Machine Learning

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

1)d, 2)d, 3)a, 4)b, 5)a, 6)c, 7)d, 8)c, 9)a, 10)b, 11)a, 12)b

Q13 to Q14 are subjective answers type questions, Answers them in their own words briefly.

13. What is the importance of clustering?

When we are trying to learn about something, say music, one approach might be to look for meaningful groups or collections. We might organize music by genre, while someone might organize music by decade. However, to group items helps us to understand more about them as individual pieces of music. We might find that we have a deep affinity for punk rock and further break down the genre into different approaches or music from different locations. On the other hand, our friend might look at music from the 1980's and be able to understand how the music across genres at that time was influenced by the socio-political climate. In both cases, we and our friend have learned something interesting about music, even though we took different approaches.

In machine learning too, we often group examples as a first step to understand a subject (data set) in a machine learning system. Grouping unlabelled examples is called clustering.

As the examples are unlabelled, clustering relies on unsupervised machine learning. If the examples are labelled, then clustering becomes classification. For a more detailed discussion of supervised and unsupervised methods see Introduction to Machine Learning Problem Framing.

Before we group similar examples, we first need to find similar examples. We can measure similarity between examples by combining the examples' feature data into a metric, called a **similarity measure**. When each example is defined by one or two features, it's easy to measure similarity. For example, we can find similar books by their authors. As the number of features increases, creating a similarity measure becomes more complex. We'll later see how to create a similarity measure in different scenarios.

What are the Uses of Clustering?

Clustering has a myriad of uses in a variety of industries. Some common applications for clustering include the following:

- market segmentation
- social network analysis

- search result grouping
- medical imaging
- image segmentation
- anomaly detection

After clustering, each cluster is assigned a number called a **cluster ID**. Now, we can condense the entire feature set for an example into its cluster ID. Representing a complex example by a simple cluster ID makes clustering powerful. Extending the idea, clustering data can simplify large datasets.

14. How can I improve my clustering performance?

Clustering is an unsupervised machine learning methodology that aims to partition data into distinct groups, or clusters. There are a few different forms including hierarchical, density, and similarity based. Each have a few different algorithms associated with it as well. One of the hardest parts of any machine learning algorithm is feature engineering, which can especially be difficult with clustering as there is no easy way to figure out what best segments your data into separate but similar groups.

STATISTICS WORKSHEET-3

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1)b, 2)c, 3)a, 4)a, 5)b, 6)a, 7)b, 8)d, 9)a

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

Q10. What Is Bayes' Theorem?

Bayes' theorem describes the probability of occurrence of an event related to any condition. It is also considered for the case of conditional probability. Bayes theorem is also known as the formula for the probability of "causes". For example: if we have to calculate the probability of taking a blue ball from the second bag out of three different bags of balls, where each bag contains three different colour balls viz. red, blue, black. In this case, the probability of occurrence of an event is calculated depending on other conditions is known as conditional probability.

Bayes Theorem Statement

Let $E_1, E_2, \ldots, E_n$ be a set of events associated with a sample space S, where all the events $E_1, E_2, \ldots, E_n$ have nonzero probability of occurrence and they form a partition of S. Let A be any event associated with S, then according to Bayes theorem,

$$P(Ei|A) = P(Ei)P(A|Ei)\sum_{k=1}^{n}P(Ek)P(A|Ek)$$

for any k = 1, 2, 3, …., n

Bayes Theorem Applications

One of the many applications of Bayes' theorem is Bayesian inference, a particular approach to statistical inference. Bayesian inference has found application in various activities, including medicine, science, philosophy, engineering, sports, law, etc. For example, we can use Bayes' theorem to define the accuracy of medical test results by considering how likely any given person is to have a disease and the test's overall accuracy. Bayes' theorem relies on consolidating prior probability distributions to generate posterior probabilities. In Bayesian statistical inference, prior probability is the probability of an event before new data is collected.

Q11. What is z-score?

A Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score. A Z-score of 1.0 would indicate a value that is one standard deviation from the mean. Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.

In finance, Z-scores are measures of an observation's variability and can be used by traders to help determine market volatility. Z-scores reveal to statisticians and traders whether a score is typical for a specified data set or if it is atypical. Z-scores also make it possible for

analysts to adapt scores from various data sets to make scores that can be compared to one another more accurately.

Edward Altman, a professor at New York University, developed and introduced the Z-score formula in the late 1960s as a solution to the time-consuming and somewhat confusing process investors had to undergo to determine how close to bankruptcy a company was.1 In reality, the Z-score formula that Altman developed actually ended up providing investors with an idea of the overall financial health of a company.

Over the years, Altman continued to re-evaluate his Z-score. From 1969 until 1975, Altman looked at 86 companies in distress. From 1976 to 1995, he observed 110 companies. Finally, from 1997 to 1999, he evaluated an additional 120 companies. From his findings, it was revealed that the Z-score had an accuracy of between 82% and 94%.2

In 2012, Altman released an updated version of the Z-score, which is called the Altman Z-score Plus. It can be used to evaluate public and private companies, manufacturing and non-manufacturing companies, and U.S. and non-U.S. companies.

Q12. What is t-test?

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.

Assumptions of t-test:

The measurement scale used for such hypothesis testing follows a set of continuous or ordinal patterns. The accounted parameters and variants influencing the samples and surrounding the groups are based on the standard consideration.

The tests are completely based on random sampling. As no individuality is maintained in the samples, the reliability is often questioned.

When the data is plotted with respect to the T-test distribution, it should follow a normal distribution and bring about a bell-curved graph.

For a clearer bell curve, the sample size needs to be bigger.

The variance should be such that the standard deviations of the samples are almost equal.

Types of t-test:

One-Sample T-Test

Independent Two-Sample T-Test

Paired Sample T-Test

Equal Variance T-Test

Unequal Variance T-Test

Q13 What is percentile?

In statistics, a percentile is a term that describes how a score compares to other scores from the same set. While there is no universal definition of percentile, it is commonly expressed as the percentage of values in a set of data scores that fall below a given value.

Q14. What is ANOVA?

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.

Q15. How can ANOVA help?

An ANOVA ("Analysis of Variance") is a statistical technique that is used to determine whether or not there is a significant difference between the means of three or more independent groups. The two most common types of ANOVAs are the one-way ANOVA and two-way ANOVA.

Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.

A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.

If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.