WORKSHEET SOLUTIONS

Machine learning:

1. c) Between -1 and 1
2. b) PCA
3. c) hyperplane
4. a) Logistic Regression
5. b) same as old coeffciient of X
6. c) decreases
7. c) Random Forests are easy to interpret
8. a) and c)
9. a) and d)
10. a) , b) and d)
11. 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 interquartile 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. Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.

13. 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.

You only need R2 when working with samples. In other words, R2 isn't necessary when you have data from an entire population.

The formula is:

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1} \right]$$

where:

- N is the number of points in your data sample.
- K is the number of independent regressors, i.e. the number of variables in your model, excluding the constant.

14. Standardisation :

The result of standardization (or Z-score normalization) is that the features will be rescaled to ensure the mean and the standard deviation to be 0 and 1, respectively. The equation is shown below:

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

This technique is to re-scale features value with the distribution value between 0 and 1 is useful for the optimization algorithms, such as gradient descent, that are used within machine learning algorithms that weight inputs (e.g., regression and neural networks). Rescaling is also used for algorithms that use distance measurements, for example, K-Nearest-Neighbours (KNN).

Normalisation :
This technique is to re-scales features with a distribution value between 0 and 1. For every feature, the minimum value of that feature gets transformed into 0, and the maximum value gets transformed into 1. The general equation is shown below:

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

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.

Statistics Answers :

1. The central limit theorem in statistics states that, given a sufficiently large sample size, the sampling distribution of the mean for a variable will approximate a normal distribution regardless of that variable's distribution in the population.The central limit theorem tells us that no matter what the distribution of the population is, the shape of the sampling distribution will approach normality as the sample size (N) increases.This is useful, as the research never knows which mean in the sampling distribution is the same as the population mean, but by selecting many random samples from a population the sample means will cluster together, allowing the research to make a very good estimate of the population mean.Thus, as the sample size (N) increases the sampling error will decrease.
2. Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined. It enables data scientists, predictive modelers and other data analysts to work with a small, manageable amount of data about a statistical population to build and run analytical models more quickly, while still producing accurate findings.

   Probability based sampling are as follows :
   - ➔ Simple random sampling: Software is used to randomly select subjects from the whole population
   - ➔ .Stratified sampling: Subsets of the data sets or population are created based on a common factor, and samples are randomly collected from each subgroup.

- ➔ Cluster sampling: The larger data set is divided into subsets (clusters) based on a defined factor, then a random sampling of clusters is analyzed.
- ➔ Multistage sampling: A more complicated form of cluster sampling, this method also involves dividing the larger population into a number of clusters. Second-stage clusters are then broken out based on a secondary factor, and those clusters are then sampled and analyzed. This staging could continue as multiple subsets are identified, clustered and analyzed.
- ➔ Systematic sampling: A sample is created by setting an interval at which to extract data from the larger population -- for example, selecting every 10th row in a spreadsheet of 200 items to create a sample size of 20 rows to analyze.

Nonprobability data sampling methods include:

- Convenience sampling: Data is collected from an easily accessible and available group.
- Consecutive sampling: Data is collected from every subject that meets the criteria until the predetermined sample size is met.
- Purposive or judgmental sampling: The researcher selects the data to sample based on predefined criteria.
- Quota sampling: The researcher ensures equal representation within the sample for all subgroups in the data set or population.

3. Type I error vs 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.

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 bell curve.A normal distribution is the proper term for a probability bell curve.

- In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
- Normal distributions are symmetrical, but not all symmetrical distributions are normal.
- In reality, most pricing distributions are not perfectly normal.

5.

Covariance	Correlation
Covariance is a measure to indicate the extent to which two random variables change in tandem.	Correlation is a measure used to represent how strongly two random variables are related to each other.
Covariance is nothing but a measure of correlation.	Correlation refers to the scaled form of covariance.

Covariance indicates the direction of the linear relationship between variables.	Correlation on the other hand measures both the strength and direction of the linear relationship between two variables.
Covariance can vary between -∞ and +∞	Correlation ranges between -1 and +1
Covariance is affected by the change in scale. If all the values of one variable are multiplied by a constant and all the values of another variable are multiplied, by a similar or different constant, then the covariance is changed.	Correlation is not influenced by the change in scale.
Covariance assumes the units from the product of the units of the two variables.	Correlation is dimensionless, i.e. It's a unit-free measure of the relationship between variables.
Covariance of two dependent variables measures how much in real quantity (i.e. cm, kg, liters) on average they co-vary.	Correlation of two dependent variables measures the proportion of how much on average these variables vary w.r.t one another.
Covariance is zero in case of independent variables (if one variable moves and the other doesn't) because then the variables do not necessarily move together.	Independent movements do not contribute to the total correlation. Therefore, completely independent variables have a zero correlation.

