Worksheet 4 Machine Learning

1. The value of correlation coefficient will always be:

   Ans. C) between -1 and 1

2. Which of the following cannot be used for dimensionality reduction?

   Ans. A) Lasso Regularisation

3. Which of the following is not a kernel in Support Vector Machines?

   Ans. C) hyperplane

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?

   Ans. D) Support Vector Classifier

5. In a Linear Regression problem, ‘X’ is independent variable and ‘Y’ is dependent variable, where ‘X’ represents weight in pounds. If you convert the unit of ‘X’ to kilograms, then new coefficient of ‘X’ will be?

   Ans. B) same as old coefficient of ‘X’

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?

   Ans. C) decreases

7. Which of the following is not an advantage of using random forest instead of decision trees?

   Ans. C) Random Forests are easy to interpret

8. Which of the following are correct about Principal Components?

   Ans. B) & C)

9. Which of the following are applications of clustering?

   Ans. A) & D)

10. Which of the following is(are) hyper parameters of a decision tree?

    Ans. A), B) & D)

11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.

Ans. A data point that varies greatly from other observations is referred to as an outlier. An outlier may be caused by measurement uncertainty or by experimental error, the latter of which is often omitted from the data set. In statistical analyses, an outlier can cause serious problems.

The interquartile range (IQR), also known as the middle 50 percent or midspread, is a statistical dispersion measure that is equal to the gap between the 75th and 25th percentiles, or upper and lower quartiles, IQR = Q3-Q1.

The IQR is calculated by subtracting the first quartile from the third quartile.

It is a measure of the dispersion similar to standard deviation or variance, but is much more robust against outliers.

12. What is the primary difference between bagging and boosting algorithms?

Ans. The primary difference between bagging and boosting algorithms as follow:-

a. Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions.

b. Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.

c. In Bagging, each model receives an equal weight. In Boosting, models are weighed based on their performance.

d. Models are built independently in Bagging. New models are affected by a previously built model’s performance in Boosting.

e. In Bagging, training data subsets are drawn randomly with a replacement for the training dataset. In Boosting, every new subset comprises the elements that were misclassified by previous models.

13. What is adjusted R2 in linear regression. How is it calculated?

Ans. The adjusted R-squared is a modified version of R-squared that accounts for predictors that are not significant in a regression model. In other words, the adjusted R-squared shows whether adding additional predictors improve a regression model or not.

The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable. In doing so, we can determine whether adding new variables to the model actually increases the model fit.

Let’s have a look at the formula for adjusted R-squared to better understand it’s working.

$$R2 = 1 – [(1-R2) * (n-1)/ (n-k-1)]$$

N - represents the number of data points in our dataset k - represents the number of independent variables, and

R - represents the R-squared values determined by the model

14. What is the difference between standardization and normalization?

Ans. Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.

Ans. Cross-Validation dataset: It is used to overcome the disadvantage of train/test split by splitting the dataset into groups of train/test splits, and averaging the result. It can be used if we want to optimize our model that has been trained on the training dataset for the best performance.

Advantages: -

- Reduces Overfitting
- Hyperparameter Tuning

Disadvantages: -

- Increases Training Time
- Needs Expensive Computation

Worksheet 4 SQL

1. Write a SQL query to show average number of orders shipped in a day (use Orders table).

Ans. SELECT ordernumber, shippeddate FROM Orders WHERE ordernumber = (SELECT

avg(ordernumber) FROM Orders);

2. Write a SQL query to show average number of orders placed in a day.

Ans. SELECT ordernumber, orderdate FROM Orders WHERE ordernumber = (SELECT

avg(ordernumber) FROM Orders);

3. Write a SQL query to show the product name with minimum MSRP (use Productstable).

Ans. SELECT productname, MSRP FROM Products WHERE MSRP = (SELECT MIN(MSRP)

FROM Products);

4. Write a SQL query to show the product name with maximum value ofstockQuantity.

Ans. SELECT productname, quantityinstock FROM Products WHERE quantityinstock = (SELECT

MAX(quantityinstock) FROM Products);

5. Write a query to show the most ordered product Name (the product with maximum number of

orders).

