Worksheet Set-7 Machine Learning

1. Which of the following in sk-learn library is used for hyper parameter tuning?
   A) GridSearchCV()
2. In which of the below ensemble techniques trees are trained in parallel?
   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?
   B) The regularization will decrease4.
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?
   C) It denotes the number of children a node can have.
5. Which of the following is true regarding Random Forests?
   D)None of the above
6. What can be the disadvantage if the learning rate is very high in gradient descent?
   C) Both of them
7. As the model complexity increases, what will happen?
   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 Which of the following is true regarding the model?

B) model is overfitting

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

Gini index=0.48
Entropy =0.31

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

Ans. Advantages are as follow:

- It reduces overfitting in decision trees and helps to improve the accuracy.
- It works well with both categorical and continuous values.
- It automates missing values present in the data.
- Normalizing of data is not required as it uses a rule-based approach.
- It is Robust to outliers.
- It Works well with non-linear data.
- It runs efficiently on a large dataset.
- 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?

Ans. Feature scaling is essential for machine learning algorithms that calculate distances between data. If not scale, the feature with a higher value range starts dominating when calculating distances.

Two techniques used for scaling:-

1. Min Max Scaler
2. Standard Scaler

Worksheet Set-7 Machine Learning

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

Ans. The advantages are as follow:

- It makes the training faster.
- It prevents the optimization from getting stuck in local optima.
- It gives a better error surface shape.
- Weight decay and Bayes optimization can be done more conveniently.
- 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?

Ans. In case of a highly imbalanced dataset for a classification problem accuracy is not at all good metric to measure the performance of the model.
Achieving 90 percent classification accuracy, or even 99 percent classification accuracy, may be trivial on an imbalanced classification problem.
This means that intuitions for classification accuracy developed on balanced class distributions will be applied and will be wrong, misleading the practitioner into thinking that a model has good or even excellent performance when it, in fact, does not.

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

Ans. In statistical analysis of binary classification, the F-score or F-measure is a measure of a test's accuracy. 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()?

Ans. 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 we have to use .fit() in front of the transformer object. Suppose we initialize the StandardScaler object O and we do .fit() then what will it do that, it takes the feature F and it will just compute the mean (μ) and standard deviation (σ) of feature F. That has happened in the fit method.
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.
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.

Suppose, we create the StandarScaler object, and then we perform .fit_transform() then it will calculate the mean(μ) and standard deviation(σ) of the feature F at a time it will transform the data points of the feature F.

Worksheet Set-7 SQL

1. The primary key is selected from the
   Ans. B) Candidate keys
2. Which is/are correct statements about primary key of a table?
   Ans. B. Primary keys cannot contain NULL values… & C. A table can have only one primary key with single or multiple fields….
3. Which one of the following sorts rows in SQL?
   Ans. C. ORDERBY
4. The SQL statement that queries or reads data from a table is
   Ans. C. SELECT
5. Which SQL command is used to insert a row in a table?
   Ans. C. Insert
6. Which normal form is considered adequate for relational database design?
   Ans. C. 3NF
7. SQL can be used to
   Ans. A. Create database structures only
8. SQL query and modification commands make up
   Ans. B. DML
9. The result of a SQL SELECT statement is a(n).
   Ans. B. Table
10. Second normal form should meet all the rules for
    Ans. A. 1 NF
11. What are joins in SQL?
    Ans. SQL Join statement is used to combine data or rows from two or more tables based on a common field between them.
12. What are the different types of joins in SQL?
    Ans. Different types of joins are:
    - **INNER JOIN-** The INNER JOIN keyword selects all rows from both the tables as long as the condition is satisfied. This keyword will create the result-set by combining all rows from both the tables where the condition satisfies i.e value of the common field will be the same.
    - **LEFT JOIN-** This join returns all the rows of the table on the left side of the join and matches rows for the table on the right side of the join. For the rows for which there is no matching row on the right side, the result-set will contain *null*. LEFT JOIN is also known as LEFT OUTER JOIN.
    - **RIGHT JOIN-** RIGHT JOIN is similar to LEFT JOIN. This join returns all the rows of the table on the right side of the join and matching rows for the table on the left side of the join. For the rows for which there is no matching row on the left side, the result-set will contain *null*. RIGHT JOIN is also known as RIGHT OUTER JOIN.

