectures 3–4: Clustering • Difficulty: Hard

For a Gaussian mixture model, the E-step in EM computes responsibilities r_{nk}. Which expression is correct?

A. $r_{nk} = \pi_k / \Sigma_j \pi_j$

B. $r_{nk} = N(x_n \mid \mu_k, \Sigma_k) / \Sigma_j N(x_n \mid \mu_j, \Sigma_j)$

C. $r_{nk} = \pi_k N(x_n \mid \mu_k, \Sigma_k) / \Sigma_j \pi_j N(x_n \mid \mu_j, \Sigma_j)$

D. $r_{nk} = \text{argmax}_k \pi_k N(x_n \mid \mu_k, \Sigma_k)$

Q13 / 40 • Lectures 3–4: Clustering • Difficulty: Medium

Which criterion is most directly motivated as an approximation to the marginal likelihood for selecting the number of mixture components?

A. Elbow method on SSE

B. BIC

C. Random restart count

D. Silhouette score only

In a classification tree, a split is typically chosen to:

A. Maximize validation accuracy at each node

B. Minimize impurity (e.g., Gini/entropy) after the split

C. Maximize the number of leaves

D. Make all features equally used

Q15 / 40 • Lecture 5: Decision Tree & Random Forest • Difficulty: Hard

Which statement about post-pruning a decision tree is most accurate?

A. Pruning typically decreases bias and increases variance

B. Pruning typically increases bias but decreases variance

C. Pruning always improves both training and validation accuracy

D. Pruning is only meaningful for regression trees, not classification

In bagging / random forests with bootstrap samples of size n drawn from n training points, approximately what fraction of points are out-of-bag (OOB) for a given tree?

A. ≈ 0%

B. ≈ 36.8%

C. ≈ 50%

D. ≈ 63.2%

Q17 / 40 • Lecture 5: Decision Tree & Random Forest • Difficulty: Hard

Which issue most commonly affects impurity-based feature importance (MDI) in random forests?

A. It is unbiased for all feature types

B. It can be biased toward features with many possible split points / high cardinality

C. It cannot be computed without a validation set

D. It always matches permutation importance exactly

Q18 / 40 • Lecture 6: Neural Networks • Difficulty: Medium

For $\mathrm{ReLU}(x) = \max(0, x)$, what is $\partial \mathrm{ReLU}/\partial x$ for $x < 0$?

A. 1

B. 0

C. x

D. Undefined for all x

Q19 / 40 • Lecture 6: Neural Networks • Difficulty: Hard

Which change most directly mitigates vanishing gradients in very deep feedforward networks?

A. Using sigmoid activations everywhere

B. Using residual connections (skip connections)

C. Removing non-linearities

D. Training with batch size 1 only

Q20 / 40 • Lecture 6: Neural Networks • Difficulty: Hard

With softmax p and one-hot label y, cross-entropy $L = -\Sigma_i y_i \log p_i$. What is $\partial L/\partial z$ (logits z)?

A. $p + y$

B. $y - p$

C. $p - y$

D. $\log p - \log y$

Q21 / 40 • Lecture 6: Neural Networks • Difficulty: Medium

With “inverted dropout” (most common implementation), what happens at test time?

A. Dropout remains active but with a smaller drop probability

B. No dropout is applied; activations are already scaled during training

C. Weights are randomly zeroed instead of activations

D. You must multiply activations by the keep probability at test time

Q22 / 40 • Lecture 6: Neural Networks • Difficulty: Hard

Batch normalization is often said to reduce “internal covariate shift”. Practically, its most consistent training benefit is:

A. Guaranteeing zero generalization gap

B. Allowing larger learning rates and stabilizing optimization

C. Replacing the need for data normalization entirely

D. Making the loss convex

Q23 / 40 • Lecture 6: Neural Networks • Difficulty: Hard

If a network severely overfits, which intervention is LEAST likely to help?

A. Increase weight decay

B. Increase dropout

C. Early stopping

D. Train for more epochs with the same settings

Q24 / 40 • Lecture 7: Support Vector Machines • Difficulty: Hard

Which objective corresponds to a soft-margin linear SVM (primal) with hinge loss?

A. $\min_w ||w||^2$ subject to $y_i(w^T x_i) \geq 1$ for all i

B. $\min_w \Sigma_i (y_i - w^T x_i)^2$

C. $\min_w (1/2)||w||^2 + C \Sigma_i \max(0, 1 - y_i(w^T x_i))$

D. $\min_w (1/2)||w||^2 + C \Sigma_i |1 - y_i(w^T x_i)|$

Increasing C in a soft-margin SVM typically:

A. Increases regularization, widening the margin

B. Decreases regularization, penalizing violations more strongly

C. Has no effect on the solution

D. Makes the kernel unnecessary

The “kernel trick” is possible primarily because the SVM optimization depends on data only through:

A. Euclidean distances to the origin

B. Dot products between examples

C. Class priors

D. Feature normalization constants

Which points are support vectors in a trained soft-margin SVM?

