Google Developer Student Clubs
National Higher School of Artificial Intelligence

Challenge 3

Clustering for Heart Disease Analysis

Google Developer Student Club - ENSIA

This page is intentionally left blank

Unsupervised learning
Challenge 3 - AiFest'23
K-Means Clustering for Heart Disease Analysis

Introduction:

Clustering algorithms play a crucial role in grouping similar items together. The application of unsupervised learning algorithms can provide valuable insights in various industries. For instance, retailers can effectively target customers by clustering them based on similarities, biologists can identify plants with shared characteristics, clustering can also be used to group similar documents together based on their content, allowing for effective document organization, etc. In this particular challenge, we will investigate the suitability of employing clustering algorithms to group medical patients.

Description:

Doctors often analyze past cases to enhance their understanding of optimal treatment approaches for their patients. When a patient exhibits similar health history or symptoms to a previous case, they may benefit from undergoing a similar treatment regimen.

This challenge aims to explore the possibility of grouping patients based on common characteristics using unsupervised learning techniques. Specifically, we will investigate the use of the k-means algorithm. By employing this method, we can identify clusters of patients with similar attributes, facilitating targeted treatment strategies.

Objectives:

In this challenge, our focus is on examining anonymized patients who have received a diagnosis of heart disease. By identifying patients who share similar characteristics, we can gain insights into the effectiveness of specific treatments. This information would be valuable for doctors, as they could learn from the outcomes of patients similar to those they are treating. The dataset originates from the V.A. Medical Center in Long Beach, California.

Tasks:

1. **Targeting treatment for heart disease patients**

Before beginning a project, it is important to get an idea of what the patient data looks like. In addition, the clustering algorithms used below require that the data be numeric, so it is necessary to ensure the patient data does not need any transformations. You will also be brushing up on your base Python skills for some analysis.

2. **Quantifying patient differences**

It is important to conduct some exploratory data analysis to familiarize ourselves with the data before clustering. This step enables us to understand the variables better and make an informed choice regarding whether data scaling is necessary. Since k-means clustering measures similarity using a distance formula, it tends to prioritize variables with larger scales, which can result in greater differences between data points.

Exploratory data analysis helps us to understand the characteristics of the patients in the data. We need to get an idea of the value ranges of the variables and their distributions. This will also be helpful when we evaluate the clusters of patients from the algorithms. For example, are there more patients of one gender? What might an outlier look like?

3. Let's start grouping patients

Once we have figured out if we need to modify the data and made any necessary changes, we can now start the clustering process. For the k-means algorithm, it is necessary to select the number of clusters in advance. It is also important to make sure that your results are reproducible when conducting a statistical analysis. This means that when someone runs your code on the same data, they will get the same results as you reported. Therefore, if you are conducting an analysis that has a random aspect, it is necessary to set a seed to ensure reproducibility.

4. Another round of k-means

Because the k-means algorithm initially selects the cluster centers by randomly selecting points, different iterations of the algorithm can result in different clusters being created. If the algorithm is truly grouping together similar observations, then cluster assignments will be somewhat robust between different iterations of the algorithm.

Regarding the heart disease data, this would mean that the same patients would be grouped together even when the algorithm is initialized at different random points. If patients are not in similar clusters with various algorithm runs, then the clustering method is not picking up on meaningful relationships between patients.

We are going to explore how the patients are grouped together with another iteration of the k-means algorithm. We will then be able to compare the resulting groups of patients.

5. Comparing patient clusters

It is important that the clusters resulting from k-means are stable. Even though the algorithm begins by randomly initializing the cluster centers, if the k-means algorithm is the right choice for the data, then different initialization of the algorithm will result in similar clusters.

The clusters from different iterations may not be exactly the same, but the clusters should be roughly the same size and have similar distributions of variables. If there is a lot of change in clusters between different iterations of the algorithm, then k-means clustering is not a good choice for the data.

It is not possible to validate that the clusters obtained from an algorithm are ground truth accurate, since there is no true labeling for patients. Thus, it is necessary to examine how the clusters change between different iterations of the algorithm. We are going to use some visualizations to get an idea of the cluster stability. That way, we can see how certain patient characteristics may have been used to group patients together.

6. Comparing clustering results

The doctors are interested in grouping similar patients together in order to determine appropriate treatments. Therefore, they want to have clusters with more than a few patients to see different treatment options. While it is possible for a patient to be in a cluster by themselves, this means that the treatment they received might not be recommended for someone else in the group.

As with the k-means algorithm, the way to evaluate the clusters is to investigate which patients are being grouped together. Are there patterns evident in the cluster assignments, or do they seem to be groups of noise ?

7. Visualizing the cluster contents

In addition to looking at the distributions of variables in each of the clusters, we will make visualizations to evaluate the algorithms. Even though the data has more than two dimensions, we can get an idea of how the data clusters by looking at a scatter plot of two variables. We want to look for patterns that appear in the data and see which patients get clustered together.

8. Conclusion

For the k-means algorithm, it is imperative that similar clusters are produced for each iteration of the algorithm. We want to make sure that the algorithm is clustering signal as opposed to noise.

For the sake of the doctors, we also want to have multiple patients in each group, so they can compare treatments. We only did some preliminary work to explore the performance of the algorithms, and it is necessary to explore further before making a recommendation. Based on the above analysis, are there any algorithms that you would want to investigate further to group patients?

Remember that it is important the k-mean algorithm seems stable when running multiple iterations. This means that we would see similar groups of patients showing up in the plots from the different iterations of the algorithm.

Global Dataset Description

N°	Feature	Type	Description	Units
1	id	Discrete	id of the patient	
2	age	Discrete	age	years
3	sex	Categorical	sex of the patient (1=male, 0=female)	
4	dataset	Categorical	origin of instance	
5	cp	Categorical	chest pain type	
6	trestbps	Discrete	resting blood pressure (on admission to the hospital)	mm Hg
7	chol	Discrete	cholesterol measurement	mg/dl
8	fbs	Categorical	fasting blood sugar > 120 mg/dl	
9	restecg	Categorical	resting electrocardiographic results	
10	thalach	Discrete	maximum heart rate achieved	
11	exang	Categorical	exercise induced angina (1 = yes, 0 = no)	
12	oldpeak	Discrete	ST depression induced by exercise relative to rest	

N°	Feature	Type	Description	Units
13	slope	Categorical	slope of the peak exercise ST segment	
14	ca	Discrete	number of major vessels (0-3)	
15	thal	Categorical	a blood disorder called thalassemia	

Dataset Detailed Description

dataset: origin of the instance: Cleveland, Hungary, VA Long Beach, Switzerland

cp: chest pain type

— Value 0: asymptomatic

— Value 1: atypical angina

— Value 2: non-anginal pain

— Value 3: typical angina

fbs: fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)

restecg: resting electrocardiographic results

— Value 0: showing probable left ventricular hypertrophy by Estes' criteria

— Value 1: normal

— Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

oldpeak: ST depression induced by exercise relative to rest (‘ST’ relates to positions on the ECG plot. See more here)

slope: the slope of the peak exercise ST segment
— 0: down sloping
— 1: flat
— 2: up sloping

thal: A blood disorder called thalassemia
— Value 0: NULL (dropped from the dataset previously
— Value 1: fixed defect (no blood flow in some part of the heart)
— Value 2: normal blood flow
— Value 3: reversible defect (a blood flow is observed, but it is not normal)