- **FULL JOIN-** FULL JOIN creates the result-set by combining results of both LEFT JOIN and RIGHT JOIN. The result-set will contain all the rows from both tables. For the rows for which there is no matching, the result-set will contain *NULL* values.

13. What is primary key in SQL?
    Ans. In SQL, a Primary Key is a **special relational database table field** or a combination of fields that uniquely identifies a record in the table of multiple records. The main feature of the primary key is, it holds a unique value for each row of table data in the database.
14. What is SQL Server?
    Ans. SQL Server is an application software for Relational Database Management System (RDBMS), from Microsoft, that can be used for creating, maintaining, managing, and implementing the RDBMS systems. It is an extensively used application as it enables multiple users simultaneously to work on the database systems, where users can range from minor office-based machines to huge Internet-based servers. Provisions any variety of SQL programming extending from ANSI SQL (for traditional SQL) through SQL to T-SQL (Transact-SQL) used for advanced relational databases.
15. What is ETL in SQL?
    Ans. ETL stands for Extract, Transform and Load, which is a process used to collect data from various sources, transform the data depending on business rules/needs and load the data into a destination database.

Worksheet Set-7 Statistics

1. A die is thrown 1402 times. The frequencies for the outcomes 1, 2, 3, 4, 5 and 6 are given in the following table:
   Find the probability of getting 6 as outcome:
   Ans. b) 0.135
2. A telephone directory page has 400 telephone numbers. The frequency distribution of their unit place digit (for example, in the number 25827689, the unit place digit is 9 is given in table below: First row refers to the digits Second row to their frequencies.
   What will be the probability of getting a digit with unit place digit odd number that is 1, 3,5,7,9?
   Ans. d) 0.53
3. A tyre manufacturing company which keeps a record of the distance covered before a tyre needed to be replaced. The table below shows the results of 1100 cases.
   Ans. c) 0.745
4. Please refer to the case and table given in the question No. 3 and determine what is the probability that if we buy a new tyre then it will last in the interval [4000-14000] miles?
   Ans. b) 0.577
5. We have a box containing cards numbered from 0 to 9. We draw a card randomly from the box. If it is told to you that the card drawn is greater than 4 what is the probability that the card is odd?
   Ans. c) 0.6
6. We have a box containing cards numbered from 1 to 8. We draw a card randomly from the box. If it is told to you that the card drawn is less than 4 what is the probability that the card is even?
   Ans. a) 0.33
7. A die is thrown twice and the sum of the numbers appearing is observed to be 7. What is the conditional probability that the number 6 has appeared at least on one of the die?
   Ans. c) 0.33
8. Consider the experiment of tossing a coin. If the coin shows tail, toss it again but if it shows head, then throw a die. Find the conditional probability of the event that 'the die shows a number greater than 4' given that 'there is at least one Head'.
   Ans. b) 0.22
9. There are three persons Evan, Ross and Michelle. These people lined up randomly for a picture. What is the probability of Ross being at one of the ends of the line?
   Ans. a) 0.66
10. Let us make an assumption that each born child is equally likely to be a boy or a girl. Now suppose, if a family has two children, what is the conditional probability that both are girls given that at least one of them is a girl?
    Ans. a) 0.33
11. Consider the same case as in the question no. 10. It is given that elder child is a boy. What is the conditional probability that both children are boys?
    Ans. c) 0.5
12. We toss a coin. If we get head, we toss a coin again and if we get tail, we throw a die. What is the probability of getting a number greater than 4 on die?
    Ans. a) 0.166
13. We toss a coin. If we get head, we toss a coin again and if we get tail, we throw a die. What is the probability of getting an odd number on die?
    Ans. d) 0.25

Worksheet Set-7 Statistics

14. Suppose we throw two dice together. What is the conditional probability of getting sum of two numbers found on the two dice after throwing is less than 4, provided that the two numbers found on the two die are different?
    Ans. d) 0.06
15. A box contains three coins: two regular coins and one fake two-headed coin, you pick a coin at random and toss it. What is the probability that it lands heads up?
    Ans. c) 1/2