Question Set | ML

Tamojit Maiti
Masters in Applied Statistics and Operations Research, ISI Kolkata
Data Scientist at Sixt R&D, previously at Rapido & AB InBev

Agenda

- Feedback about the test
- Question Discussions

Test Feedback

Question

In AdaBoost, we re-weight points, giving more weight to the misclassified data points. If we introduced an upper limit on the weight that any point can take, which among the following would be an effect of such a modification?

- It makes the final classifier robust to outliers.
- We may observe the performance of the classifier reduce as the number of stages increases.
- It may result in lower overall performance.
- None of the above

Question

Which of the following are advantages of the K-Means clustering algorithm?

(Note: More than one option may be correct)

- K-Means algorithm is easy to implement and comprehend.
- K-Means algorithm is scale-invariant.
- K-Means algorithm is not sensitive to outliers.
- K-Means algorithm always converges on a data set.

Question

Which of the following are advantages of the K-Means clustering algorithm?

(Note: More than one option may be correct)

- K-Means algorithm is easy to implement and comprehend.
- K-Means algorithm is scale-invariant.
- K-Means algorithm is not sensitive to outliers.
- K-Means algorithm always converges on a data set.

Question

Which of the following is NOT a mandatory data preprocessing step for the XGBOOST algorithm?

- Treating missing values.
- Removing extreme correlated predictors.
- One-hot encoding or dummy variable creation of categorical predictors.
- None of the above.

Question

Which of the following statements is true regarding the XGBoost method?

- In XGBoost, models are added sequentially to correct the errors made by the previous models.
- In XGBoost, models are multiplied sequentially to correct the errors made by the the previous weights.
- XGBoost is an extended version of gradient boosting, where it uses regularization and parallel computing to tune the model and find the best fit.
- All the above

Question

Which of the following options is NOT true with respect to random forest?

- In a random forest, all of the trees are independent of each other.
- Each tree trains on a constant number of randomly selected (with replacement) datapoints.
- Each tree trains on a constant number of randomly selected features.
- The individual trees communicate with each other during the training process.

Question

Which of the following statements is true about standardization and min-max scaling?

- Standardization brings data to a range of 0 to 1, while min-max scaling does the same to a range of -1 to 1.
- Standardization brings data to a range of -1 to 1, while min-max scaling does the same to a range of 0 to 1.
- Both standardization and min-max bring data to a range of 0 to 1.
- Only min-max scaling brings data to a range of 0 to 1.

Question

What happens when the error terms in a linear regression model are not normally distributed?

- The p-values obtained during the hypothesis test to determine the significance of the coefficients become unreliable.
- It becomes difficult to draw certain conclusions from the model if the error terms are not normally distributed.
- Neither a nor b
- Both a and b

Question

When calculating the Log Odds for a logistic regression model, let's say that the equation for the log odds was found to be

$$\ln(\frac{P}{1-P}) = -13.5 + 0.06x$$

For $x = 220$, the log odds are equal to -0.3. For $x = 231.5$, log odds will be equal to

- 0.4
- 0.3
- -0.3
- 0

Question

A classification model is used to identify patients with a certain disease that does not transmit, is not deadly if untreated but has a painful treatment. The model must ensure that the people marked as infected are actually infected. The most important metric to track for this model will be _______.

- Sensitivity
- Specificity
- Precision
- False Positive Rate

Question

Which of the following features for the random forest algorithm increase the randomness in the individual tree models and ultimately reduce the variance of the model?

- Bootstrap sampling
- Aggregation
- Random feature selection
- All of the above

Question

The primary issue with models having high variance is

- Model produced is too sensitive to the input data – it will also learn what shouldn’t be learned from the data given.
- Model becomes too complex to deal with.
- Takes a lot of time and computational effort to train such models.
- All of the above

Question

Let’s say that you are representing a vector P using [100 35] using a new set of basis vectors [1 a] and [b 5] where a and b are unknowns.

In this new representation, that vector in the original standard basis is now written as [30 10].

From this information, find a and b.

- A = 2, b = 7
- A = 1, b = 5
- A = 1, b = 7
- None of the above

Question

Which of the following statements is NOT why Principal Component Analysis (PCA) should be used?

- PCA should be used when the attributes of the data set are highly correlated.
- PCA should be used when a lesser number of dimensions or variables should be used.
- PCA should be used to do better data visualization, as there will be a lesser number of dimensions/features.
- None of the above

Question

Let g(x) = $x^2 + e^{-x}$. Which direction should we move to minimize g(x) at a point with the x-coordinate x?

- $e^{-x} - 2x$
- $e^{-x} + 2x$
- $2x - e^{-x}$
- $2x - e^{x}$

Question

In an election, ten candidates are competing and voters must vote for one of the ten candidates. Voters communicate with each other while casting their votes.

If you consider each voter as an individual model, which of the below techniques does this election procedure represent?

- Bagging
- Boosting
- Both bagging and boosting
- None of the above

Question

Let's say that you are building a telecom churn prediction model with the business objective that your company wants to implement an aggressive customer retention campaign to retain the high churn-risk customers. This is because a competitor has launched extremely low-cost mobile plans, and you want to avoid churn as much as possible by incentivizing the customers. Assume that budget is not a constraint.

If the odds of churn for a customer are equal to 1/3, then that means

- The probability of the customer not churning is 3 times the probability of the customer churning.
- The probability of the customer churning is 3 times more than the probability of the customer not churning.
- The probability of the customer not churning is 4 times the probability of the customer churning.
- The probability of the customer churning is 4 times more than the probability of the customer not churning.

Question

Select all that apply with regard to OOB or Out of Bag Error. (More than one option may be correct)

- All the data points considered while calculating OOB belong to the training set.
- The data points considered while calculating OOB error can be from both train and test sets.
- All data point in the training set is considered for all the trees in the random forest while calculating the OOB error.
- A data point p is considered OOB for only those trees which did not have this data point p in their training set.

Question

Which of the following effects are caused by increasing the value of the regularization coefficient in the case of lasso regression?

- Shrinking the magnitude of coefficients for uninformative features.
- Increasing the gap between train and test error.
- Making the model more complex.
- Selecting features by making coefficients of uninformative features 0.

Question

Which of the following is correct regarding bagging & boosting?

- Bagging is known to reduce variance.
- Boosting models can be trained faster in parallel.
- Boosting algorithm trains k-th classifier sequentially by minimizing the residual error from (k-1)-th classifier.
- Bagging models can be trained faster in parallel.

Question

Which of the following describes an advantage of the random forest (RF) algorithm over the decision tree (DT) algorithm?

- An RF model has better interpretability as compared to a DT.
- An optimally fit RF model has less variance than an optimally fit DT.
- An RF model is usually more complex than a DT.
- An optimally fit RF model has more variance than an optimally fit DT.