Dhirubhai Ambani Institute of
Information and Communication Technology

Surprise Quiz III

IT496: Introduction to Data Mining

Date: November 7, 2023 | Timings: 10:15 PM – 10:45 PM

READ THE TEXT BELOW CAREFULLY

- *Note that many questions have multiple correct answers.*
- For each question, write what you think is (are) the correct option(s) and support your answer with **brief** and **justified** explanations in 1-2 sentences in your answer sheets.
- For each question, you will be given **3** points for the correct answer (only if you select all the correct options with the correct reason) and **-1** point *otherwise*.

10 questions | **30** total points | **20** minutes | **+3** if correct **-1** otherwise.

---

1. `PolynomialFeatures(degree=d, include_bias=False)` transforms a dataset that had *n* features into one that has _____ features:

   (a) (n+d)!/n!d!
   (b) (n/d)!/n!d!
   (c) (n+d)!/(n!+d!)
   (d) (n+d)!/d!

   PolynomialRegression transforms a dataset by adding more interaction and higher-order features to make the regression equation perfect polynomial of degree d.

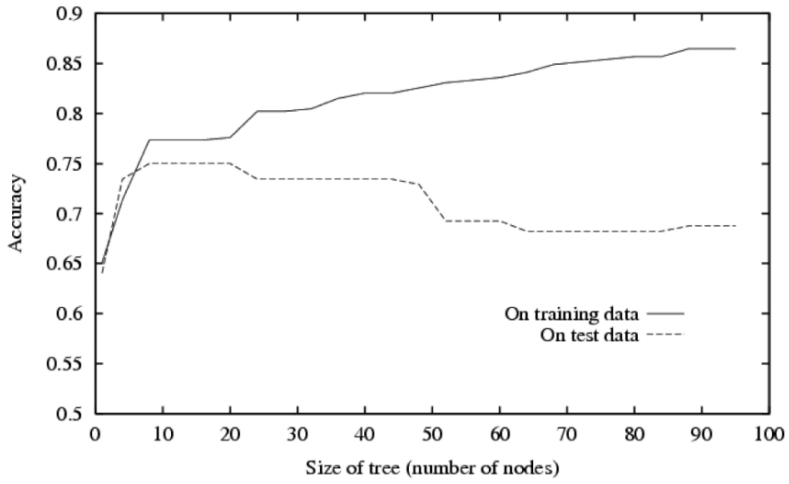
2. In Principal Component Analysis (PCA) for data compression, when reducing the dimensionality of a dataset by retaining only a subset of the principal components, what is the most important factor to consider when choosing the number of components to keep?

   (a) The total variance retained in the reduced dimensionality dataset.
   (b) The original dataset size before compression.
   (c) The number of data points in the dataset.
   (d) The first principal component's eigenvalue.

PCA finds the principal components that capture the maximum variance in the data. When you reduce the dimensionality of the dataset by retaining only a subset of these principal components, you want to ensure that you retain a significant portion of the variance in the data. Retaining a higher percentage of the total variance typically results in better preservation of the information in the dataset, which is the primary goal of data compression in PCA.

3. Consider the plot below showing training and test set accuracy for decision trees of different sizes, using the same training data set to train each tree. Describe how the training data curve (solid line) and the test data curve (dotted line) will change if the number of training examples approaches infinity.

    (a) The training curve will be unchanged.
    (b) The testing curve will be unchanged.
    (c) Both curves will be unchanged
    (d) None of the above



The new training accuracy curve should be below the original training curve (since it's impossible for the trees to overfit infinite training data); the new testing accuracy curve should be above the original testing curve and become identical to the new training curve (since trees learned from infinite training data should perform well on test data and do not overfit at all).

4. Suppose you are building a decision tree using the CART algorithm. At a particular node, you have 80 data points that belong to either class A or class B. The Gini impurity for this node is 0.4. After splitting the node into two child nodes, you find that one child node contains 40 data points, and the Gini impurity for that child node is 0.3. What is the Gini impurity for the other child node?

    (a) 0.2
    (b) 0.25
    (c) 0.35
    (d) None of the above

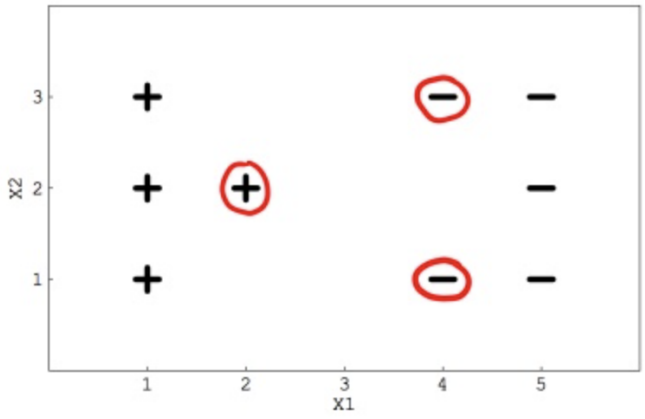
Apply the Gini impurity formula.

