MACHINE LEARNING – WORSHEET 4

1. C
2. C
3. C
4. A
5. A
6. B
7. C
8. D
9. A & D
10. A, B & D

11. An observation which differs from an overall pattern on a sample dataset is called an outlier. The outliers may suggest experimental errors, variability in a measurement, or an anomaly. The age of a person may wrongly be recorded as 200 rather than 20 Years. Such an outlier should definitely be discarded from the dataset. IQR is the range between the first and the third quartiles namely Q1 and Q3: *IQR = Q3 – Q1*. The data points which fall below *Q1 – 1.5 IQR* or above *Q3 + 1.5 IQR* are outliers.

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. Before abnormal observations can be singled out, it is necessary to characterize normal observations. The inter quartile range rule is useful in detecting the presence of outliers. Outliers are individual values that fall outside of the overall pattern of a data set. This definition is somewhat vague and subjective, so it is helpful to have a rule to apply when determining whether a data point is truly an outlier—this is where the interquartile range rule comes in.

a) Calculate the interquartile range for the data, i.e (IQR = Q3 - Q1)

b) Multiply the interquartile range (IQR) by 1.5 (a constant used to discern outliers).

c) Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier.

d) Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.

12. In Bagging the result is obtained by averaging the responses of the N learners (or majority vote). However, Boosting assigns a second set of weights, this time for the N classifiers, in order to take a weighted average of their estimates.

13. **Adjusted R-Squared** measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable. Adjusted R-Squared can be calculated mathematically in terms of sum of squares. The only difference between R-square and Adjusted R-square equation is degree of freedom. Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size.

The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative, but it's usually not. It is always lower than the R-squared. Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size. Adjusted R2 also indicates how well terms fit a curve or line, but adjusts for the number of terms in a model. If you add more and more useless variables to a model, adjusted r-squared will decrease. If you add more useful variables, adjusted r-squared will increase.

14. The difference between standardization and normalization

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.

S.NO.	Normalization	Standardization
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

15. Cross Validation in Machine Learning is a great technique to deal with over fitting 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 Over fitting: In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from over fitting the training dataset. So, in this way, the model attains the generalization capabilities which are a good sign of a robust algorithm.

Note: 1. Chances of over fitting 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. Hyper parameter Tuning: Cross Validation helps in finding the optimal value of hyper parameters 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 hyper parameters. If you have multiple choices 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 – WORKSHEET 4

Q1. Write a SQL query to show average number of orders shipped in a day (use Orders table).

selectshippedDate, Avg (orderNumber) from Orders GROUP BY shippedDate;

Q2. Write a SQL query to show average number of orders placed in a day.

selectorderDate, Avg (orderNumber) from Orders GROUP BY orderDate;

Q3. Write a SQL query to show the product name with minimum MSRP (use Products table).

select productname, Min(MSRP) from product;

Q4. Write a SQL query to show the product name with maximum value of stockQuantity.

select productName, Max (quantityInStock) from products;

Q5. Write a query to show the most ordered product Name (the product with maximum number of orders).

Select productCode, count (orderNumber),max(quantityOrdered) from orderdetails group by productCode order by quantityOrdered desc;

Q6. Write a SQL query to show the highest paying customer Name.

Select C.customerName, p.amount from customers c join payments p on c.customerNumber = p.cuetomerNumber where amount = (select max(amount)from p.payments);

Q7. Write a SQL query to show cutomerNumber, customerName of all the customers who are from Melbourne city.

selectcustomerNumber, customerName from customers where city='Melbourne';

Q8. Write a SQL query to show name of all the customers whose name start with “N”.

selectcustomerName from customers where customerName like 'N%';

Q9. Write a SQL query to show name of all the customers whose phone start with ‘7’ and are from city ‘Las Vegas’.

selectcustomerName from customers where phone like '7_' and city='Las Vegas';

Q10. Write a SQL query to show name of all the customers whose creditLimit< 1000 and city is either “Las Vegas” or ”Nantes” or “Stavern”.

selectcustomerName from customers where creditLimit< 1000 and(city = 'Las Vegas' or city = 'Nantes' or city = 'Stavern');

Q11. Write a SQL query to show all the orderNumber in which quantity ordered <10.

selectorderNumber from orderdetails where quantityOrdered< 10;

Q12. Write a SQL query to show all the orderNumber whose customer Name start with letter ‘N’.

selectorderNumber from orders where customerName LIKE ‘N%’ INNER JOIN customers ON orders.customerNumber = customer.customerNumbere;

Q13. Write a SQL query to show all the customerName whose orders are “Disputed” in status.

selectcustomerName fromordersINNER JOIN customers ON orders.customerNumber= customers.CustomerNumberwhere status = 'Disputed';

Q14. Write a SQL query to show the customerName who made payment through cheque with checkNumber starting with H and made payment on “2004-10-19”.

selectcustomerName frompaymentsINNER JOIN customers ON payments.customerNumber = customers.customerNumberwhere checkNumber like 'H%' and paymentDate='2004-10-19';

Q15. Write a SQL query to show all the checkNumber whose amount > 1000.

selectcheckNumber from payments where amount > 1000;

STATISTICS - WORSHEET 4

Q1) What is central limit theorem and why is it important?

