ML Project

Evaluation of Turkiye student dataset

A PROJECT BY

LAXITA SODHANI	2001103
NAMAN SINGH NAYAL	2001117
PANCHADI CHANDANA SAI	2001131
PRITI KUMARI	2001147

INDEX

Introduction

- We are clustering Turkiye student dataset using K-means and hierarchical clustering algorithms
- We have used Principal Component Analysis to reduce the dimensionality of the dataset.

Back to Agenda Page

About the Dataset

Number of features/ attributes

The dataset has a total of 33 features, out of which 28 are feedback questions and others are about the instructor, course code, attendance, difficulty, and the number of times students repeated the course.

Number of instances

The data has feedback from 5820 students.

Provided By

EGazi University in Ankara (Turkey)

Principal Component Analysis

Large datasets are increasingly common and are often difficult to interpret. Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss.

PCA is an adaptive analysis technique as the new variables are defined by the dataset at hand, not a priori. Also, variants of the technique have been developed that are tailored to various different data types and structures

- PCA reduces the dimensionality of a dataset, while preserving as much ‘variability’ (i.e. statistical information) as possible.

- finding new variables that are linear functions of those in the original dataset, that successively maximize variance and that are uncorrelated with each other.

- PCA as a descriptive tool needs no distributional assumptions and, as such, is very much an adaptive exploratory method that can be used on numerical data of various types

BASIC METHOD

- We have a n x p data matrix.
- We seek a linear combination of the columns of matrix X with maximum variance
- Such linear combinations are given by

  $$\sum_{j=1}^{p} a_j x_j = Xa,$$

  where a is a vector of constants a1, a2, … , ap
- The variance of any such linear combination is given by $\mathrm{var}(Xa) = a'Sa,$

  where S is the sample covariance matrix associated with the dataset ‘ and denotes the transpose
- maximizing $a'Sa - \lambda(a'a - 1)$, where λ is a Lagrange multiplier. $a'a = 1$

Back to Agen

- Differentiating with respect to the vector a

$$Sa - \lambda a = 0 \Longleftrightarrow Sa = \lambda a$$

Thus, a must be a (unit-norm) eigenvector, and λ the corresponding eigenvalue, of the covariance matrix S.

- It is these linear combinations Xak that are called the principal components of the dataset
- the elements of the eigenvectors ak are commonly called the PC loadings, whereas the elements of the linear combinations Xak are called the PC scores

KEY ISSUES

- the properties of PCA have some undesirable features when these variables have different units of measurement
- To overcome this undesirable feature, it is common practice to begin by standardizing the variables.
- Since the covariance matrix of a standardized dataset is merely the correlation matrix R of the original dataset, a PCA on the standardized data is also known as a correlation matrix PCA
- Correlation matrix PCs are invariant to linear changes in units of measurement and are therefore the appropriate choice for datasets where different changes of scale are conceivable for each variable

Back to Agen

Methods Of Clustering

Back to Agenda Page

K-Means Clustering

- What is K-Means
- It's Algorithm Properties
- Algorithm Process
- Strength Of K-Means
- Weakness Of K-Means

Hierarchical Clustering

- What is Hierarchial
- Types of Hierarchial
- Strength Of Hierarchial
- Weakness Of Hierarchial

K-Means

- K-Means Clustering is an Unsupervised Learning algorithm.
- It groups the unlabeled dataset into different clusters

Back to Agenda Page

K-Means Algorithm Properties

There are always K clusters.

There is always at least one item in each cluster

The clusters are non-hierarchical and they do not overlap

Back to Agenda Page

K-Means Algorithm Process

The dataset is partitioned into K clusters

The data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.

Back to Agenda Page

Weaknesses of k-means

The algorithm is only applicable if the mean is defined

The algorithm is sensitive to outliers

The user needs to specify k.

The k-means algorithm is not suitable for discovering clusters that are not hyperellipsoids (or hyper-spheres).

Back to Agenda Page

Strengths of K-Means

Simple: Easy to understand and to implement.

Efficient:
Time complexity: O(tkn),
where , n = number of data points,
k = number of clusters
t = number of iterations

Since both k and t are small. k-Means is considered a linear algorithm

Back to Agenda Page

Hierarchical

- A hierarchical method creates a hierarchical
- decomposition of the given set of data objects. A
- dendrograms is built due to Tree of clusters. Every cluster node contains child clusters, sibling cluster partition the points covered by their common parents

Back to Agenda Page

Hierarchial Clustering

In hierarchical clustering, Researchers assign each item to a cluster such that if Researchers have N items then Researchers have N clusters.

Find closest pair of clusters and merge them into a single cluster. Compute distance between new cluster and each of old clusters. Researchers have to repeat these steps until all items are clustered into K no. of clusters

Back to Agenda Page

Strengths of Hierarchial

Conceptually Simple.

Theoretical properties are well understood.

When Clusters are merged /split, the decision is permanent => the number of different alternatives that need to be examined is reduced.

Back to Agenda Page

Weakness of Hierarchial

Merging /splitting of clusters is permanent
Erroneous decisions are impossible to correct later.
Divisive methods can be computational hard.
Methods are not (necessarily) scalable for large datasets.

Back to Agenda Page

Types of Hierarchical Clustering

Agglomerative (bottom up)

- Start with each document being a single cluster.
- Eventually all documents belong to the same cluster

Divisive (top down)-

- Starts with all documents belong to the same cluster.
- Eventually each node forms a cluster on its own.

Back to Agenda Page

K-Means Vs Hierarchical

PROPERTIES	KMEANS	HIERARCHIAL
Definition	K Means Clustering generates a specific number of disjoint, flat (non-hierarchical) Clusters	Hierarchical Clusteringmethod construct a hierarchy of Clustering, not just a single partition of objects.
Clustering Criteria	It is well suited for generating globular Cluster.	Use a distance matrix as Clustering Criteria. Atermination Condition can be used .Example –A number of Clusters.
Performance	The performance of K- mean algorithm is better than Hierarchical Clustering Algorithm.	Hierarchical Clustering Algorithm performance is less as compare to K- mean algorithm.

Categorial Data	K- Means can be used in categorical data is first converted into numeric by assigning rank.	Hierarchical algorithm was adopted for categorical data, and due to its complexity a new approach for assigning rank value to each categorical attribute.
Sensitive to Noise	k-Means is very sensitive to noise in the dataset.	It is less sensitive to noise in the dataset
Cluster	There are always K clusters	The number of Clusters k is not required as an input.

Execution Time	K -mean algorithm also increases its time of execution.	Hierarchical algorithm its performance is better.
Quality	K-Means algorithms Shows less quality	Hierarchical algorithm shows more quality.
Dataset	k -mean algorithm is good for large dataset.	Hierarchical is good for small datasets

Conclusion 

Generally using huge dataset, K-Means algorithm is faster than other clustering algorithm and also produces quality clusters.

Back to Agenda Page

Resources

Topics	Links
Resources of K-means and Hierarchical	https://www.researchgate.net/publication/293061584_Comparative_Study_of_K-Means_and_Hierarchical_Clustering_Techniques
Resources of PCA	https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2015.0202

Back to Agenda Page