5. In a bagging ensemble with 100 base decision trees, each trained on a different bootstrap sample of the training data, what is the typical proportion of the training data included in each bootstrap sample?

   (a) 100%
   (b) 75%
   (c) 63.2%
   (d) 37.5%

The typical proportion of the training data included in each bootstrap sample is approximately 63.2% (1 - 1/e), where "e" is the base of the natural logarithm. This value ensures that, on average, about 63.2% of the training data is included in each bootstrap sample.

6. Suppose you are using a Linear SVM classifier on a two-class classification problem: you are given the following data in which some points are circled red representing support vectors.



   I. On removing any of the red points, will the decision boundary change?

      (a) Yes (b) May be (c) May not be (d) No

These three examples are positioned such that removing one introduces slack in the constraints. So, the decision boundary would completely change.

   II. On removing any of the non-red circled points from the data, will the decision boundary change?

      (a) Yes (b) May be (c) May not be (d) No

On the other hand, the rest of the points in the data won't affect the decision boundary much.

7. In an AdaBoost algorithm, you train a series of decision trees as weak learners. After the first tree is trained, you notice an error rate of 0.2 on the training data. You then apply the AdaBoost algorithm to create an ensemble. How much weight is assigned to this first

decision tree in the final ensemble, assuming the default weight update formula is used? [Hint: Use base 2 for log]

(a) Weight of 0.2
(b) Weight of 1.0
(c) Weight of 0.5
(d) Weight of 10.0

In the AdaBoost algorithm, weights are assigned to the weak learners based on their performance in classifying the training data. The weight update formula typically used in AdaBoost is the Weight of the weak learner = 0.5 * log((1 - error rate) / error rate). In this case, the error rate of the first decision tree is 0.2, so the weight assigned to it would be: Weight of the first decision tree = 0.5 * log2((1 - 0.2) / 0.2) = 0.5 * log2(0.8 / 0.2) = 0.5 * log2(4) = 0.5 * 2 = 1

---

8. In a Support Vector Machine (SVM) classification problem, you have a dataset with 200 data points and features in a 10-dimensional space. After training an SVM, you find that the model has 40 support vectors. Calculate the dimensionality (number of features) of the hyperplane used for classification.

(a) 2
(b) 3
(c) 10
(d) 40

In a Support Vector Machine (SVM), the dimensionality (number of features) of the hyperplane used for classification is the same as the original feature space. The SVM constructs a decision boundary (hyperplane) in the same feature space where the data resides. In this case, you have a dataset with features in a 10-dimensional space. Regardless of the number of support vectors, the dimensionality of the hyperplane remains the same as the original feature space.
So, the correct answer is C) 10, which represents the number of features in the original 10-dimensional space. The number of support vectors does not affect the dimensionality of the hyperplane; it only impacts the structure and location of the hyperplane within the feature space.

9. Suppose you are working on a binary classification problem using Gradient Boosting, and you notice that your model is overfitting the training data. Which of the following hyperparameters should you consider adjusting to reduce overfitting?

(a) Learning Rate
(b) Number of Estimators (Trees)
(c) Max Depth of Trees
(d) Min Samples Split

A) Learning Rate: A lower learning rate can help prevent the model from making too large of adjustments in each iteration, which can lead to overfitting.
B) Number of Estimators (Trees): Reducing the number of estimators (trees) can limit the model's complexity and prevent it from overfitting the training data.
C) Max Depth of Trees: Lowering the maximum depth of trees restricts the complexity of individual trees, thereby reducing the overall complexity of the model and mitigating overfitting.
D) Min Samples Split: Increasing the minimum number of samples required to split a node forces the model to consider more data before making splits, preventing it from overfitting to specific data points.

10. In a bagging ensemble with 50 base decision tree classifiers, each tree is trained on a random subset of the training data. If the original dataset contains 800 samples, and each base tree is trained on a random subset that contains, on average, 70% of the original samples, how many samples, on average, are included in each base tree's training set?

    (a) 400 samples
    (b) 480 samples
    (c) 560 samples
    (d) 600 samples

Given that the original dataset contains 800 samples, and each base tree is trained on an average of 70% of the original samples, the average number of samples included in each base tree's training set can be calculated as follows:
Number of samples in each base tree's training set = 800 * 0.70 = 560 samples

***** End of the Paper *****