Dhirubhai Ambani Institute of
Information and Communication Technology

Surprise Quiz II

IT496: Introduction to Data Mining
Date: Oct 27, 2023 | Timings: 11:15 AM – 11:45 AM

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. Suppose you have a training set with millions of features; which Linear Regression training algorithm can you use for prediction?

(a) Linear Regression with Normal Equation
(b) Linear Regression with Stochastic Gradient Descent
(c) Linear Regression with Batch Gradient Descent
(d) Polynomial Regression with Stochastic Gradient Descent

Gradient Descent's complexity depends on the number of instances, not the number of features, while Normal Equation and SVD's complexity grows more than quadratically with the number of features.

2. Suppose the features in your training set have very different scales. Which of the following algorithm(s) will work fine in this case?

(a) Linear Regression with Normal Equation
(b) Linear Regression with Stochastic Gradient Descent
(c) Linear Regression with Lasso Regularization
(d) Linear Regression with Ridge Regularization

Normal equation and SVD will work fine without scaling (taking inverse involves dividing with the determinant of the matrix, which somewhat scales the values).

3. Can Gradient Descent get stuck in a local minimum when training a Logistic Regression model?

   (a) Yes, it generally happens
   (b) Yes, but it rarely happens
   (c) No, it does not happen
   (d) None of the above

No, it can't, as the loss function of the logistic regression (cross-entropy) is convex.

4. Which of the following statements is (are) correct about SVM's sensitivity to feature scales?

   (a) `fit` and `transform` the `StandardScaler` on training, validation, and test sets individually
   (b) `fit` and `transform` the `StandardScaler` on training and validation, and just `transform` on the test sets
   (c) `fit` and `transform` the `StandardScaler` on training and just `transform` on the validation and test sets
   (d) `fit` and `transform` the `StandardScaler` on training, validation, and test sets all-together

We should use the mean and standard deviation of the training set that includes the validation set.

5. Suppose you have a training set with millions of instances and hundreds of features. Which of the following is (are) correct in this context?

   (a) You should train a model using a primal form of the SVM problem.
   (b) You should train a model using a dual form of the SVM problem.
   (c) You can use any of the two forms of the SVM problem to train a model.
   (d) You should not use SVM in this case.

The computational complexity of the primal form of SVM is proportional to the number of training instances m, while the dual form is proportional to a number between m^2 and m^3.

6. What is the appropriate depth of a Decision Tree trained (without restrictions) on a training set with one million instances?

   (a) ~10 (b) ~20
   (c) ~30 (d) ~40

$\log_2 10^6 = 6 \times \log_2 10 \approx 20$ (a bit more, as the tree will not be perfectly balanced).

7. Is a node's Gini impurity generally lower or greater than its parents?

(a) always greater
(b) always lower
(c) generally lower
(d) generally greater

The CART's cost function splits each node in a way that minimizes the weighted sum of its children's Gini impurities.

---

8. If you train five models, each of them gives 95% precision. Which of the following will be considered the best ensemble out of these individual models?

(a) Five completely different models with the same training data
(b) Five different models of the same learner algorithm with the same training data
(c) Five different models of the same learner algorithm with different training batches
(d) Five completely different models with different training batches

That's the whole point of bagging and pasting ensembles, i.e., diversity of models.

---

9. Which of the following scenarios is problematic for decision trees in machine learning?

(a) Imbalanced class distribution
(b) Continuous and highly correlated features
(c) Small training dataset
(d) Low tree depth

They might make suboptimal splits and become overly complex. Preprocessing, like feature selection and dimensionality reduction, is often necessary in such situations to improve decision tree performance.

---

10. In AdaBoost, what happens if the base classifier's error rate is greater than 0.5 (i.e., worse than random guessing)?

(a) AdaBoost cannot work with classifiers with error rates greater than 0.5.
(b) AdaBoost still works, but the contribution of the classifier to the final ensemble is negatively weighted.
(c) AdaBoost always works perfectly regardless of the base classifier's error rate.
(d) AdaBoost throws an error and cannot be used with such a classifier.

AdaBoost is robust to this scenario because it will focus on the training instances that are misclassified by the base classifier, attempting to correct its mistakes.

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