MACHINE LEARNING ASSIGNMENT SOLUTION

1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:

ANS. b) 4

2. In which of the following cases will K-Means clustering fail to give good results?

1. Data points with outliers

2. Data points with different densities

3. Data points with round shapes

4. Data points with non-convex shapes

ANS. d) 1, 2 and 4

3. The most important part of is selecting the variables on which clustering is based.

ANS. d) formulating the clustering problem

4. The most commonly used measure of similarity is the or its square.

ANS. a) Euclidean distance

5. is a clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.

ANS. a) Non-hierarchical clustering

6. Which of the following is required by K-means clustering?

ANS. d) All answers are correct

7. The goal of clustering is to

ANS. a) Divide the data points into groups

8. Clustering is a

ANS. b) Unsupervised learning

9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?

ANS. d) All of the above

10. Which version of the clustering algorithm is most sensitive to outliers?

ANS. a) K-means clustering algorithm

11. Which of the following is a bad characteristic of a dataset for clustering analysis

ANS. d) All of the above

12. For clustering, we do not require

ANS. a) Labeled data

13. How is cluster analysis calculated?

ANS. Cluster analysis is calculated by using only three steps:

- Copy your data into the table
- Select more than one variable
- Select the number of clusters you want to calculate

Clusters can be calculated using various grouping methods. These can be divided into

- graph-theoretical
- hierarchically
- partitioning
- optimizing

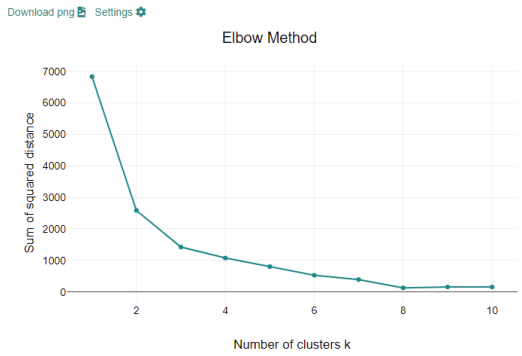
Optimal cluster number

The number of clusters in the k-Means method must be determined before the start and is therefore not determined by the cluster method. But what is

the optimal number of clusters in the k-Means method? The elbow method is a common way to determine the appropriate number of clusters.

Elbow curve

When you want to calculate a cluster analysis, often the big question is how many clusters should I take, The Elbow Method helps with this question! With each new cluster, the total variation in each cluster becomes smaller and smaller. In the extreme case, when there are as many clusters as there are points, the result is zero. However, in most cases, the reduction of the total variation becomes smaller after a certain point. This point is then used as the optimal cluster number.



14. How is cluster quality measured?

ANS. The following methods can be used to assess the quality of clustering algorithms based on internal criterion:

- Davies–Bouldin index

  Consider the following equation defined by Davies, D., & Bouldin, D. (1979) that calculates the dispersion of cluster i:

  $$Si=\{1Ti\sum j=1Ti|Xj–Ai|q\}1q$$

  where:
  - i : particular identified cluster
  - Ti : number of vectors (observations) in cluster i
  - Ti : number of vectors (observations) in cluster i
  - Xj : j-th vector (observation) in cluster i
  - Ai : centroid of cluster i

Basically, to get the intra-cluster dispersion, we calculate the average distance between each observation within the cluster and its centroid.

Davies–Bouldin index is considered the best algorithm based on this criterion.

- Silhouette Index – Silhouette analysis refers to a method of interpretation and validation of consistency within clusters of data.

  The Silhouette score measures how close each point is to his cluster and how far it is from the closest cluster.
  For each point i:
  a_i - the average distance of point i from all his cluster's points.
  b_i - the average distance of point i from all the points in the closest cluster.

  $$S_i = \frac{b_i - a_i}{Max(a_i, b_i)}$$

  The Silhouette range is [-1,1].
  High Silhouette value - the point is close to his cluster, and far from other clusters.
  Zero Silhouette value - moving the point to the closest cluster will not have a big change on the SSE.
  Negative Silhouette value - the point may be an outlier, or was assigned to the wrong cluster.

