Ensemble Methods

Advanced Quiz Questions

10 Conceptual Questions

Test Your Understanding

Question 21

You can build two ensembles of 21 base classifiers:

- Ensemble A: Base models are very accurate but highly correlated
- Ensemble B: Base models are slightly less accurate individually but much less correlated

Assuming both use majority voting, which is MORE likely to perform better?

A. Ensemble A, because base accuracy is all that matters

B. Ensemble B, because lower correlation often matters as much as accuracy

C. Both will perform the same

D. It depends only on the number of models, not their correlation

Question 22

For which base learner is bagging LEAST likely to give a big performance gain?

A. Deep, unpruned decision trees

B. Shallow decision trees

C. High-capacity neural nets with small data

D. Ordinary least squares linear regression on a well-conditioned dataset

Question 23

Which statement best summarizes how bagging and boosting typically affect bias and variance?

A. Bagging ↓variance, Boosting ↓bias (often ↑variance slightly)

B. Bagging ↓bias, Boosting ↓variance

C. Bagging ↑bias, Boosting ↑variance

D. Both mainly reduce bias

Question 24

You have a SMALL, noisy dataset and a high-variance decision tree model. Computational resources are limited. Which method is the most DANGEROUS choice if you push it too aggressively?

A. Bagging with a modest number of trees

B. Random forest with limited depth

C. Gradient boosting with many trees and high learning rate

D. Stacking simple linear models with cross-validation

Question 25

You decrease the learning rate (α) in a gradient boosting model from 0.1 to 0.01 but keep the same number of trees. What is the most likely effect?

A. Massive overfitting

B. Underfitting unless you increase the number of trees

C. No change, learning rate doesn't matter

D. The model becomes non-interpretable

Question 26

You're training a random forest on a dataset with 1,000 features. You notice trees are very similar and heavily correlated. Which change is most likely to INCREASE diversity and improve performance?

A. Increase the number of features considered at each split

B. Decrease the number of features considered at each split

C. Reduce the number of trees

D. Train each tree on the full dataset without bootstrapping

Question 27

You have many noisy, irrelevant features. You switch from a standard random forest to ExtraTrees (Extremely Randomized Trees). What behavior do you expect?

A. ExtraTrees will always overfit more than RF

B. ExtraTrees may reduce variance further by injecting more randomness into splits

C. ExtraTrees will behave identically to RF

D. ExtraTrees cannot handle noisy features at all

Question 28

You apply AdaBoost with many iterations to a dataset that contains several mislabeled outliers. What is the most likely outcome?

A. AdaBoost will learn to ignore the outliers completely

B. AdaBoost will heavily focus on the outliers and may overfit them

C. AdaBoost will behave like bagging and just average them away

D. AdaBoost will fail to converge

Question 29

You design a stacking pipeline:

- 1. Train several base models on the full training data
- 2. Use these models to predict on the SAME training data
- 3. Train a meta-model on these predictions
- 4. Evaluate on the test set

What is the main methodological flaw?

A. You didn't use enough base models

B. The meta-model is linear

C. You have data leakage because base models and meta-model both see the same training targets on the same examples

D. You used classification instead of regression

Question 30

Your single decision tree model UNDERFITS (high bias) the training data: even training error is quite large. You want to fix this with an ensemble. Which method is MORE likely to help first?

A. Bagging

B. Random forests with shallow trees

C. Gradient boosting with deeper trees

D. Decreasing the training data size