FLIP ROBO

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:
a) 0.34
b) 0.135
c) 0.45
d) 0.78

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?
a) 0.67
b) 0.60
c) 0.45
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)	<4000	4000-9000	9001-14000	>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?
a) 0.67
b) 0.459
c) 0.745
d) 0.73

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?

a) 0.56
b) 0.577
c) 0.745
d) 0.73

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) 0.5
b) 0.8
c) 0.6
d) 0.7

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
b) 0.40
c) 0.56
d) 0.89

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
b) 0.37
c) 0.33
d) 0.89

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'.

a) 0.1
b) 0.22
c) 0.38
d) 0.45

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
b) 0.45
c) 0.56
d) 0.26

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
b) 0.23
c) 0.5
d) 0.76

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
b) 0.34
c) 0.78
d) 0.34

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?

a) 0.345
b) 0.79
c) 0.2
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?

a) 0.3
b) 0.56
c) 0.24
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 tossit. What is the probability that it lands heads up?

a) 1/3
b) 2/3
c) 1/2
d) ¾

MACHINE LEARNING

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 labelspredicted 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

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.

Answer :

Entropy=-P(A)*log(P(A)- P(B)*logP(B)

P(A)=4/10=2/5 ; P(B)=6/10=3/5

Entropy= - 2/5*log(2/5)-

3/5*log(3/5)Entropy=0.9

2) Gini Index: 1-[P(A)^2 +P(B)^2]=1-0.6=0.4

Gini Index=0.4

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

Answer :

1. **Reduction in overfitting**: by averaging several trees, there is a significantly lower risk ofoverfitting.

2. **Less variance**: By using multiple trees, you reduce the chance of stumbling across a classifier that doesn't perform well because of the relationship between the train and test data.

3. High predictive accuracy.

4. Efficient on large datasets

5. Works well with missing data still giving a better predictive accuracy

11. What is the need of scaling all numerical features in a dataset? Name any two techniquesused forscaling.

Answer :

1. The gradient descent algorithm which is used to reach the optimal solution in most of the cases, it reached the optimal solution much faster if all the features are at the same scale. That's why scaling helps to reach the optimal solution.

2· If the features in the training dataset are on different scales, then during training the features with large scales will be favored over there in order to minimize the loss. That's why we do Scaling to puts all the features on the same scale.

Two techniques used for scaling are:
1. Standard Scalar
2. Min – Max Scalar

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

Answer: Scaling input helps gradient decent converge faster

13. **In case of a highly imbalanced dataset for a classification problem, is accuracy a goodmetric tomeasure the performance of the model. If not, why?**

Answer : In the framework of imbalanced data-sets, accuracy is no longer a proper measure, since it does notdistinguish between the numbers of correctly classified examples of different classes. Hence, it may lead to erroneous conclusions

.

14. **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 toevaluate binary classification systems, which classify examples into ‘positive’ or ‘negative’.

Mathematical Formula of the F-score:

F-Score = (2 * Precision * Recall) / (Precision + Recall)

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

Answer :

A) **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 touse **.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.

B) 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

C) 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.

WORKSHEET 7 SQL

Q1 and Q2 have one or more correct answer. Choose all the correct option to answer your question.

1. The primary key is selected from the

   A. Composite keys

   B. Candidate keys

C. Foreign keys
D. Determinants

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

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

Q3 to Q10 have only one correct answer. Choose the correct option to answer your question.

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

    A. Select
    B. Create
    C. Insert
    D. Drop

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

    A. SORTBY
    B. ALIGNBY
    C. ORDERBY
    D. GROUPBY

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

    A. QUERY
    B. READ
    C. SELECT
    D. QUERY

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

    A. 1NF
    B. 2NF
    C. 3NF
    D. 4NF

7. SQL can be used to

    A. Create database structures only
    B. Modify database data only
    C. All of the above can be done by SQL
    D. Query database data only

8. SQL query and modification commands make up
   A. DDL
   B. **DML**
   C. HTML
   D. XML

9. The result of a SQL SELECT statement is a(n).
   A. File
   B. **Table**
   C. Report
   D. F
   or
   m

10. Second normal form should meet all the rules for
   A. **1 NF**
   B. 2 NF
   C. 3 NF
   D. 4 NF

11. What are joins in SQL?

Answer: JOINS in SQL are commands which are used to combine rows from two or more tables,based on a related column between those tables. There are predominantly used when a user is trying to extract data from tables which have one-to-many or many-to-many relationships between them.

12. What are the different types of joins in SQL?

Answer: There are mainly four types of joins that you need to understand. They are:

- INNER JOIN
- FULL JOIN
- LEFT JOIN
- RIGHT JOIN

13. What is SQL Server?

Answer: SQL Server is a relational database management system, or RDBMS, developed and marketed byMicrosoft.

Similar to other RDBMS software, SQL Server is built on top of SQL, a standard programming language for interacting with the relational databases. SQL server is tied to Transact-SQL, or T-SQL, the Microsoft'simplementation of SQL that adds a set of proprietary programming constructs

14. What is primary key in SQL?

Answer:

- The *PRIMARY KEY* constraint uniquely identifies each record in a table.

- Primary keys must contain UNIQUE values, and cannot contain NULL values.

- A table can have only ONE primary key; and in the table, this primary key can consist of single ormultiple columns (field).

15. What is ETL in SQL?

Answer: ETL stands for Extract, Transform and Load. It is the process in which the Data is extracted fromany data sources and transformed into a proper format for storing and future reference purposes. Finally, this data is loaded into the database