6. Univariate statistics summarize only one variable at a time.

- Bivariate statistics compare two variables.
- Multivariate statistics compare more than two variables.

7. Sensitivity (True Positive rate) measures the proportion of positives that are correctly identified (i.e. the proportion of those who have some condition (affected) who are correctly identified as having the condition).

Sensitivity = TP / (TP+FN). Where TP = True Positive, FN = false negative

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.Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data.

- The test provides evidence concerning the plausibility of the hypothesis, given the data.
- Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed.

8. Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed. All analysts use a random population sample to test two different hypotheses: the null hypothesis and the alternative hypothesis.
The null hypothesis is usually a hypothesis of equality between population parameters; e.g., a null hypothesis may state that the population mean return is equal to zero. The alternative hypothesis is effectively the opposite of a null hypothesis (e.g., the population mean return is not equal to zero). Thus, they are mutually exclusive, and only one can be true. However, one of the two hypotheses will always be true.
H. An alternative hypothesis that specified that the parameter can lie on either side of the value specified by H0 is called a two-sided (or two-tailed) test, e.g. H0: $\mu = 100$ HA: $\mu <> 100$

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

10. The range is the difference between the smallest and highest numbers in a list or set. To find the range, first put all the numbers in order. Then subtract (take away) the lowest number from the highest. The answer gives you the range of the list.
The interquartile range is a measure of where the "middle fifty" is in a data set. Where a range is a measure of where the beginning and end are in a set, an interquartile range is a measure of where the bulk of the values lie. That's why it's preferred over many other measures of spread when reporting things like school performance or SAT scores.
The interquartile range formula is the first quartile subtracted from the third quartile:
IQR = Q3 – Q1.

11. A bell curve is a common type of distribution for a variable, also known as the normal distribution. The term "bell curve" originates from the fact that the graph used to depict a normal distribution consists of a symmetrical bell-shaped curve.
The highest point on the curve, or the top of the bell, represents the most probable event in a series of data (its mean, mode, and median in this case), while all other possible occurrences are symmetrically distributed around the mean, creating a downward-sloping curve on each side of the peak. The width of the bell curve is described by its standard deviation.

12. IQR can be used to find outliers.

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. P is also described

14. The binomial distribution formula is:
$b(x; n, P) = nCx * Px * (1 – P)n – x$

15. 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.

A researcher might, for example, test students from multiple colleges to see if students from one of the colleges consistently outperform students from the other colleges. In a business application, an R&D researcher might test two different processes of creating a product to see if one process is better than the other in terms of cost efficiency.

The type of ANOVA test used depends on a number of factors. It is applied when data needs to be experimental. Analysis of variance is employed if there is no access to statistical software resulting in computing ANOVA by hand. It is simple to use and best suited for small samples. With many experimental designs, the sample sizes have to be the same for the various factor level combinations. ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources. It is employed with subjects, test groups, between groups and within groups.

SQL Solutions :

1. Select avg(orderNumber) from (select count(*) as orderCount, distinct(shipped date) from orders);
2. Select avg(orderNumber) from (select count(*) as orderCount, distinct(orderDate) from orders) ;
3. Select productName from products where MSRP = (select min(MSRP) from products) ;
4. Select productName from products where quantitInStock= (select max(quantityInStock) from products) ;
5. Select productCode, count (orderNumber),max(quantityOrdered) from orderdetails group by productCode order by quantityOrdered desc;
6. Select C.customerName, p.amount from customers c join payments p on c.customerNumber = p.cuetomerNumber where amount = (select max(amount)from p.payments);
7. Select customerNumber, customerName from customers where city = 'Melbourne';
8. Select customerName from customers where customerName like 'N%';
9. Select * from customers where phone like '7%' and city = 'LasVegas';
10. Select customerName from customers where creditLimit < 1000 and city in ('LasVegas','Nantes','Stavern');
11. Select orderNumber from orderdetails where quantityOrdered < 10;
12. Select O.orderNumber from orders O join customers C on O.customerNumber = C.customerNumber where C.customerName like 'N%';
13. Select O.orderNumber from orders O join customers C on O.customerNumber = C.customerNumber where O.status = 'Disputed';
14. Select C.customerName from customers C join payments P on P.customerNumber = C.customerNumber where p.checkNumber like 'H%' and paymentDate = '2004-10-19';
15. Select checkNumber from payments where amount > 1000;