Université Gustave Eiffel

Épreuve écrite - CONCOURS EXTERNE D'INGÉNIEUR DE RECHERCHE N° 2020 – IR – CE – 01– Partie 1 (QCM) (5 points)

BAREME :

- *0.5 point par question. Plusieurs réponses peuvent être correctes pour une même question.*
- *Pour avoir juste à une question, toutes les réponses correctes doivent être cochées **ET** aucune réponse incorrecte ne doit être cochée.*

Q1- A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college.

Which of the following statement is true in following case?

a) Feature F1 is an example of nominal variable.
b) Feature F1 is an example of ordinal variable.
c) It doesn't belong to any of the above category.
d) Both of these

Q2- In analyzing a data set, suppose you have identified multi-collinear features.

Which of the following action(s) would you perform next?.

1. Remove both collinear variables.
2. Instead of removing both variables, we can remove only one variable.
3. Removing correlated variables might lead to loss of information. In order to retain those variables, we can use penalized regression models like ridge or lasso regression.

a) Only 1
b) Only 2
c) Only 3
d) Either 1 or 3
e) Either 2 or 3

Q3- Which of the following is an example of a deterministic algorithm:

a) PCA
b) *K-means*
c) None of the above

Q4- Which of the following statement(s) is / are true for Gradient Decent (GD) and Stochastic Gradient Decent (SGD)?

1. In GD and SGD, you update a set of parameters in an iterative manner to minimize the error function.
2. In SGD, you have to run through all the samples in your training set for a single update of a parameter in each iteration.
3. In GD, you either use the entire data or a subset of training data to update a parameter in each iteration.

a) Only 1
b) Only 2
c) Only 3
d) 1 and 2
e) 2 and 3
f) 1, 2 and 3

Université Gustave Eiffel

Q5- Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?

1. Number of Trees
2. Depth of Tree
3. Learning Rate

a) Only 1
b) Only 2
c) Only 3
d) 1 and 2
e) 2 and 3
f) 1,2 and 3

Q6- Imagine, you are solving a classification problems with highly imbalanced class. The majority class is observed 99% of times in the training data.

Your model has 99% accuracy after taking the predictions on test data. Which of the following is true in such a case?

1. Accuracy metric is not a good idea for imbalanced class problems.
2. Accuracy metric is a good idea for imbalanced class problems.
3. Precision and recall metrics are good for imbalanced class problems.
4. Precision and recall metrics aren't good for imbalanced class problems.

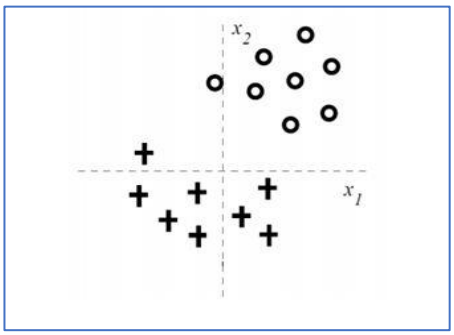
a) 1 and 3
b) 1 and 4
c) 2 and 3
d) 2 and 4

Q7- Adding a non-important feature to a linear regression model may result in.?

1. Increase in R-square
2. Decrease in R-square

a) Only 1 is correct
b) Only 2 is correct
c) Either 1 or 2
d) None of these

Q8- Suppose you are given the below data and you want to apply a logistic regression model for classifying it in two given classes.



You are using logistic regression with L1 regularization.

$$\sum_{i=1}^{n} \log P(y_i|x_i, w_0, w_1, w_2) - C(|w_1| + |w_2|).$$

Which of the following option is correct when you increase the value of C from zero to a very large value?

a) First w2 becomes zero and then w1 becomes zero
b) First w1 becomes zero and then w2 becomes zero
c) Both becomes zero at the same time
d) Both cannot be zero even after very large value of C

Q9- _____ is general-purpose computing model and runtime system for distributed data analytics.

a) Hadoop
b) Sparks
c) Flume
d) None of the mentioned

Q10- According to analysts, for what can traditional IT systems provide a foundation when they're integrated with big data technologies like Hadoop?

a) Big data management and data mining
b) Data warehousing and business intelligence
c) Management of Hadoop clusters
d) Collecting and storing unstructured data