3.0 - Unsupervised Models

Giovanni Della Lunga
giovanni.dellalunga@unibo.it

Introduction to Machine Learning for Finance

Bologna - February, 2022

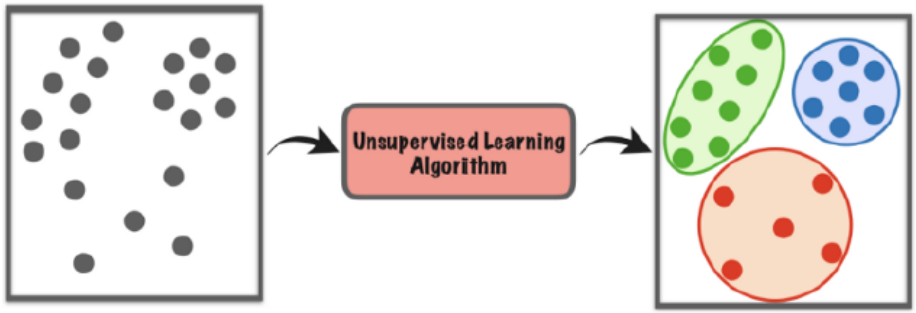
Outline

Subsection 1

What is Unsupervised Learning?

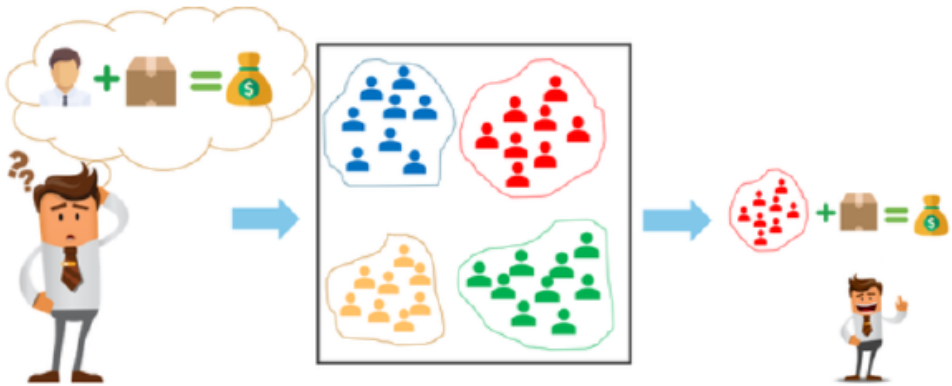
Unsupervised Learning

- Unsupervised learning algorithms, on the other hand, work with data that isn't explicitly labelled.
- Unsupervised algorithms attempt to find some sort of underlying structure in the data.
- Are some observations clustered into groups? Are there interesting relationships between different features? Which features carry most of the information?



Unsupervised Learning

- In unsupervised learning we are not trying to predict anything
- The objective is to cluster data to increase our understanding of the environment
- **Example - Clustering Customers**
  - Suppose you are a bank and have hundreds of thousands of customers and 100 features describing each one
  - Unsupervised learning algorithms can be used to divide your customers into clusters so that you can anticipate their needs and communicate with them more effectively



Unsupervised Learning

- Also in contrast to supervised learning, assessing performance of an unsupervised learning algorithm is somewhat subjective and largely depend on the specific details of the task.
- Unsupervised learning is commonly used in tasks such as text mining and dimensionality reduction.
- K-means is an example of an unsupervised learning algorithm.

Subsection 2

Example: k-Means Clustering

The *k*-Means Algorithm

- In this section we explain a simple clustering procedure known as the *k-means algorithm*;
- *k*-means clustering is one of the simplest and popular unsupervised machine learning algorithms.
- Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes.
- The objective of K-means is simple: group similar data points together and discover underlying patterns.
- To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset.

