Machine Learning

1. C
2. D
3. C
4. D
5. D
6. C
7. C
8. B & C
9. A,B,C,D
10. A & D
11. An outlier is a data point that lies outside the overall pattern in a distribution.

The interquartile range is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q3) and the firstquartile (Q1). IQR = Q3 – Q1. The IQR can help determine outliers.

1. Bagging tries to solve over-fitting problem. Boosting tries to reduce bias.

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

1. 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. Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size. Every time you add a independent variable to a model, the R-squared increases, even if the independent variable is insignificant.

SQL

1. select AVG(shippedDate,) from orders GROUP BY orderDate;
2. select AVG(oderNumbers) from orders GROUP BY orderDate;
3. select productName from products where MSRP = (select MIN(MSRP) from products);
4. select productName from products where quantityinstocks = (select MAX(Quantityinstock) from products);
5. select productName from products, COUNT(DISTINCT productCode), MAX(quantityOrder) from orderDetails GROUP BY productCode;
6. select a.customerName from customers a, where b.amount =(select MAX(amount)) from payments AND a.customerNumber = b.customerNumber;
7. select customerName, customernumber from customers where city = ‘Melbourne’;
8. select customerName from customers where customerName LIKE ‘N%’ order by customerName;
9. select customerName from customers where customerNumber LIKE ‘7%’ AND city=’LasVegas’ order by customerName;
10. select customerName from customers where creditLimit<1000 AND city=’LasVegas’ OR city=’Nantes’ OR city=’Stavern’;
11. select orderNumber from orderdetails where quantityOrder <10;
12. select orderNumber from orders a, where b.customerName LIKE ‘N%’ from customers AND a.customerNumber=b.customerNumber;
13. select a.customerName from customers where b.status=’Disputed’ from orders AND a.customerNumber=b.customerNumber;
14. select a.customerName from customers where b.checkNumber LIKE ‘H%’ AND b.paymentdate=’2004-10-19’ from payments AND a.customerNumber=b.customerNumber;
15. select checkNumber from payments where amount>1000;

Statistics

1. 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.
2. Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. The methodology used to sample from a larger population depends on the type of analysis being performed, but it may include simple random sampling or systematic sampling.

There are two types of sampling methods:

[Probability sampling](https://www.scribbr.com/methodology/sampling-methods/#probability-sampling) involves random selection, allowing you to make strong statistical inferences about the whole group.

[Non-probability sampling](https://www.scribbr.com/methodology/sampling-methods/#non-probability-sampling) involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

1. 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. Type I error is equivalent to false positive.
2. 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.
3. Covariance is a measure of how much two random variables vary together. Correlation is a statistical measure that indicates how strongly two variables are related.
4. Univariate statistics summarize only one variable at a time. Bivariate statistics compare two variables. Multivariate statistics compare more than two variables.
5. 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, sensitivityanalyses study how various sources of uncertainty in a mathematical model contribute to the model's overall uncertainty. The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.
6. A statistical hypothesis is an assertion or conjecture concerning one or more populations.

Hypothesis testing is formulated in terms of two hypotheses: H0: the null hypothesis; H1: the alternate hypothesis. 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. Null hypothesis (H0): The null hypothesis here is what currently stated to be true about the population. ... Alternate hypothesis (H1): The alternate hypothesis is always what is being claimed. “In our case, Tedd believes(Claims) that the actual value has changed”.

1. 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.
2. To find the interquartile range (IQR), ​first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.
3. 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.
4. The most effective way to find all of your outliers is by using the interquartile range (IQR). The IQR contains the middle bulk of your data, so outliers can be easily found once you know the IQR. Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier.Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.
5. 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 oneyou found if the null hypothesis was true.
6. Binomial probability refers to the probability of exactly x successes on n repeated trials in an experiment which has two possible outcomes (commonly called a binomial experiment). If the probability of success on an individual trial is p , then the binomial probability is nCx⋅px⋅(1−p)n−x .
7. 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. Another measure to compare the samples is called a t-test.

ANOVA models can realistically be used in numerous industries and applications:

1. Comparing the gas mileage of different vehicles, or the same vehicle under different fuel types, or road types.
2. Understanding the impact of temperature, pressure or chemical concentration on some chemical reaction (power reactors, chemical plants, etc)
3. Understanding the impact of different catalysts on chemical reaction rates
4. Studying whether advertisements of different kinds solicit different numbers of customer responses
5. Understanding the performance, quality or speed of manufacturing processes based on number of cells or steps they’re divided into