A. All correctly classified points

B. Only misclassified points

C. Points with non-zero dual coefficients ($\alpha_i > 0$), typically on/inside the margin

D. Only points closest to the centroid of each class

Q28 / 40 • Lecture 8: Performance Evaluation • Difficulty: Hard

For a highly imbalanced dataset (positive class rare), which metric is usually more informative for ranking probabilistic classifiers?

A. Accuracy

B. ROC-AUC

C. PR-AUC

D. R^2

Q29 / 40 • Lecture 8: Performance Evaluation • Difficulty: Medium

To avoid leakage in k-fold cross-validation with feature standardization, you should:

A. Fit the scaler on the full dataset once, then split into folds

B. Fit the scaler on each training fold only, then apply to its validation fold

C. Fit the scaler on the validation fold only

D. Never standardize because it changes the data distribution

Q30 / 40 • Lecture 8: Performance Evaluation • Difficulty: Hard

You tune hyperparameters using cross-validation and then report the best CV score as the final test performance. The main issue is:

A. The score is unbiased if you used enough folds

B. It is optimistically biased; nested CV or a held-out test set is needed

C. It is pessimistically biased; you should add 5% to correct it

D. It is only an issue for neural networks, not for linear models

Q31 / 40 • Lecture 8: Performance Evaluation • Difficulty: Hard

A classifier has good ROC-AUC but poor calibration. Which action most directly improves calibration without changing ranking much?

A. Threshold at 0.5

B. Platt scaling or isotonic regression on a validation set

C. Increase model depth

D. Train longer

Q32 / 40 • Lecture 8: Performance Evaluation • Difficulty: Medium

Which metric is threshold-dependent?

A. ROC-AUC

B. PR-AUC

C. Accuracy

D. Log-loss (cross-entropy)

Naive Bayes can work surprisingly well even when its conditional independence assumption is violated because:

A. It exactly models all feature dependencies

B. Classification can still be correct if errors in probability estimates cancel in the argmax

C. It always overfits less than logistic regression

D. It produces calibrated probabilities by design

A key guarantee of EM (with exact E- and M-steps) is that each iteration:

A. Decreases the data log-likelihood

B. Increases (or leaves unchanged) the data log-likelihood

C. Finds the global maximum likelihood solution

D. Produces unbiased parameter estimates

Which assumption distinguishes LDA from QDA in Gaussian discriminant analysis?

A. LDA assumes different class covariance matrices; QDA assumes shared covariance

B. LDA assumes shared class covariance matrix; QDA allows class-specific covariances

C. LDA assumes Bernoulli features; QDA assumes Gaussian features

D. QDA assumes linear decision boundaries; LDA assumes quadratic boundaries

Q36 / 40 • Lecture 10: Ensemble Learning • Difficulty: Medium

Which pairing is most accurate?

A. Bagging mainly reduces bias; Boosting mainly reduces variance

B. Bagging mainly reduces variance; Boosting often reduces bias (and can also reduce variance)

C. Both bagging and boosting only reduce variance

D. Both bagging and boosting only reduce bias

Q37 / 40 • Lecture 10: Ensemble Learning • Difficulty: Hard

In AdaBoost, what happens to the weights of misclassified training examples after an iteration?

A. They decrease, so the next learner focuses on easy points

B. They increase, so the next learner focuses on hard points

C. They are set to zero permanently

D. They are randomized uniformly at each iteration

Q38 / 40 • Lecture 11: Reinforcement Learning • Difficulty: Hard

Which is the Bellman optimality equation for Q*?

A. $Q^*(s,a) = E[r_t]$

B. $Q^*(s,a) = E[r_{t+1} + \gamma \max_{a'} Q^*(s',a') \mid s,a]$

C. $Q^*(s,a) = \max_a E[r_{t+1} \mid s]$

D. $Q^*(s,a) = E[r_{t+1} + \gamma Q^*(s,a)]$

Q39 / 40 • Lecture 11: Reinforcement Learning • Difficulty: Hard

Which statement best contrasts Q-learning and SARSA?

A. Both are on-policy methods

B. Q-learning is off-policy; SARSA is on-policy

C. Q-learning requires a model; SARSA does not

D. SARSA converges faster because it is off-policy

Q40 / 40 • Lecture 12: Regularization • Difficulty: Hard

Lasso solves: $\min_w ||y - Xw||^2 + \lambda||w||_1$. Compared to ridge, lasso is more likely to:

A. Shrink all coefficients equally but keep them non-zero

B. Set some coefficients exactly to zero (sparsity)

C. Increase sensitivity to multicollinearity

D. Always have a closed-form solution