MACHINE LEARNING

Note: correct answer highlighted

1. Which of the following in sk-learn library is used for hyper parameter tuning?
   A) GridSearchCV()
   B) RandomizedCV()
   C) K-fold Cross Validation
   D) All of the above

2. In which of the below ensemble techniques trees are trained in parallel?
   A) Random forest
   B) Adaboost
   C) Gradient Boosting
   D) All of the above

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?
   A) The regularization will increase
   B) The regularization will decrease
   C) No effect on regularization
   D) kernel will be changed to linear

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)*
   Which of the following is true regarding max_depth hyper parameter?
   A) It regularizes the decision tree by limiting the maximum depth up to which a tree can be grown.
   B) It denotes the number of children a node can have.
   C) both A & B
   D) None of the above

5. Which of the following is true regarding Random Forests?
   A) It's an ensemble of weak learners.
   B) The component trees are trained in series
   C) In case of classification problem, the prediction is made by taking mode of the class labels predicted by the component trees.
   D) None of the above

6. What can be the disadvantage if the learning rate is very high in gradient descent?
   A) Gradient Descent algorithm can diverge from the optimal solution.
   B) Gradient Descent algorithm can keep oscillating around the optimal solution and may not settle.
   C) Both of them
   D) None of them

7. As the model complexity increases, what will happen?
   A) Bias will increase, Variance decrease
   B) Bias will decrease, Variance increase
   C) both bias and variance increase
   D) Both bias and variance decrease.

8. 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?
   A) model is underfitting
   B) model is overfitting
   C) model is performing good
   D) None of the above

MACHINE LEARNING

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.

Impurity - Entropy & Gini

There are three commonly used impurity measures used in binary decision trees: **Entropy**, **Gini index**, and **Classification Error**.

Entropy (a way to measure impurity):

Entropy=$-\sum .4\log_2(.4) - \sum .6\log_2(.6)$

Entropy=$-\sum .4 * -1.321 - \sum .6*-.737$

Entropy=$-(-.5284) -(-.4422)$

Entropy=$-.9706$

Gini index (a criterion to minimize the probability of misclassification):

Gini=$1-\sum p_j^2$

Gini=$1-\sum(pi)2$

Gini = $1-\sum(.4)^2 + (.6)^2$

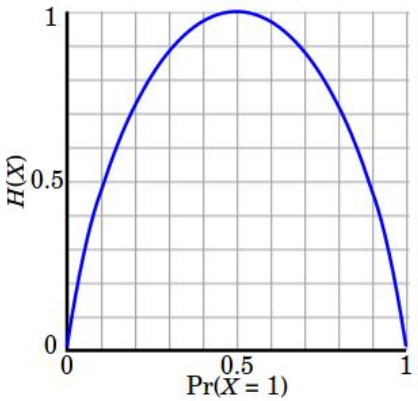
Gini = $1-\sum(.16 + .36)$

Gini = $1-\sum(.52)$

Gini = 1-.52

Gini = .48

The entropy is 0 if all samples of a node belong to the same class, and the entropy is maximal if we have a uniform class distribution. In other words, the entropy of a node (consist of single class) is zero because the probability is 1 and log (1) = 0. Entropy reaches maximum value when all classes in the node have equal probability.



MACHINE LEARNING

Similar to entropy, the **Gini index** is maximal if the classes are perfectly mixed, for example, in a binary class:

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

Comparison basis	Decision Tree	Random Forest
Speed	It is fast	It is slow
Interpretation	It is easy to interpret	It is quite complex to interpret
Time	Takes less time	Takes more time
Linear problems	It is best to build solutions for linear patterns of data	It cannot handle data with linear patterns
Overfitting	There is a possibility of overfitting of data	There is a reduced risk of overfitting, because of the multiple trees
Computation	It has less computation	It has more computation
Visualization	Visualization is quite simple	Visualization is quite complex
Outliers	Highly prone to being affected by outliers	Much less likely to be affected by outliers
Implementation	Rapid model building as it fits the dataset quickly.	Slow to build the model depending on the size of the dataset
Accuracy	It gives less accurate results	It gives more accurate results

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

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values

Techniques to perform Feature Scaling

Consider the two most important ones:

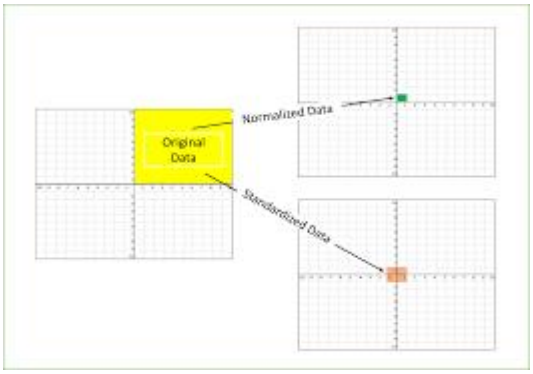
- **Min-Max Normalization:** This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{\text{new}} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- **Standardization:** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{\text{new}} = \frac{X_i - X_{\text{mean}}}{\text{Standard Deviation}}$$

MACHINE LEARNING



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

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 before feeding it to the model. Having features on a similar scale can help the gradient descent converge more quickly towards the minima

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?

When working with imbalanced data, The minority class is our interest most of the time. Like when detecting "spam" emails, they number quite a few compared to "not spam" emails. So, **the machine learning algorithms favor the larger class and sometimes even ignore the smaller class if the data is highly imbalanced**

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

An F-score is the harmonic mean of a system's precision and recall values. It can be calculated by the following formula: **2 x [(Precision x Recall) / (Precision + Recall)]**.

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

A. The **fit(data)** method is used to compute the mean and std dev for a given feature to be used further for scaling.

B. The **transform(data)** method is used to perform scaling using mean and std dev calculated using the .fit() method.

C. The **fit_transform()** method does both fits and transform.