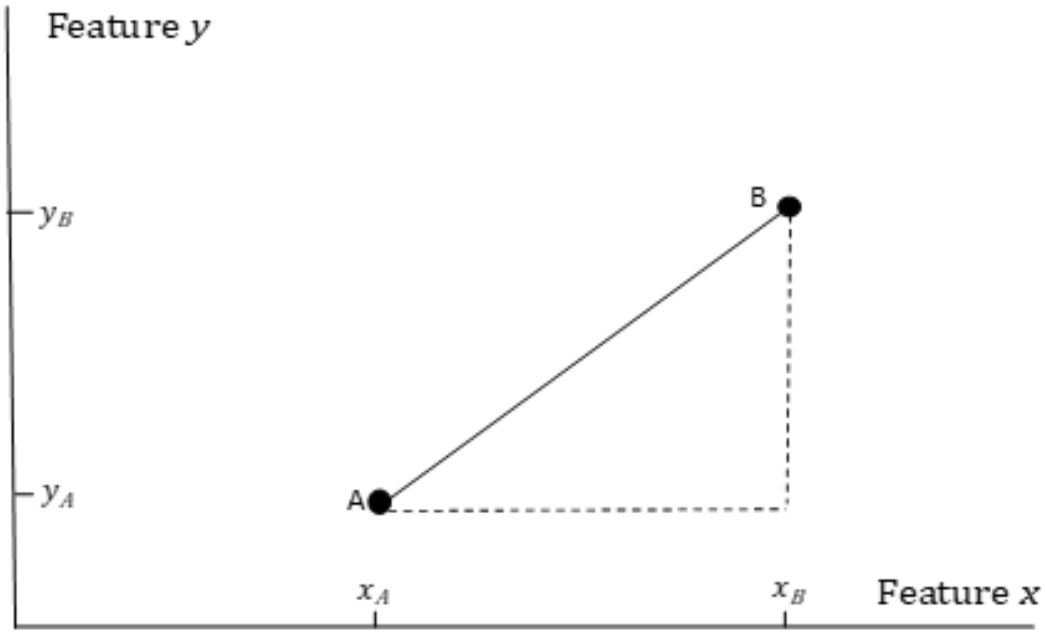
The *k*-Means Algorithm

- A cluster refers to a collection of data points aggregated together because of certain similarities.
- You'll define a target number k, which refers to the number of centroids you need in the dataset.
- A centroid is the imaginary or real location representing the center of the cluster.
- Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares.
- In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the **nearest** cluster, while keeping the centroids as small as possible.
- The 'means' in the K-means refers to averaging of the data; that is, finding the centroid.

The k-Means Algorithm

A Distance Measure

- For clustering we need a distance measure
- The simplest distance measure is the Euclidean distance measure.
  Distance $= \sqrt{(x_B - x_B)^2 + (y_B - y_A)^2}$