The Central Limit Theorem tells us that as sample sizes get larger, the sampling distribution of the mean will become normally distributed, even if the data within each sample are not normally distributed.

The Central Limit Theorem is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases. This means that we can take advantage of statistical techniques that assume a normal distribution

Q2) What is sampling? How many sampling methods do you know?

The **sample** is the specific group of individuals that you will collect data from. The sampling frame is the actual list of individuals that the sample will be drawn from. Ideally, it should include the entire target population (and nobody who is not part of that population).
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.

Q3) What is the difference between type1 and type II error?

Basis for comparison	Type I error	Type II error
Definition	Type 1 error, in statistical hypothesis testing, 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.
Also termed	Type I error is equivalent to false positive.	Type II error is equivalent to a false negative.
Meaning	It is a false rejection of a true hypothesis.	It is the false acceptance of an incorrect hypothesis.
Symbol	Type I error is denoted by α.	Type II error is denoted by β.
Probability	The probability of type I error is equal to the level of significance.	The probability of type II error is equal to one minus the power of the test.
Reduced	It can be reduced by decreasing the level of significance.	It can be reduced by increasing the level of significance.
Cause	It is caused by luck or chance.	It is caused by a smaller sample size or a less powerful test.
What is it?	Type I error is similar to a false hit.	Type II error is similar to a miss.
Hypothesis	Type I error is associated with rejecting the null hypothesis.	Type II error is associated with rejecting the alternative hypothesis.

When does it happen?	It happens when the acceptance levels are set too lenient.	It happens when the acceptance levels are set too stringent.

Q4) What do you understand by the term Normal distribution?

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 bell **curve**.

Q5) What is correlation and covariance in statistics?

Covariance and **Correlation** are two mathematical concepts which are commonly used in the field of probability and statistics. Both concepts describe the relationship between two variables.

Covariance –

1. It is the relationship between a pair of random variables where change in one variable causes change in another variable.
2. It can take any value between -infinity to +infinity, where the negative value represents the negative relationship whereas a positive value represents the positive relationship.

Correlation –

1. It show whether and how strongly pairs of variables are related to each other.
2. Correlation takes values between -1 to +1, wherein values close to +1 represents strong positive correlation and values close to -1 represents strong negative correlation.
3. In this variable are indirectly related to each other.
4. It gives the direction and strength of relationship between variables.

Q6) Differentiate between univariate , Bivariate and multivariate analysis.

Univariate statistics summarize only **one variable** at a time. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.
Bivariate statistics compare **two variables**.
Multivariate statistics compare **more than two variables**.

Q7) What do you understand by sensitivity and how would you calculate it?

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. The **sensitivity** is calculated by dividing the percentage change in output by the percentage change in input.

Q8) What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter.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.
All analysts use a random population sample to test two different hypotheses: the null hypothesis and the alternative hypothesis.
The null hypothesis would be represented as Ho: P = 0.5. The alternative hypothesis would be denoted as "H1" and be identical to the null 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.

Q9) What is quantitative data and qualitative data?

Quantitative data is information about quantities, and therefore numbers, **examples** are length, mass, temperature, and time whereas **qualitative data** is descriptive, and regards phenomenon which can be observed but not measured, such as language.

Q10) How to calculate range and inter quartile range?

The **Range** is the difference between the lowest and highest values. Example: In {4, 6, 9, 3, 7} the lowest value is 3, and the highest is 9. So the **range** is $9 - 3 = 6$.

We can find the interquartile range or IQR in four simple steps:

1. Order the data from least to greatest
2. Find the median
3. Calculate the median of both the lower and upper half of the data
4. The IQR is the difference between the upper and lower medians

Q11) What do you understand by bell curve distribution?

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.

Q12) Mention one method to find outliers.

One of the methods to Calculate the Outlier is by using the Interquartile Range

1. Take your IQR and multiply it by 1.5 and 3. We'll use these values to obtain the inner and outer fences. ...
2. **Calculate** the inner and outer lower fences. Take the Q1 value and subtract the two values from step 1. ...
3. **Calculate** the inner and outer upper fences.

Q13) What is p-value in hypothesis testing?

The **p-value**, or probability **value**, **tells you** how likely it is that your data could have occurred under the null hypothesis.The **p-value** is a proportion: if your **p-value** is 0.05, that means that 5% of the time **you** would see a test statistic at least as extreme as the one **you** found if the null hypothesis was true.

Q14) What is the Binomial Probability Formula?

The binomial distribution formula is:

$$\mathbf{B\ (x;\ n,\ P) = {}_nC_x * P^x * (1 - P)^{n-x}}$$

Where:

b = binomial probability

x = total number of “successes” (pass or fail, heads or tails etc.)

P = probability of a success on an individual trial

n = number of trials

Q15) Explain ANOVA and it's applications.

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 many industries that can use the ANOVA test to identify issues or variances between samples. The ANOVA is a good statistical technique for testing. Businesses that might consider the use of the ANOVA include manufacturing, healthcare, service, food, and more.