Machine Learning Set-3

Question No	Answers
1)	d) All of the above.
2)	d) None
3)	c) Reinforcement learning and unsupervised learning.
4)	b) The tree representing how close the data points are to each other.
5)	d) None.
6)	c) k-nearest neighbour is same as k-means.
7)	d) 1, 2 and 3.
8)	a) 1 only.
9)	a) 2
10)	b) Given a database of information about your users, automatically group them into different market segments.
11)	a)
12)	b)

Q13) Answer: -

The primary use of clustering in machine learning is to extract valuable inferences from many unstructured datasets. If you are working with large amount of data that are also not structured, it is only logical to organize that data to make it helpful in so many other ways and clustering help us do that. Clustering and classification allow you to take sweeping glance at your data. And then form some logical structures based on what you find there before going deeper analysis. Clustering is a significant component of machine learning and its importance is highly significant in providing better machine learning techniques. Following are some uses of clustering in machine learning,

I. Medical imaging.
II. Privacy preservation.
III. Data compression.
IV. Anomaly detection.
V. Image segmentation.
VI. Search result grouping.
VII. Social network analysis.

Q14) Answer: -

Three methodology findings are as below,

I. Graph base clustering performance can easily be improved by applying ICA (Independent component analysis) blind source separation during the Laplacian embedding step.
II. Applying unsupervised feature learning to input data using either RICA (Reconstruction Independent component analysis) or SFT (Shrink and fine tune) improves clustering performance.
III. Surprisingly for some cases high clustering performance can be achieved by simply performing k-means clustering on the ICA components after PCA (Principal component analysis) dimension reduction on the input data.

STATISTICS WORKSHEET-3

Question No	Answers
1)	b) Total variation = Residual variation + Regression variation.
2)	c) binomial.
3)	a) 2
4)	a) Type-I error
5)	b) size of the test.
6)	b) Increase.
7)	b) Hypothesis.
8)	d) All of the mentioned.
9)	a) 0

Q10 Answer: -

Bayes theorem calculates the conditional probability of an event, based on the values of specific related known probabilities. A bayes theorem calculator figures the probability of an event A conditional on another event B, given the prior probabilities of A & B and the probability of B conditional on A. It calculates conditional probability based on known probabilities. Bayes theorem relies on incorporating prior probabilities distributions in order to generate posterior probabilities. Prior probabilities, in Bayesian statistical inference is the probability of an event occurring before new data is collected. In other words, it represents the best rational assessment of the probability of a particular outcome based on current knowledge before an experiment is performed. In statistical term posterior probability is the probability of event A occurring given that event B has occurred.

Bayes Formula

$$P(x, y) = P(x \mid y)P(y) = P(y \mid x)P(x)$$

$$\Rightarrow$$

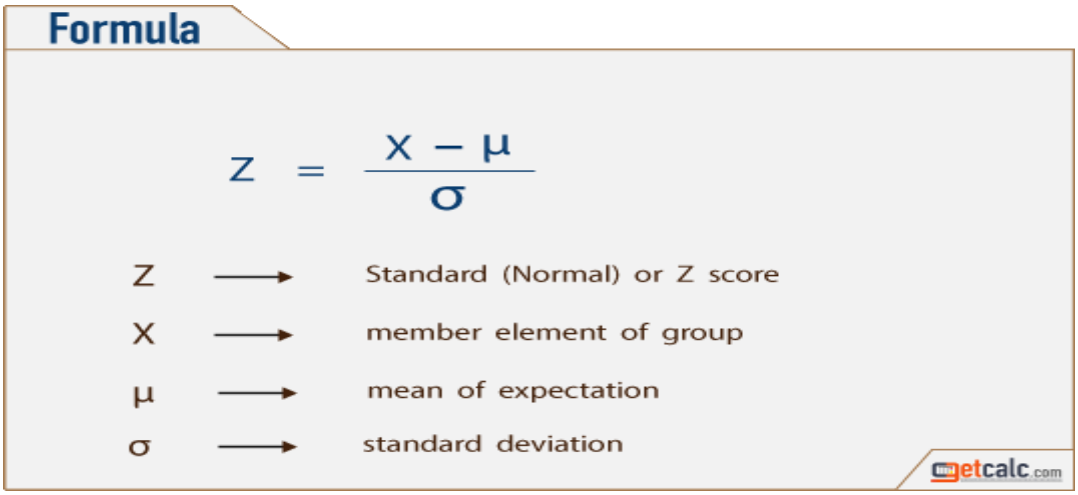
$$P(x \mid y) = \frac{P(y \mid x)\ P(x)}{P(y)} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

Q11 Answer: -

A z-score gives us idea of how far from the mean a data point is. A measure of how many standard deviations below or above the population mean a raw score is called as z-score. It will be positive if the values lie above the mean and negative if it lies below the mean. It is also known as standard score. It indicates that how many standard deviations as entity is, from the mean. In order to calculate z-score, the mean and the standard deviation should be known.