Ans. SELECT productName, COUNT(DISTINCT productCode) FROM Products GROUP BY

productName ORDER BY 2 DESC;

6. Write a SQL query to show the highest paying customer Name.

Worksheet 4 SQL

Ans. SELECT * Top 1 customername from customers join payements on

customers.customernumber=payments.customernumber group by customername order by count(*)

DESC;

. Write a SQL query to show cutomerNumber, customerName of all the customers who are from

Melbourne city.

Ans. SELECT cutomerNumber, customerName FROM Customers WHERE city= 'Melbourne city';

8. Write a SQL query to show name of all the customers whose name start with “N”.

Ans. SELECT * FROM Customers WHERE customername LIKE 'N%';

9. Write a SQL query to show name of all the customers whose phone start with ‘7’ and are from

city ‘LasVegas’.

Ans. SELECT * customerName FROM Customers WHERE city= 'Las Vegas'and WHERE

phone=‘^7*’ ;

10. Write a SQL query to show name of all the customers whose creditLimit < 1000 and city is

either “Las Vegas” or ”Nantes” or “Stavern”.

Ans. SELECT customer name from Customer where creditlimit>1000 and city=' “Las Vegas”or

” Nantes” or “Stavern”;

11. Write a SQL query to show all the order Number in which quantity ordered <10.

Ans. SELECT order number from Orderdetails where quantityorderd < 10 ;

Worksheet 4 SQL

12. Write a SQL query to show all the orderNumber whose customer Name start with letter ‘N’.

Ans. select ordernumber from Orders join Customers on

customers.customernumber=ordernumber.customernumber WHERE customername LIKE 'N%' ;

13. Write a SQL query to show all the customerName whose orders are “Disputed” in status.

Ans. Select ordernumber from Orders join Customers on

customers.customernumber=status.customernumber WHERE (status = "Disputed") ;

14. Write a SQL query to show the customerName who made payment through cheque with

checkNumber startingwith H and made payment on “2004-10-19”

Ans.Select customerNumber from payments Customers on

customers.customernumber=checkNumber.customernumber WHERE (paymentDate = “2004-10-

19” AND checkNumber LIKE ‘H%’) ;

15. Write a SQL query to show all the checkNumber whose amount > 1000.

Ans. SELECT checkNumber FROM Payments WHERE amount > 1000;

Worksheet 4 Statistics

1. What is central limit theorem and why is it important?

   Ans. The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed. Sufficiently large samples are term as sample set over 30 samples. Therefore, as a sample size increases, the sample mean and standard deviation will be closer in value to the population mean μ and standard deviation σ.

   Central limit theorem is important as in the real world and practical datasets, the samples size is sufficiently larger, thus we can infer the distribution of the sample set will be a Gaussian/normal distribution.

2. What is sampling? How many sampling methods do you know?

   Ans. Sampling is a process of selecting, manipulate and analyze a subset of the entire population. The samples are selected such that they represent the population. As a data scientist we analyze and find patterns data on the representative sample subset and validate those finding on the entire population using hypothesis testing. Sampling allows the data scientist to work on a smaller data set representing the entire population. Sampling methods can be classified into 2 types, probabilistic sampling methods and non-probabilistic types. Some of the sampling techniques I know are from probability sampling are simple random sampling, systematic, stratified and cluster sampling. From non-probabilistic sampling method, Convenience and Judgmental are known to me.

3. What is the difference between type1 and type II error?

   Ans. Type 1 error, is the error caused by rejecting a null hypothesis when it is true. Type II error is the error that occurs when the null hypothesis is accepted when it is not true. Type 1 error is the conclusion for false positives while type 2 error concludes false negative. Type 1 error is also called as the significance of the test. Type 2 error is also called as the beta error. As a result of type 1 error, then we might end up believing that the hypothesis works even when it doesn't. Whereas, as result of type 2 error, we might end up believing that the hypothesis works even when it doesn't.