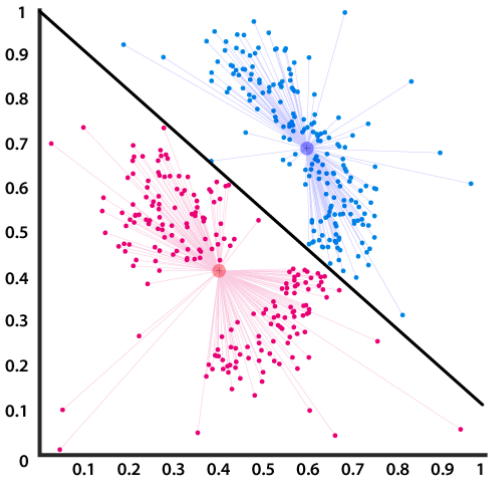
15. What is cluster analysis and its types?

ANS. Cluster analysis is a multivariate data mining technique whose goal is to groups objects (eg. products, respondents, or other entities) based on a set of user selected characteristics or attributes. It is the basic and most important step of data mining and a common technique for statistical data analysis, and it is used in many fields such as data compression, machine learning, pattern recognition, information retrieval etc.

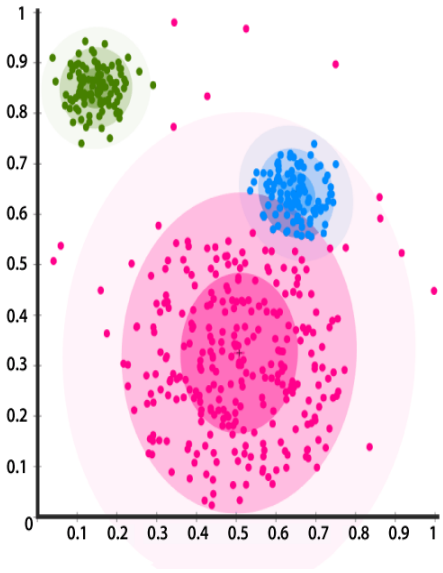
Types of Cluster Analysis

It should be noted that an algorithm that works on a particular set of data will not work on another set of data. There are a number of different methods to perform cluster analysis. Some of them are

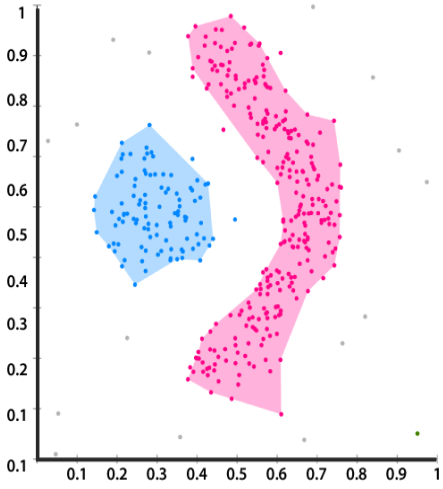
- Hierarchical Cluster Analysis- In this method, first, a cluster is made and then added to another cluster (the most similar and closest one) to form one single cluster. This process is repeated until all subjects are in one cluster. This particular method is known as **Agglomerative method**.
- Centroid-based Clustering- In this type of clustering, clusters are represented by a central entity, which may or may not be a part of the given data set. K-Means method of clustering is used in this method, where k are the cluster centers and objects are assigned to the nearest cluster centers



- Distribution-based Clustering- It is a type of clustering model closely related to statistics based on the modals of distribution. Objects that belong to the same distribution are put into a single cluster. This type of clustering can capture some complex properties of objects like correlation and dependence between attributes.



- Density-based Clustering-In this type of clustering, clusters are defined by the areas of density that are higher than the remaining of the data set. Objects in sparse areas are usually required to separate clusters. The objects in these sparse points are usually noise and border points in the graph. The most popular method in this type of clustering is DBSCAN.



