MACHINE LEARNING

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?

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.- (b) Divisive clustering

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

Ans.-(b) Number of clusters

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 an exploratory analysis that tries to identify structures within the data. Cluster analysis is also called segmentation analysis or taxonomy analysis. More specifically, it tries to identify homogenous groups of cases if the grouping is not previously known. Because it

is exploratory, it does not make any distinction between dependent and independent variables. The different cluster analysis methods that SPSS offers can handle binary, nominal, ordinal, and scale (interval or ratio) data.

Cluster analysis is often used in conjunction with other analyses (such as discriminant analysis). The researcher must be able to interpret the cluster analysis based on their understanding of the data to determine if the results produced by the analysis are actually meaningful.The hierarchical cluster analysis follows three basic steps: 1) calculate the distances, 2) link the clusters, and 3) choose a solution by selecting the right number of clusters.

First, we have to select the variables upon which we base our clusters. In the dialog window we add the math, reading, and writing tests to the list of variables. Since we want to cluster cases we leave the rest of the tick marks on the default.

In the dialog box Plots… we should add the Dendrogram. The Dendrogram will graphically show how the clusters are merged and allows us to identify what the appropriate number of clusters is.

The dialog box Method… allows us to specify the distance measure and the clustering method. First, we need to define the correct distance measure. SPSS offers three large blocks of distance measures for interval (scale), counts (ordinal), and binary (nominal) data.

For interval data, the most common is Square Euclidian Distance. It is based on the Euclidian Distance between two observations, which is the square root of the sum of squared distances. Since the Euclidian Distance is squared, it increases the importance of large distances, while weakening the importance of small distances

Conduct and Interpret a Cluster Analysis

What is the Cluster Analysis?

Cluster analysis is an exploratory analysis that tries to identify structures within the data. Cluster analysis is also called segmentation analysis or taxonomy analysis. More specifically, it tries to identify homogenous groups of cases if the grouping is not previously known. Because it is exploratory, it does not make any distinction between dependent and independent variables. The different cluster analysis methods that SPSS offers can handle binary, nominal, ordinal, and scale (interval or ratio) data.

Cluster analysis is often used in conjunction with other analyses (such as discriminant analysis). The researcher must be able to interpret the cluster analysis based on their understanding of the data to determine if the results produced by the analysis are actually meaningful.

Aligning theoretical framework, gathering articles, synthesizing gaps, articulating a clear methodology and data plan, and writing about the theoretical and practical implications of your research are part of our comprehensive dissertation editing services.

Schedule Your FREE Consultation

with a Dissertation Expert Today

Bring dissertation editing expertise to chapters 1-5 in timely manner.

Track all changes, then work with you to bring about scholarly writing.

Ongoing support to address committee feedback, reducing revisions.

Typical research questions the cluster analysis answers are as follows:

Medicine – What are the diagnostic clusters? To answer this question the researcher would devise a diagnostic questionnaire that includes possible symptoms (for example, in psychology, anxiety, depression etc.). The cluster analysis can then identify groups of patients that have similar symptoms.

Marketing – What are the customer segments? To answer this question a market researcher may conduct a survey covering needs, attitudes, demographics, and behavior of customers. The researcher then may use cluster analysis to identify homogenous groups of customers that have similar needs and attitudes.

Education – What are student groups that need special attention? Researchers may measure psychological, aptitude, and achievement characteristics. A cluster analysis then may identify what homogeneous groups exist among students (for example, high achievers in all subjects, or students that excel in certain subjects but fail in others).

Biology – What is the taxonomy of species? Researchers can collect a data set of different plants and note different attributes of their phenotypes. A cluster analysis can group those observations into a series of clusters and help build a taxonomy of groups and subgroups of similar plants.

Other techniques you might want to try in order to identify similar groups of observations are Q-analysis, multi-dimensional scaling (MDS), and latent class analysis.

The Cluster Analysis in SPSS

Our research question for this example cluster analysis is as follows:

What homogenous clusters of students emerge based on standardized test scores in mathematics, reading, and writing?

In SPSS Cluster Analyses can be found in Analyze/Classify…. SPSS offers three methods for the cluster analysis: K-Means Cluster, Hierarchical Cluster, and Two-Step Cluster.

K-means cluster is a method to quickly cluster large data sets. The researcher define the number of clusters in advance. This is useful to test different models with a different assumed number of clusters.

