Assignment – Clustering and PCA Part 2

KAVITHA M

Question 1: Assignment Summary

Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly(why you took that many numbers of principal components, which type of Clustering produced a better result and so on)

Note: You don't have to include any images, equations or graphs for this question. Just text should be enough.

Solution

- **Problem Statement:** HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.
- A lot of highly correlated variables exist, hence the usage of PCA is justified. Rescaling the data and applying PCA, over 90% of the data is properly explained by the first 3 principal components
- Analyzing the 1st 3 components if they suffice and implementing K-means and Hierarchical clustering on the data, **we use the clusters formed during K-means clustering to find the countries that we require since Hierarchical clustering is not showing proper clusters here. For K-means part, we got Cluster 2 and 4 might be the ones which has a proper need of aid**

Question 2: Clustering

a) Compare and contrast K-means Clustering and Hierarchical Clustering.

- The most important difference is the **hierarchy**. Actually, there are two different approaches that fall under this name: top-down and bottom-up.
- In top-down hierarchical clustering, we divide the data into 2 clusters (using k-means with k=2 for eg) Then, for each cluster, we can repeat this process, until all the clusters are too small or too similar for further clustering to make sense, or until we reach a preset number of clusters
- In bottom-up hierarchical clustering, we start with each data item having its own cluster. We then look for the two items that are most similar, and combine them in a larger cluster. We keep repeating until all the clusters we have left are too dissimilar to be gathered together, or until we reach a preset number of clusters.
- In k-means clustering, we try to identify the best way to divide the data into k sets simultaneously. A good approach is to take k items from the data set as initial cluster representatives, assign all items to the cluster whose representative is closest, and then calculate the cluster mean as the new representative, until it converges (all clusters stay the same).

Question 2: Clustering

b) Briefly explain the steps of the K-means clustering algorithm

- *Step one: Initialize cluster centers*
- *Step two: Assign observations to the closest cluster center*
- Once we have these cluster centers, we can assign each point to the clusters based on the minimum distance to the cluster center
- *Step three: Revise cluster centers as mean of assigned observations.*
- *Step four: Repeat step 2 and step 3 until convergence*

Question 2: Clustering

c) How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well as the business aspect of it

The basic idea behind this method is that it plots the various **values** of cost with changing **k**. As the **value of K** increases, there will be fewer elements in the **cluster**. So average distortion will decrease. The lesser number of elements **means** closer to the centroid

Question 2: Clustering

d) Explain the necessity for scaling/standardization before performing Clustering .

The idea **is** that if different components of **data** (features) have different **scales**, then derivatives tend to align along directions with higher variance, which leads to poorer/slower convergence. Standardization **is** an important step of **Data** preprocessing

Question 2: Clustering

e) Explain the different linkages used in Hierarchical Clustering.

- In complete-link (or complete linkage) hierarchical clustering, we merge in each step the two clusters whose merger has the smallest diameter (or: the two clusters with the smallest **maximum** pairwise distance). Complete-link clustering can also be described using the concept of clique. Let dn be the diameter of the cluster created in step n of complete-link clustering. Define graph G(n) as the graph that links all data points with a distance of at most dn. Then the clusters after step n are the cliques of G(n). This motivates the term complete-link clustering.
- In single-link (or single linkage) hierarchical clustering, we merge in each step the two clusters whose two closest members have the smallest distance (or: the two clusters with the smallest **minimum** pairwise distance). Single-link clustering can also be described in graph theoretical terms. If dn is the distance of the two clusters merged in step n, and G(n) is the graph that links all data points with a distance of at most dn, then the clusters after step n are the connected components of G(n). A single-link clustering also closely corresponds to a weighted graph's minimum spinning tree
- Average-link is a compromise between the sensitivity of complete-link clustering to outliers and the tendency of single-link clustering to form long chains that do not correspond to the intuitive notion of clusters as compact, spherical objects.

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.

- The primary application of PCA is dimension reduction
- It can be used in different domains as below,
- Facial recognition
- Image compressions in image processing
- A major stand in computer visions

Question 3: Principal Component Analysis

b) Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.

- **Basis Transformation PCA** simply takes points expressed in the standard **basis** and transforms them into points expressed in an eigenvector **basis**. In this process of **transformation**, some dimensions with low variance are discarded and hence the resulting dimensional reduction

- **Variance as information** "**variance**" means summative **variance** or multivariate **variability** or overall **variability** or total **variability**. Below is the **covariance** matrix of some 3 variables. Their **variances** are on the diagonal, and the sum of the 3 values (3.448) is the overall **variability**.

Question 3: Principal Component Analysis

c) State at least three shortcomings of using Principal Component Analysis.

- **Relies on linear assumptions** if the data is not linearly correlated, PCA is not enough
- **Relies on orthogonal tranformations** principal components are orthogonal to the others it's a restriction to find projections with the highest variance
- PCA has no explicit model! It can be like taking a raw mean (which is usually a great and meaningful measure of central tendency, but not always)