Machine Learning WORKSHEET-3

Answer-13)

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). Clustering is required because:

• Organizing data into clusters shows internal structure of the data

• Sometimes the partitioning is the goal

• Prepare for other AI techniques – Ex. Summarize news (cluster and then find centroid)
• Techniques for clustering is useful in knowledge discovery in data – Ex. Underlying rules, reoccurring patterns, topics, etc

Answer-14)

Clustering performance can be improved by using independent component analysis, unsupervised feature learning, Feature Weight Learning and performing K-means clustering

STATISTICS WORKSHEET-3

Answer-10)

The Bayes' theorem (also known as the Bayes' rule) is a mathematical formula used to determine the conditional probability of events. Essentially, the Bayes' theorem describes the probability of an event based on prior knowledge of the conditions that might be relevant to the event.

The Bayes' theorem is expressed in the following formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where:

P(A|B) – the probability of event A occurring, given event B has occurred

P(B|A) – the probability of event B occurring, given event A has occurred

P(A) – the probability of event A

P(B) – the probability of event B

Answer-11)

A z-score (also called a standard score) gives you an idea of how far from the mean a data point is. But more technically it's a measure of how many standard deviations below or above the population mean a raw score is.

Answer-12)

A t-test is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another.

Answer-13)

A percentile is a number where **a certain percentage of scores fall below that number.** We might know that we have scored 67 out of 90 on a test. But that figure has no real meaning unless we know what percentile we fall into. If we know that our score is in the 90th percentile that means we have scored better than 90% of people who took the test.

Answer-14)

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

Answer-15)

An **ANOVA** test is a way to find out if survey or experiment results are significant. In other words, they help you to figure out if you need to reject the null hypothesis or accept

the alternate hypothesis. Basically, **we're testing groups to see if there's a difference between them. 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.