PFA WORKSHEET 4 solutions

Machine learning

1. c

2. d

3. c

4. a

5. c

6. b

7. c

8. b,c

9. a,d

10.a,b

11) An *outlier* is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

Any set of data can be described by its five-number summary. These five numbers, which give you the information you need to find patterns and outliers, consist of (in ascending order):

* The minimum or lowest value of the dataset
* The first quartile *Q*1, which represents a quarter of the way through the list of all data
* The median of the data set, which represents the midpoint of the whole list of data
* The third quartile *Q*3, which represents three-quarters of the way through the list of all data
* The maximum or highest value of the data set.

12) Bagging and Boosting are two types of Ensemble Learning. These two decrease the variance of single estimate as they combine several estimates from different models. So the result may be a model with higher stability.

* If the difficulty of the single model is over-fitting, then Bagging is the best option.
* If the problem is that the single model gets a very low performance, Boosting could generate a combined model with lower errors as it optimises the advantages and reduces pitfalls .

13) Every time you add a independent variable to a model, the **R-squared** **increases**, even if the independent variable is insignificant. It never declines. Whereas **Adjusted R-squared** increases only when independent variable is significant and affects dependent variable.

14) **Normalization** typically means rescales the values into a range of [0,1]. **Standardization** typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

15) Cross Validation in Machine Learning is a great technique to deal with overfitting problem in various algorithms. Instead of training our model on one training dataset, we train our model on many datasets. Below are some of the advantages and disadvantages of Cross Validation in Machine Learning:  
  
Advantages of Cross Validation  
  
1. Reduces Overfitting: In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from overfitting the training dataset. So, in this way, the model attains the generalization capabilities which is a good sign of a robust algorithm.  
  
Note: Chances of overfitting are less if the dataset is large. So, Cross Validation may not be required at all in the situation where we have sufficient data available.  
  
2. Hyperparameter Tuning: Cross Validation helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm.  
  
Disadvantages of Cross Validation  
  
1. Increases Training Time: Cross Validation drastically increases the training time. Earlier you had to train your model only on one training set, but with Cross Validation you have to train your model on multiple training sets.   
  
For example, if you go with 5 Fold Cross Validation, you need to do 5 rounds of training each on different 4/5 of available data. And this is for only one choice of hyperparameters. If you have multiple choice of parameters, then the training period will shoot too high.  
  
2. Needs Expensive Computation: Cross Validation is computationally very expensive in terms of processing power required.

Sql

1)SELECT shippedDate, COUNT(orderNumber) AS NoOfOrders FROM Orders

SELECT AVG(NoOfOrders) AS AverageNumOfOrdersShipped

2)SELECT orderDate, COUNT(derNumber) AS NoOfOrders FROM Orders

SELECT AVG(`NoOfOrders`) AS `AverageNumOfOrdersPlaced

3) SELECT productName FROM Products

Where productName=(SELECT MIN(MSRP) from Products);

4)SELECT productName FROM products

Where productName=(SELECT MIN(Stock Quantity) FROM poducts  
5) SELECT productName

FROM OrderDetails AS a INNER JOIN Products AS b ON

a.productCode= b.productCode

GROUP BY b.productCode

ORDER BY COUNT(orderNumber) ASC LIMIT 1:

6) SELECT \* FROM employees SELECT salary in MAX(salary) FROM employees

7) SELECT customerNumber, customerName FROM Customers

WHERE city= Melbourne;

8) SELECT customerName FROM Customers

WHERE customerName REGEXP `^N\*`;

9) SELECT customerName FROM Customers

WHERE phones REGEXP “^7.\*” AND city = “Las Vegas”;

10) SELECT customerNameFROM Customers

WHERE creditLimit< 1000 AND city IN ("Las Vegas", "Nantes", "Stavern");

11) SELECT orderNumber FROM orderDetails

WHERE quantityOrdered< 10;

12) SELECT orderNumber FROM Customers AS a INNER JOIN orders AS b ON a.customerNumber = b.customerNumber WHERE customer Name REGEXP "^B.\*";

13) SELECT customer Name FROM Orders AS a INNER JOIN Customers AS b ONa.customerNumber= b.customerNumber WHERE status= "Disputed";

14) SELECT customer Name FROM Payments INNER JOIN Customers USING (customer Number) WHERE payment Date = "2004-10-19" AND check Number REGEXP "^H.\*";

15) SELECT check Number FROM Payments WHERE amount> 1000;

STATISTICS

1) The Central Limit Theorem is at the core of what every data scientist does daily: make statistical inferences about data.The theorem gives us the ability to quantify the likelihood that our sample will deviate from the population without having to take any new sample to compare it with. We don’t need the characteristics about the whole population to understand the likelihood of our sample being representative of it.

