MACHINE LEARNING

(ASSIGNMENT – Worksheet7 Answers – Vivek Kumar Sahu – Internship 35)

(Marked answers in Bold)

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

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.

Ans.Calculate the Gini index and entropy of the dataset, we need to know the proportion of each class in the dataset. Let's assume that we have a dataset with 100 samples, where 40 samples belong to class A and 60 samples belong to class B.

The proportion of class A is 40/100 = 0.4, and the proportion of class B is 60/100 = 0.6.

Gini Index:

The Gini index is a measure of the impurity of a node in a decision tree. For a binary classification problem like this one, the Gini index is calculated as:

Gini index = 1 - (proportion of class A)^2 - (proportion of class B)^2

Plugging in the values, we get:

Gini index = 1 - 0.4^2 - 0.6^2

= 1 - 0.16 - 0.36

= 0.48

So the Gini index of the dataset is 0.48.

Entropy:

Entropy is another measure of the impurity of a node in a decision tree. For a binary classification problem, the entropy is calculated as:

Entropy = - (proportion of class A) * log2(proportion of class A) - (proportion of class B) * log2(proportion of class B)

Plugging in the values, we get:

Entropy = - 0.4 * log2(0.4) - 0.6 * log2(0.6)

= - 0.4 * (-1.32) - 0.6 * (-0.74)

= 0.528 + 0.444

= 0.972

So the entropy of the dataset is 0.972.

10. What are the advantages of Random Forests over Decision Tree?
Ans. Random Forests have several advantages over Decision Trees:

1. Reduced overfitting: Random Forests reduce overfitting by using an ensemble of decision trees, where each tree is trained on a random subset of the data and a random subset of the features. This helps to reduce the variance and improve the generalization performance of the model.

2. Better accuracy: Random Forests generally have better accuracy than individual decision trees, especially for large and high-dimensional datasets.

3. Robustness to noise: Random Forests are robust to noise and outliers in the data, as each tree is trained on a different subset of the data.

4. Feature importance: Random Forests can provide an estimate of feature importance, which can be useful for feature selection and understanding the underlying data.

5. Easy to use: Random Forests are easy to use and require minimal hyperparameter tuning compared to other models, making them a good choice for a wide range of applications.

MACHINE LEARNING

6. Overall, Random Forests are a powerful and versatile machine learning model that can provide high accuracy, reduce overfitting, and handle noisy data.

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

Ans. Most of the times, a dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidean distance between two data points in their computations, this is a problem because, the results would vary greatly between different units, 5 km and 5000 m. The features with high magnitudes will weight a lot more in the distance calculations than features with low magnitudes. To avoid this effect, we use scaling which transforms all features to the same level of magnitudes. The two major scaling techniques used now a days are standardization and normalization.

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

Ans. Scaling provides several advantages in optimization using gradient descent algorithm:

1. Faster convergence: Scaling the features can help the algorithm to converge faster to the optimal solution. This is because scaling ensures that the gradients have a similar scale, and hence the algorithm can take larger steps towards the minimum.
2. Better performance: Scaling can improve the performance of the algorithm by ensuring that all the features have a similar impact on the outcome. This is especially important when the features have different scales, as larger scale features can dominate the optimization process.
3. Avoiding numerical errors: Scaling can prevent numerical errors that can occur when the features have a large range of values. For example, when the features have very large or very small values, the algorithm may encounter numerical instability or underflow/overflow errors.

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?

Ans. In case of a highly imbalanced dataset for a classification problem, accuracy may not be a good metric to measure the performance of the model. This is because accuracy can be misleading when the classes are imbalanced, as it can give a false impression of high performance even when the model is not performing well.

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

Ans. The F-score, also known as the F1 score, is an evaluation metric that combines precision and recall into a single score. It is commonly used to evaluate the performance of a binary classification model.

The mathematical formula for F-score is as follows:

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

where Precision is the proportion of true positive predictions among all positive predictions, and Recall is the proportion of true positive predictions among all actual positive samples.

Precision = True Positive / (True Positive + False Positive)

Recall = True Positive / (True Positive + False Negative)

The F1 score is the harmonic mean of precision and recall. It ranges from 0 to 1, where a score of 1 indicates perfect precision and recall, and a score of 0 indicates the worst possible performance. A

MACHINE LEARNING

high F1 score indicates a model with both high precision and high recall, which is desirable in most classification problems.

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

Ans. fit(): The fit() method is used to estimate the parameters of a model on the training data. In other words, it is used to "fit" the model to the training data. It takes the input data as its argument and learns the parameters of the model based on that data.

transform(): The transform() method is used to apply the learned transformation to a new set of data. It takes the input data as its argument and applies the transformation that was learned during the fit() method to that data. The transform() method does not change the parameters of the model, it only applies the learned transformation to the new data.

fit_transform(): The fit_transform() method is a combination of fit() and transform() methods. It is used to learn the parameters of a transformation on the training data and apply the transformation to that data in a single step. It takes the input data as its argument, learns the parameters of the transformation based on that data using the fit() method, and then applies the learned transformation to the data using the transform() method.

MACHINE LEARNING