Interpretation based on z-score: -

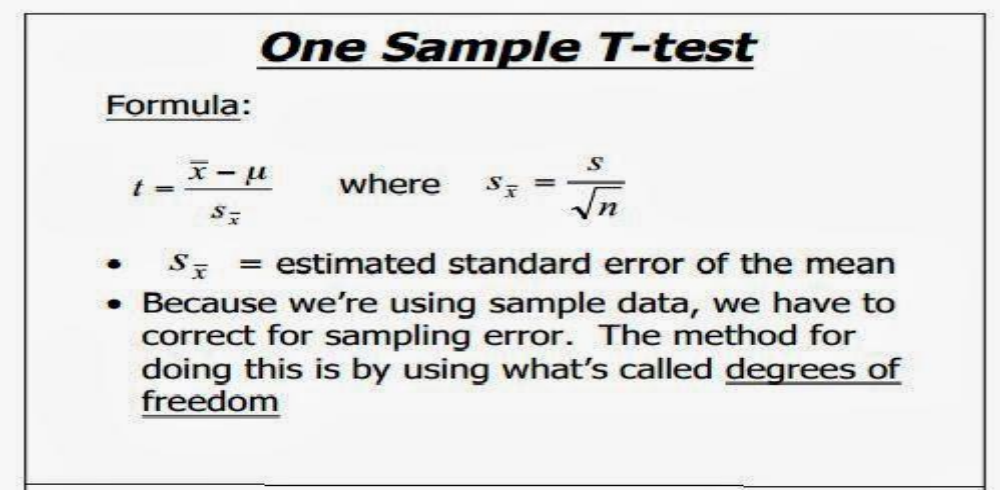
I. If a z-score is equal to -1, then it denotes an element which is 1 standard deviation less than the mean.
II. If the z-score is equal to 1, it denotes an element which is 1 standard deviation greater than the mean.
III. If the z-score is equal to 2 signifies 2 standard deviations greater than the mean.
IV. If a z-score is less than 0, then it denotes an element less than the mean.
V. If a z-score is greater than 0, then it denotes an element greater than mean.
VI. If a z-score is equal to 0, then it denotes an element equal to mean.



Q12 Answer: -

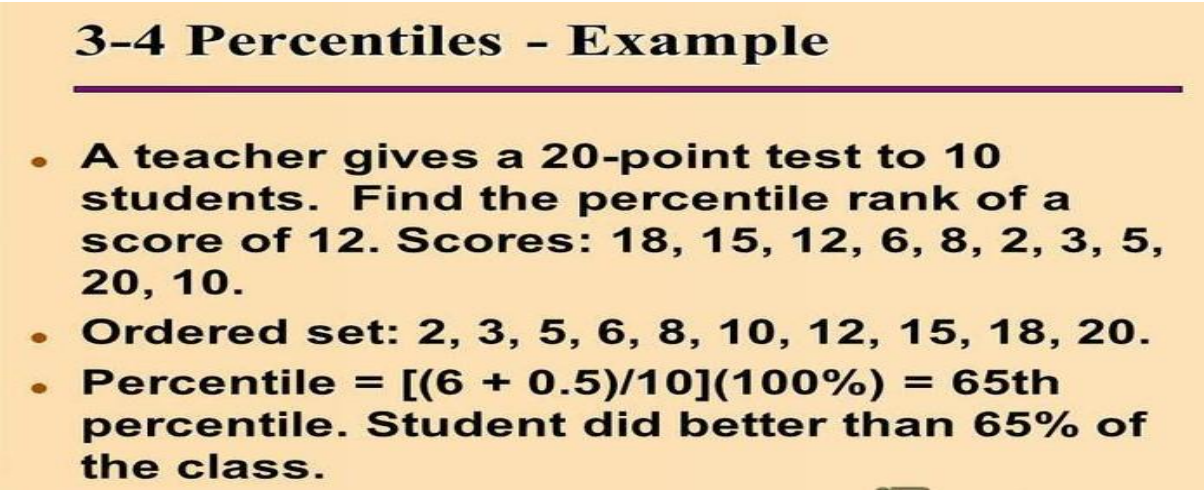
A t-test is statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment influences the population of interest or whether two groups are different from one another. Calculating t-test requires three fundamental data values including the difference between the mean values from each dataset, the std deviation of each group, and the number of data values. Mathematically the t-test takes sample from each of the two sets and establish the problem statement. It assumes a null

hypothesis that two means are equal. Using the formulas values are calculated and compared against the standard values. The assume null hypothesis is accepted or rejected accordingly. If the null hypothesis qualifies to be rejected it indicates that the data readings are strong and are probably not due to chance.



Q13 Answer: -

In statistics, a k-th percentile is a score below which a given percentage of scores in its frequency distribution falls. The 25th percentile is also known as first quartile (Q1), the 50th percentile is also known as median or second quartile (Q2), and the 75th percentile also known as third quartile (Q3).



Q14 Answer: -

Analysis of variance, or ANOVA is a statistical method that separates observed variance data into different components to use for additional tests. ANOVA splits the observed aggregate variability found inside a dataset into two parts: systematic factors and random factors. Systematic factors have a statistical influence on the given dataset, while the random factors do not. Analyst use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study. A one-way ANOVA is used for three or more groups of data to gain information about the relationship between dependent and independent variables. If no true variance exists between the groups, the ANOVA's F ratio should equal to 1. With two-wat ANOVA there are two independents. For example, it allows company to compare worker productivity based on two independent variables such as salary and skill set.

Q15 Answer: -

Even though ANOVA involves complex statistical steps, it is beneficial technique for business via use of AI. Organization use ANOVA to make decisions about which alternative to choose among many possible options. For example, ANOVA can help to,

I. Compare the yield of two different wheat varieties under three different fertilizer brands.
II. Compare the effectiveness of different lubricants in different types of vehicles.