4. What do you understand by the term Normal distribution?

   Ans. Normal distribution is a type of probability density function in a shape of a bell curve. It is also known as a Gaussian distribution curve. For an ideal normally distributed data mean, median and the mode all lie on the same point, that is the peak of the bell curve. For a normally distributed feature 68.26% of the data lies in the 1st standard deviation, 95.44% of the data lies in the 2nd standard deviation area and 99.73% of data lies within 3 standard deviations of the feature.

5. What is correlation and covariance in statistics?

Worksheet 4 Statistics

Ans. Covariance is a measure of the joint variability of two random variables. Correlation it is obtained by dividing the covariance of the two variables by the product of their standard deviations. The covariance values of the variable can lie anywhere between –inf to +inf whereas

the values of correlation are between -1 to +1. Also, correlation is a unit-free measure whereas covariance is not a unit-free measure.

6. Differentiate between univariate, Bivariate and multivariate analysis.

   Ans. Univariate analysis is done using a single feature from the dataset, bivariate analysis is performed using 2 features whereas multi-feature analysis is performed using more than 2 variables. Plots used for visualizing univariate analysis are count plots, histograms, density curves, distribution plots etc. Plots used for visualizing bivariate analysis are bar plots, scatter plots, joint plots, strip plots etc. Multivariate analysis plots are mode by adding hued data as an indication to the bivariate plots.

7. What do you understand by sensitivity and how would you calculate it?

   Ans. Sensitivity is the proportion of people who have the disease and tested positive among those who have the disease regardless of whether they tested positive or negative, i.e., among true positive cases and false negative cases. Therefore, the sensitivity equation is:

   Sensitivity = TP / (TP + FN),

8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?

   Ans. Hypothesis testing is the process used to evaluate the strength of evidence from the sample and provides a framework for making determinations related to the population. This sample is selected using one of the various sampling methods, probabilistic or non-probabilistic.

   H0 is the notation for null hypothesis whereas H1 is the notation for alternate hypothesis.

   For a two tailed test, the null hypothesis (H0) should be rejected when the test value is in either of two critical regions on either side of the distribution of the test value and vice versa for alternate hypothesis

9. What is quantitative data and qualitative data?

   Ans. Quantitative data can be counted, measured, and expressed using numbers. Qualitative data is descriptive and conceptual. Qualitative data can be categorized based on traits and characteristics.

10. How to calculate range and interquartile range?

Worksheet 4 Statistics

Ans. Range is calculated by: highest value – lowest value

IQR is calculated by: upper quartile (Q3) – lower quartile (Q1)

11. What do you understand by bell curve distribution?

Ans. Bell curve is defined as a graphical depiction of a normal probability distribution whose standard deviations from the mean form a bell-shaped curve. A standard deviation is a measurement that helps quantify the variability of data dispersion, and the mean, on the other hand, is the average of all data points in the data set and is found on the highest point on the bell curve.

Basically, a bell curve distribution represents the normal/ Gaussian distribution.

12. Mention one method to find outliers.

Ans. Calculate the z-score

A z-score, or standard score, shows how far away a data point is from the mean of the data. To calculate the z-score, you subtract the mean from the raw measurement and divide it by the standard deviation.

The equation for calculating the z-score is:

$$\mathbf{Z = (X-\mu) \div \sigma}$$

where:

X = raw measurement

μ = the mean

σ = the standard deviation

13. What is p-value in hypothesis testing?

Ans. The P value or calculated probability is the estimated probability of rejecting the null hypothesis (H0) of a study question when that hypothesis is true. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

14. What is the Binomial Probability Formula?

Ans. Binomial Probability formula: P(X) = (n! / (n-X)! X!) * (p)^X * (q)^n-X

Where X is the total number of successes.

p is the probability caused by success of an individual trail

q is the probability caused by failure of an individual train (q = 1-p)

n Is the number of trials.

15. Explain ANOVA and its applications.

Ans. **Analysis of variance (ANOVA)** is 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.

There are two main types of ANOVA: one-way (or unidirectional) and two-way. There also variations of ANOVA.

Applications of ANOVA:

- Understanding the impact of different catalysts on chemical reaction rates

- Understanding the performance, quality or speed of manufacturing processes based on number of cells or steps they're divided into

- Comparing the gas mileage of different vehicles, or the same vehicle under different fuel types, or road type