2) Sampling is a method that allows us to get information about the population based on the statistics from a subset of the population (sample), without having to investigate every individual.

There **are** two types of **sampling methods**: Probability **sampling** involves random selection, allowing **you to** make strong statistical inferences about the whole group. Non-probability **sampling** involves non-random selection based on convenience or other criteria, allowing **you to** easily collect data

3) The points given below are substantial so far as the differences between type I and type II error is concerned:

1. Type I error is an error that takes place when the outcome is a rejection of null hypothesis which is, in fact, true. Type II error occurs when the sample results in the acceptance of null hypothesis, which is actually false.
2. Type I error or otherwise known as false positives, in essence, the positive result is equivalent to the refusal of the null hypothesis. In contrast, Type II error is also known as false negatives, i.e. negative result, leads to the acceptance of the null hypothesis.
3. When the null hypothesis is true but mistakenly rejected, it is type I error. As against this, when the null hypothesis is false but erroneously accepted, it is type II error.
4. Type I error tends to assert something that is not really present, i.e. it is a false hit. On the contrary, type II error fails in identifying something, that is present, i.e. it is a miss.
5. The probability of committing type I error is the sample as the level of significance. Conversely, the likelihood of committing type II error is same as the power of the test.
6. Greek letter ‘α’ indicates type I error. Unlike, type II error which is denoted by Greek letter ‘β’.

4) Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bellcurve

5) Covariance and Correlation are very helpful in understanding the relationship between two continuous variables. Covariance tells whether both variables vary in the same direction (positive covariance) or in the opposite direction (negative covariance). There is no meaning of covariance numerical value only sign is useful. Whereas Correlation explains the change in one variable leads how much proportion change in the second variable. Correlation varies between -1 to +1. If the correlation value is 0 then it means there is no Linear Relationship between variables however other functional relationship may exist

6) univariate analysis**:-** provides summary statistics for each field in the raw data set (or) summary only on one variable. *Ex* :- CDF,PDF,Box plot, Violin plot.  
Bivariate analysis:- is performed to find the relationship between each variable in the dataset and the target variable of interest (or) using 2 variables and finding realtionship between them. *Ex* :-Box plot,Voilin plot.

7) A sensitivity analysis determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions. In other words, sensitivity analyses study how various sources of uncertainty in a mathematical model contribute to the model's overall uncertainty. This technique is used within specific boundaries that depend on one or more input variables.

8) Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis.

Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process. The word "population" will be used for both of these cases in the following description

Alternative Hypothesis: **H1**: The hypothesis that we are interested in proving. Null hypothesis: **H0**: The complement of the alternative hypothesis. ... This is the probability of falsely rejecting the null hypothesis. Type II error: do not reject the null hypothesis when it is wrong.

Instead, hypothesis **testing** concerns on how to use a random sample to judge if it is evidence that supports or not the hypothesis. Hypothesis **testing** is formulated in terms of **two** hypotheses: **H0**: the **null hypothesis**; • **H1**: the alternate hypothesis.

9)Quantitative is information about quantities, and therefore numbers, and  **quantitative** is descriptive, and regards phenomenon which can be observed but not measured, such as language.

10) The interquartile range (IQR) is a measure of variability, based on dividing a data set into quartiles. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.

* Q1 is the “middle” value in the first half of the rank-ordered data set.
* Q2 is the median value in the set.
* Q3 is the “middle” value in the second half of the rank-ordered data set

11) The term "bell curve" is used to describe a graphical depiction of a normal probability distribution, whose underlying standard deviations from the mean create the curved bell shape. A standard deviation is a measurement used to quantify the variability of data dispersion, in a set of given values around the mean. The mean, in turn, refers to the average of all data points in the data set or sequence and will be found at the highest point on the bell curve.

12) **Boxplots display asterisks or other symbols on the graph to indicate explicitly when datasets contain outliers. These graphs use the interquartile method with fences to find outliers, which I explain later. The boxplot below displays our example dataset. It’s clear that the outlier is quite different than the typical data value.**

13) The p value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H 0) of a study question is true – the definition of 'extreme' depends on how the hypothesis is being tested.

14) A binomial experiment is one that possesses the following properties:

1. The experiment consists of *n* repeated trials;
2. Each trial results in an outcome that may be classified as a success or a failure (hence the name,binomial);
3. The probability of a success, denoted by *p*, remains constant from trial to trial and repeated trials are independent.

The number of successes *X* in *n* trials of a binomial experiment is called a binomial random vareiable

The probability distribution of the random variable *X* is called a binomial distribution and is given by the formula:

*P*(*X*)=*Cxn* *pxqn*−*x*

15)anovais a statistical technique that is used to check if the means of two or more groups are significantly different from each other. anova checks the impact of one or more factors by comparing the means of different samples. ... Another measure to compare the samples is called a t test.