MACHINE LEARNING ASSIGNMENT - 7

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

D) All of the above

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 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) Which of the following is true regarding max_depth hyper parameter?

C) both A & B

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

C) In case of classification problem, the prediction is made by taking mode of the class labels predicted by the component trees.

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?

A) Bias will increase, 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?

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.

Sol :

```
# calculate the entropy for a dataset

from math import log2

# proportion of examples in each class

classA= 40/100
classB = 60/100

# calculate entropy
entropy = -(classA * log2(classA) + classA * log2(classB))
```

entropy= 0.8235574756214273

Gini Index= 1-((4/10)^2 + (6/10)^2) = .48

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

Sol : Decision tree (DT) is non linear ML model and Random forest(RF) is basically bagging technique which is the type of ensemble model. Random forest is basically a set of decision trees formed through an algorithm to classify multi-dimensional feature vectors. So as intuition dictates, a random forest is more powerful than a decision tree for problems that deal with higher dimensional feature vectors. For problems that require fewer dimensions, a decision tree will suffice.

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

Sol : When you collect data and extract features, many times the data is collected on different scales. For example, the age of employees in a company may be between 21-70 years, and their salaries may range from Rs. 30000- 80000. In this situation if you use a simple Euclidean metric, the *age* feature will not play any role because it is several order smaller than other features. Here, you may want to normalize the features independently to the same scale, say [0,1], so they contribute equally while computing the distance. However, normalization may also result in loss of information.

Techniques: Min Max Scaler, Standard Scaler

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?

Sol : Classification accuracy is a metric that summarizes the performance of a classification model as the number of correct predictions divided by the total number of predictions. This intuition breaks down when the distribution of examples to classes is severely skewed. Intuitions developed by practitioners on balanced datasets, such as 99 percent representing a skillful model, can be incorrect and dangerously misleading on imbalanced classification predictive modeling problems.

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

F1 score is harmonic mean of Precision and Recall.

F1=1 / (Precision-1 +Recall-1)
F1=2∗(Precision∗Recall) / (Precision+Recall)

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

fit() : used for generating learning model parameters from training data

transform() : parameters generated from fit() method,applied upon model to generate transformed data set.

fit_transform() : combination of fit() and transform() on same data set

Here the fit method, when applied to the training dataset, learns the model parameters (for example, mean and standard deviation). We then need to apply
the transform method on the training dataset to get the transformed (scaled) training dataset. We could also perform both of these steps in one step by
applying fit_transform on the training dataset.

WORKSHEET 7 SQL

1. The primary key is selected from the B. Candidate keys

2. Which is/are correct statements about primary key of a table?

B. Primary keys cannot contain NULL values… C. A table can have only one primary key with single or multiple fields….

3. Which SQL command is used to insert a row in a table? C. Insert

4. Which one of the following sorts rows in SQL? C. ORDERBY

5. The SQL statement that queries or reads data from a table is C. SELECT

6. Which normal form is considered adequate for relational database design? C. 3NF

7. SQL can be used to C. All of the above can be done by SQL

8. SQL query and modification commands make up B. DML

9. The result of a SQL SELECT statement is a/n

B. Table

10. Second normal form should meet all the rules for B. 2 NF

11. What are joins in SQL?

(INNER) JOIN: Returns records that have matching values in both tables

LEFT (OUTER) JOIN: Returns all records from the left table, and the matched records from the right table

RIGHT (OUTER) JOIN: Returns all records from the right table, and the matched records from the left table

FULL (OUTER) JOIN: Returns all records when there is a match in either left or right table

13. What is SQL Server?

SQL SERVER is a relational database management system (RDBMS) developed by Microsoft. SQL Server supports ANSI SQL, which is the standard SQL (Structured Query Language) language. However, SQL Server comes with its own implementation of the SQL language, T-SQL (Transact-SQL).

14. What is primary key in SQL

The primary key is that specially selected single candidate key (among all other candidate keys of this entity) which is to be used to uniquely identify the current entity in relations.

15. What is ETL in SQL?

E = Extract

pulling data from flat files, RDBMS, maybe even web-scraping and piping to csv

T = Transform

filtering, sorting, aggregating, joining, cleaning, validating, etc

maybe ask Mike West for help on this step

L = Load

put the sqeaky clean data into the RDBMS/table

STATISTICS WORKSHEET-7

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:

Outcome	1	2	3	4	5	6
Frequency	400	300	157	180	175	190

   Find the probability of getting 6 as outcome: **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.

0	1	2	3	4	5	6	7	8	9
44	52	44	44	40	20	28	56	32	40

   What will be the probability of getting a digit with unit place digit odd number that is 1, 3,5,7,9? **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. Distance (miles) 14000 Frequency 20 260 375 445 If we buy a

new tyre of this company, what is the probability that the tyre will last more than 9000 miles? **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? **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? **a) 0.3**
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? **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? **a) 0.45**
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'. **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? a) 0.66 b) 0.45 c) 0.23 d) 0.56
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? **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? **a) 0.33**
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? **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? **d) 0.25**
14. Suppose we throw two dice together. What is the conditional probability of getting sum of two numbers found on the two die after throwing is less than 4, provided that the two numbers found on the two die are different? **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? **b) 2/3**