STATISTICS WORKSHEET-1 SOLUTION

1. Bernoulli random variables take (only) the values 1 and 0.

ANS. a) True

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

ANS. a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

ANS. b) Modeling bounded count data

4. Point out the correct statement

ANS. d) All of the mentioned

5. _____ random variables are used to model rates.

ANS. c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.

ANS. b) False

7. 1. Which of the following testing is concerned with making decisions using data?

ANS. b) Hypothesis

8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.

ANS. a) 0

9. Which of the following statement is incorrect with respect to outliers?

ANS. c) Outliers cannot conform to the regression relationship

10. What do you understand by the term Normal Distribution?

ANS. A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme. The middle of the range is also known as the mean of the distribution. The normal distribution is also known as a Gaussian distribution or probability bell curve. It is symmetric about the mean and indicates that values near the mean occur more frequently than the values that are farther away from the mean.

11. How do you handle missing data? What imputation techniques do you recommend?

ANS Missing data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical programme will make the decision for you. Your application will remove things in a listwise sequence most of the time. Depending on why and how much data is gone, listwise deletion may or may not be a good idea. Another common strategy among those who pay attention is imputation. Imputation is the process of substituting an estimate for missing values and analyzing the entire data set as if the imputed values were the true observed values

Mean imputation-Calculate the mean of the observed values for that variable for all non-missing people. It has the advantage of maintaining the same mean and sample size, but it also has a slew of drawbacks. Almost all of the methods described below are superior to mean imputation.

Substitution-Assume the value from a new person who was not included in the sample. To put it another way, pick a new subject and employ their worth instead.

Hot deck imputation-A value picked at random from a sample member who has comparable values on other variables. To put it another way, select all the sample participants who are comparable on other factors, then choose one of their missing variable values at random

12. What is A/B testing?

ANS A/B testing is a method of comparing two versions of a webpage or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

13. Is mean imputation of missing data acceptable practice?

ANS The process of replacing null values in a data collection with a data's mean is known as mean imputation. Mean imputation is typically considered as terrible practice since it ignores feature correlation. Second, mean imputation decreases the value of variance, the model is less accurate.

14. What is linear regression in statistics?

ANS Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable.

15. What are the various branches of statistics?

ANS Statistics may be divided into two main branches:

- Descriptive Statistics
- Inferential Statistics

(1) Descriptive Statistics - Descriptive statistics deals with the collection of data, its presentation in various forms, such as tables, graphs and diagrams and finding averages and other measures which would describe the data. For example: Industrial statistics, population statistics, trade statistics, etc. Businessmen make use of descriptive statistics in presenting their annual reports, final accounts, and bank statements.

(2) Inferential Statistics - Inferential statistics deals with techniques used for the analysis of data, making estimates and drawing conclusions from limited information obtained through sampling and testing the reliability of the estimates. For example: We take a sample from the population and find the proportion of illiterate individuals in the sample. With the help of probability, this sample proportion enables us to make some inferences about the population proportion

WORKSHEET SQL

1. Which of the following is/are DDL commands in SQL?

ANS A) Create, D)Alter

2. Which of the following is/are DML commands in SQL?

ANS A) Update , B)Delete

3. Full form of SQL is:

ANS B) Structured Query Language

4. Full form of DDL is:

ANS B) Data Definition Language

5. DML is:

ANS A) Data Manipulation Language

6. Which of the following statements can be used to create a table with column B int type and C floattype?

ANS C) Create Table A (B int,C float)

7. Which of the following statements can be used to add a column D (float type) to the table A created above?

ANS B) Alter Table A ADD COLUMN D float

8. Which of the following statements can be used to drop the column added in the above question?

ANS D) None of them

9. Which of the following statements can be used to change the data type (from float to int) of the column Dof table A created in above questions?