Hierarchical cluster is the most common method. It generates a series of models with cluster solutions from 1 (all cases in one cluster) to n (each case is an individual cluster). Hierarchical

cluster also works with variables as opposed to cases; it can cluster variables together in a manner somewhat similar to factor analysis. In addition, hierarchical cluster analysis can handle nominal, ordinal, and scale data; however it is not recommended to mix different levels of measurement.

Two-step cluster analysis identifies groupings by running pre-clustering first and then by running hierarchical methods. Because it uses a quick cluster algorithm upfront, it can handle large data sets that would take a long time to compute with hierarchical cluster methods. In this respect, it is a combination of the previous two approaches. Two-step clustering can handle scale and ordinal data in the same model, and it automatically selects the number of clusters.

The hierarchical cluster analysis follows three basic steps: 1) calculate the distances, 2) link the clusters, and 3) choose a solution by selecting the right number of clusters.

First, we have to select the variables upon which we base our clusters. In the dialog window we add the math, reading, and writing tests to the list of variables. Since we want to cluster cases we leave the rest of the tick marks on the default.

In the dialog box Statistics… we can specify whether we want to output the proximity matrix (these are the distances calculated in the first step of the analysis) and the predicted cluster membership of the cases in our observations. Again, we leave all settings on default.

In the dialog box Plots… we should add the Dendrogram. The Dendrogram will graphically show how the clusters are merged and allows us to identify what the appropriate number of clusters is.

The dialog box Method… allows us to specify the distance measure and the clustering method. First, we need to define the correct distance measure. SPSS offers three large blocks of distance measures for interval (scale), counts (ordinal), and binary (nominal) data.

For interval data, the most common is Square Euclidian Distance. It is based on the Euclidian Distance between two observations, which is the square root of the sum of squared distances. Since the Euclidian Distance is squared, it increases the importance of large distances, while weakening the importance of small distances.

If we have ordinal data (counts) we can select between Chi-Square or a standardized Chi-Square called Phi-Square. For binary data, the Squared Euclidean Distance is commonly used.

In our example, we choose Interval and Square Euclidean Distance.

Next, we have to choose the Cluster Method. Typically, choices are between-groups linkage (distance between clusters is the average distance of all data points within these clusters), nearest neighbor (single linkage: distance between clusters is the smallest distance between two data points), furthest neighbor (complete linkage: distance is the largest distance between two data points), and Ward's method (distance is the distance of all clusters to the grand average of the sample). Single linkage works best with long chains of clusters, while complete linkage works best with dense blobs of clusters. Between-groups linkage works with both cluster types. It is recommended is to use single linkage first. Although single linkage tends to create chains of clusters, it helps in identifying outliers. After excluding these outliers, we can move onto Ward's method. Ward's method uses the F value (like in ANOVA) to maximize the significance of differences between clusters.

A last consideration is standardization. If the variables have different scales and means we might want to standardize either to Z scores or by centering the scale. We can also transform the values to absolute values if we have a data set where this might be appropriate.

14. How is cluster quality measured?

Measuring Clustering Quality

Suppose you have assessed the clustering tendency of a given data set. You may have also tried to predetermine the number of clusters in the set. You can now apply one or multiple clustering methods to obtain clusterings of the data set. “How good is the clustering generated by a method, and how can we compare the clusterings generated by different methods?”

We have a few methods to choose from for measuring the quality of a clustering. In general, these methods can be categorized into two groups according to whether ground truth is available. Here, ground truth is the ideal clustering that is often built using human experts.

If ground truth is available, it can be used by extrinsic methods, which compare the clustering against the group truth and measure. If the ground truth is unavailable, we can use intrinsic methods, which evaluate the goodness of a clustering by considering how well the clusters are separated. Ground truth can be considered as supervision in the form of “cluster labels.” Hence, extrinsic methods are also known as supervised methods, while intrinsic methods are unsupervised methods.

Let's have a look at simple methods from each category.

Extrinsic Methods

When the ground truth is available, we can compare it with a clustering to assess the clustering. Thus, the core task in extrinsic methods is to assign a score, , to a clustering, , given the ground truth, . Whether an extrinsic method is effective largely depends on the measure, Q, it uses.

In general, a measure Q on clustering quality is effective if it satisfies the following four essential criteria:

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.

Clusters should exhibit high internal homogeneity and high external heterogeneity.

Types of Cluster Analysis

The clustering algorithm needs to be chosen experimentally unless there is a mathematical reason to choose one cluster method over another.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. Agglomerative clustering starts with single objects and starts grouping them into clusters.

The divisive method is another kind of Hierarchical method in which clustering starts with the complete data set and then starts dividing into partitions.

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 centres.

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.