Machine Learning Worksheet 7

Q1-Which of the following in sk-learn library is used for hyper parameter tuning?
Answer-D) All of the above

Q2-In which of the below ensemble techniques trees are trained in parallel?
Answer-A) Random forest

Q3-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 decreas

Q4-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) Which of the following is true regarding max_depth hyper parameter?
Answer-It regularizes the decision tree by limiting the maximum depth up to which a tree can be grown.

Q5-Which of the following is true regarding Random Forests
Answer-D)None of the above

Q6-What can be the disadvantage if the learning rate is very high in gradient descent
Answer-C) Both of them

Q7-As the model complexity increases, what will happen?
Answer-B) Bias will decrease, Variance increase

Q8-Suppose I have a linear regression model which is performing as follows: Train accuracy=0.95 and Test accuracy=0.75 Which of the following is true regarding the model?
Answer-B) model is overfitting

Q9-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 index = 1- (p(A)2 + p(B)2)
GI=1-(.4*.4)-(0.6*0.6)
GI=1-0.52
GI=0.48

Entropy= -(p(A) log2(p(A)) + p(B) log2(p(B)))
Entropy= -(0.4*log2(0.4) +0.6*log2(0.6))
Entropy= -(0.4 *(-1.3219)+0.6*(-0.7369)
Entropy= -(- 0.52877123795 - 0.4421793565))
Entropy=0.9709

Q10-What are the advantages of Random Forests over Decision Tree?
Answer-A classification algorithm consisting of many decision trees combined to get a more accurate result as compared to a single tree.Random forest algorithm avoids and prevents overfitting by using multiple trees. This gives accurate and precise results.

Q11-What is the need of scaling all numerical features in a dataset? Name any two techniques used for scaling

Answer-Feature Scaling is a method to transform the numeric features in a dataset to a standard range so that the performance of the machine learning algorithm improves.
The most common techniques of feature scaling are **Normalization and Standardization. Normalization** is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1].
While Standardization transforms the data to have zero mean and a variance of 1, they make our data unitless.

Q12-Write down some advantages which scaling provides in optimization using gradient descent algorithm?
Answer-Having features on a similar scale will help the gradient descent converge more quickly towards the minima. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features we scale the data

---

Q13-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.

Q14-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'.

$$F1\_Score=2*P*R/(P+R)$$

Where,
P=precision,
R=recall

Q15-What is the difference between fit(), transform() and fit_transform()
Answer- **fit()** : In the fit() method, where 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 (fit transform in python), we have to use fit() in front of the transformer object.
transform() : For changing the data, we probably do transform in the transform() method, where 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.

fit_transform():The fit_transform() method is basically the combination of the fit method and the transform method. This method simultaneously performs fit and transform operations on the input data and converts the data points.Using fit and transform separately when we need them both decreases the efficiency of the model. Instead, fit_transform() is used to get both works done.

- **The fit()** method helps in fitting the training dataset into an estimator (ML algorithms).
- **The transform()** helps in transforming the data into a more suitable form for the model.
- **The fit_transform()** method combines the functionalities of both fit() and transform().