MACHINE LEARNING

1) Which of the following in sk-learn library is used for hyper parameter tuning?

Answer: D) All of the above

****
****

2) In which of the below ensemble techniques trees are trained in parallel?

Answer: A) Random forest

****
****

3) In machine learning, if in the below line of code: *sklearn.svm.**SVC** (C=1.0, kernel='rbf', degree=3),* we increasing the C hyper parameter, what will happen?

Answer: B) The regularization will decrease

****
****

4) Check the below line of code and answer the following questions: *sklearn.tree.**DecisionTreeClassifier**(*criterion='gini',splitter='best',max_depth=None, min_samples_split=2)*

Answer: C) both A & B

****
****

5) Which of the following is true regarding Random Forests?

Answer: A) It's an ensemble of weak learners.

****
****

6) What can be the disadvantage if the learning rate is very high in gradient descent?

Answer: C) Both of them

****
****

7) As the model complexity increases, what will happen?

Answer: B) Bias will decrease, Variance increase

****
****

8) Suppose I have a linear regression model which is performing as follows: Train accuracy=0.95 and Test accuracy=0.75

Answer: B) model is overfitting

****
****

Q9 to Q15 are subjective answer type questions, Answer them briefly.

9) Suppose we have a dataset which have two classes A and B. The percentage of class A is 40% and percentage of class B is 60%. Calculate the Gini index and entropy of the dataset.

Answer: **Gini** = 1 - (0.40^2 + 0.60^2)

= 1 - (0.16+0.36)

= 1 – (0.52)

= 0.48

Entropy = - [0.4 * log2(0.4) + 0.6 * log2(0.6)]

= - [0.4 * -1.32192809489 + 0.6 * -0.736965594166]

= 0.97

****
****

10) What are the advantages of Random Forests over Decision Tree?

Answer: The advantages of Random Forests over Decision Tree are:

a. It reduces overfitting in decision trees and helps to improve the accuracy.

b. It works well with both categorical and continuous values.

c. It automates missing values present in the data.

d. Normalizing of data is not required as it uses a rule-based approach.

e. It is Robust to outliers.

f. It Works well with non-linear data.

g. It runs efficiently on a large dataset.

h. Better accuracy than Decision tree.

****
****

11) What is the need of scaling all numerical features in a dataset? Name any two techniques used for scaling.

Answer: Since the features have different scales, there is a chance that higher weightage is given to features with higher magnitude. This will impact the performance of the machine learning algorithm and obviously, We do not want our algorithm to be biased towards one feature. We scale our data before employing a distance based algorithm so that all the features contribute equally to the result.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Standardization is another scaling technique where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

** **

12) Write down some advantages which scaling provides in optimization using gradient descent algorithm.

Answer: Advantages of scaling in optimization using gradient descent algorithm are:

1. It makes the training faster. It prevents the optimization from getting stuck in local optima.

2. It gives a better error surface shape.

3. Weight decay and Bayes optimization can be done more conveniently.

4. It's also important to apply feature scaling if regularization is used as part of the loss function so that coefficients are penalized appropriately

** **

13) In case of a highly imbalanced dataset for a classification problem, is accuracy a good metric to measure the performance of the model. If not, why?

Answer: Accuracy is not a good metric for imbalanced datasets. Say we have an imbalanced dataset and a badly performing model which always predicts for the majority class. This model would receive a very good accuracy score as it predicted correctly for the majority of observations, but this hides the true performance of the model which is objectively not good as it only predicts for one class.

** **

14) What is “f-score" metric? Write its mathematical formula.

Answer: The F-score, also called the F1-score, is a measure of a model’s accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into ‘Positive’ or ‘Negative’. It is calculated from the precision and recall of the test, where the precision is the number of true positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of true positive results divided by the number of all samples that should have been identified as positive.

F score = 2 * (precision * recall) / (precision + recall)

****
****

15) What is the difference between fit(), transform() and fit_transform()?

Answer:

Fit() - In the fit() method, we use the required formula and perform the calculation on the feature values of input data and fit this calculation to the transformer. For applying the fit() method we have to use .fit() in front of the transformer object.

Transform() – For changing the data we probably do transform, in the transform() method, we apply the calculations that we have calculated in fit() to every data point in feature F. We have to use .transform() in front of a fit object because we transform the fit calculations. We use the example that is used above section when we create an object of the fit method then we just put it in front of the .transform and transform method uses those calculations to transform the scale of the data points, and the output will we get is always in the form of sparse matrix or array.

Fit_transform() - This fit_transform() method is basically the combination of fit method and transform method, it is equivalent to fit().transform(). This method performs fit and transform on the input data at a single time and converts the data points. If we use fit and transform separate when we need both then it will decrease the efficiency of the model so we use fit_transform() which will do both the work.

****
****