ANS B) Alter Table A Alter Column D int

10. Suppose we want to make Column B of Table A as primary key of the table. By which of the following statements we can do it?

ANS. C) Alter Table A Add Primary key B

11. What is data-warehouse?

ANS A Data Warehousing (DW) is process for collecting and managing data from varied sources to provide meaningful business insights. A Data warehouse is typically used to connect and analyze business data from heterogeneous sources. The data warehouse is the core of the BI system which is built for data analysis and reporting.

12. What is the difference between OLTP VS OLAP?

ANS OLTP and OLAP: The two terms look similar but refer to different kinds of systems. Online transaction processing (OLTP) captures, stores, and processes data from transactions in real time. Online analytical processing (OLAP) uses complex queries to analyze aggregated historical data from OLTP systems.

OLTP is operational, while OLAP is informational. A glance at the key features of both kinds of processing illustrates their fundamental differences, and how they work together.

	OLTP	OLAP
Characteristics	Handles a large number of small transactions	Handles large volumes of data with complex queries
Query types	Simple standardized queries	Complex queries
Operations	Based on INSERT, UPDATE, DELETE commands	Based on SELECT commands to aggregate data for reporting
Response time	Milliseconds	Seconds, minutes, or hours depending on the amount of data to process
Design	Industry-specific, such as retail, manufacturing, or banking	Subject-specific, such as sales, inventory, or marketing
Source	Transactions	Aggregated data from transactions
Purpose	Control and run essential business operations in real time	Plan, solve problems, support decisions, discover hidden insights
Data updates	Short, fast updates initiated by user	Data periodically refreshed with scheduled, long-running batch jobs
Space requirements	Generally small if historical data is archived	Generally large due to aggregating large datasets
Backup and recovery	Regular backups required to ensure business continuity and meet legal	Lost data can be reloaded from OLTP database as needed in lieu of regular backups

	and governance requirements	
Productivity	Increases productivity of end users	Increases productivity of business managers, data analysts, and executives
Data view	Lists day-to-day business transactions	Multi-dimensional view of enterprise data
User examples	Customer-facing personnel, clerks, online shoppers	Knowledge workers such as data analysts, business analysts, and executives
Database design	Normalized databases for efficiency	Denormalized databases for analysis

13. What are the various characteristics of data-warehouse?

ANS. Various features of data-warehouse

- Subject – oriented: A data warehouse typically provides information on a topic (such a sales inventory or supply chain) rather than company operations.
- Time -variant: Time variant keys (e.g., for the date, month, time) are typically present
- Integrated: A data warehouse combines data from various sources. These may include a cloud, relational databases, flat files, structured and semi – structured data, metadata, and master data. The sources are combined in a manner that's consistent, relatable and ideally certifiable, providing a business with confidence in the data's quality.
- Persistent and non-volatile: Prior data isn't deleted when new data is added. Historical data is preserved for comparisons, trends and analytical.

14. What is Star-Schema??

ANS A star schema is a multi-dimensional data model used to organize data in a database so that it is easy to understand and analyze. Star schemas can be applied to data warehouses, databases, data marts, and other tools. The star schema design is optimized for querying large data sets.

Introduced by Ralph Kimball in the 1990s, star schemas are efficient at storing data, maintaining history, and updating data by reducing the duplication of repetitive business definitions, making it fast to aggregate and filter data in the data warehouse

15. What do you mean by SETL?

ANS SETL is a very-high level language with dynamic typing and dynamic data structures, based on the mathematical notion of set. It was designed in the very early 1970s by J. Schwartz – a renowned mathematician, with the help of R. Dewar and others. The language introduced a fundamentally new paradigm in programming in which sets, ordered sets and maps are the principal data structures and the programs are expressed in terms of set constructors, set operations, and predicates on sets. The very name SETL is an abbreviation of ‘SET Language’. SETL programs are much more declarative than procedural. According to its author, the language should present ‘an abstract but nevertheless executable notation for describing algorithms’.