The k-Means Algorithm

Distance Measure

- In general when there are m features the distance between P and Q is

$$d = \sqrt{\sum_{j=1}^{m} (\nu_{pj} - \nu_{qj})^2} \tag{1}$$

where ν_{pj} and ν_{qj} are the values of the $j-th$ feature for P and Q

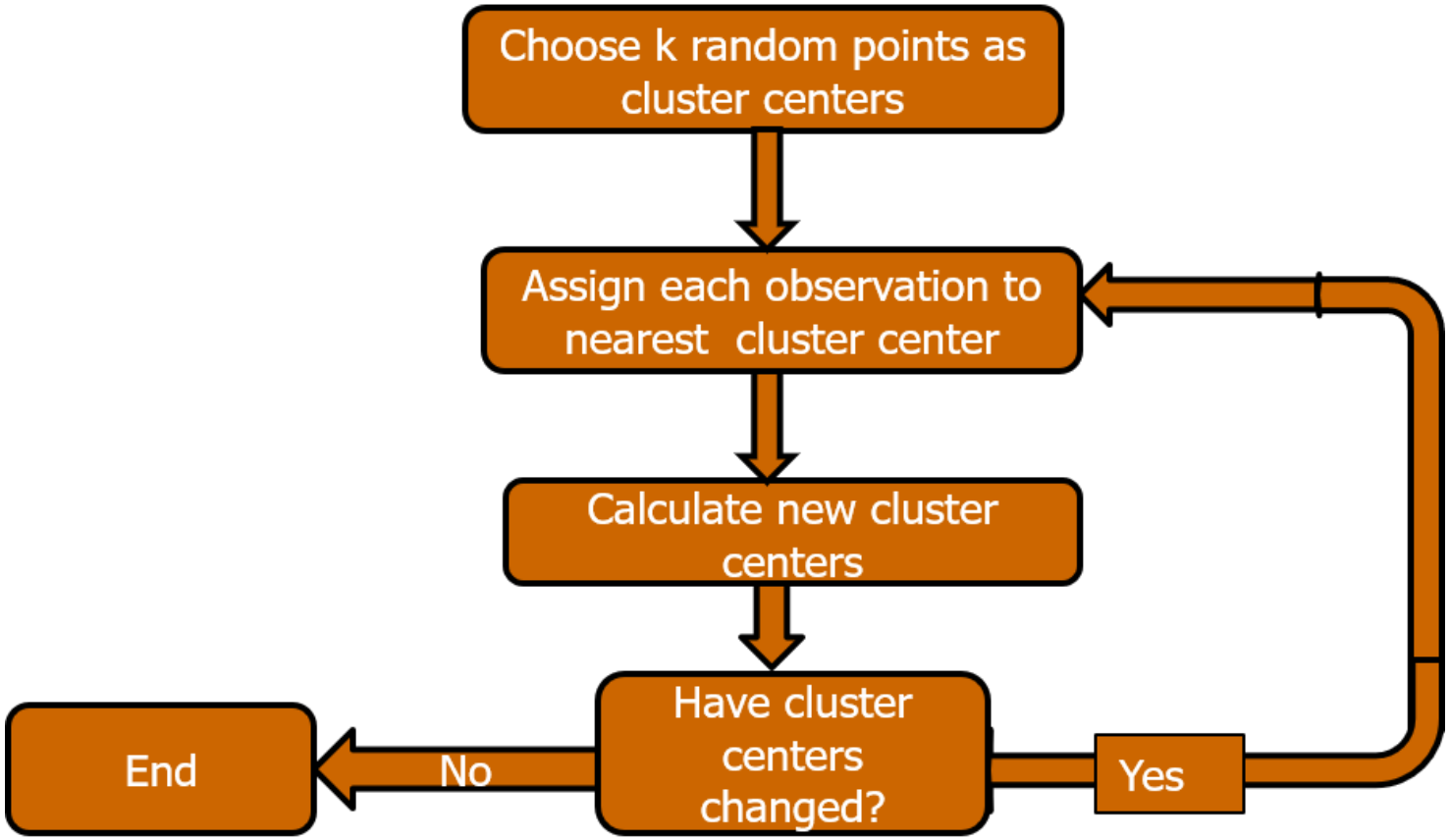
The k-Means Algorithm

Cluster Centers

The center of a cluster (sometimes called the **centroid**) is determined by averaging the values of each feature for all points in the cluster.
Example:

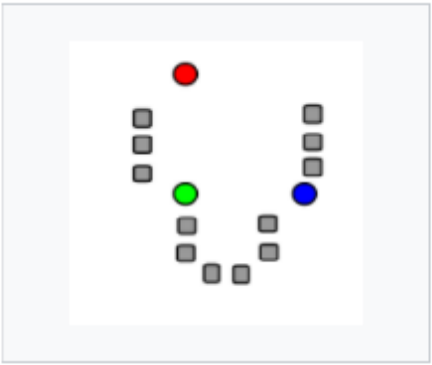
Observ.	Feature 1	Feature 2	Feature 3	Feature 4	Distance to center
1	1.00	1.00	0.40	0.25	0.145
2	0.80	1.20	0.25	0.40	0.258
3	0.82	1.05	0.35	0.50	0.206
4	1.10	0.80	0.21	0.23	0.303
5	0.85	0.90	0.37	0.27	0.137
Center	0.914	0.990	0.316	0.330	

The k-Means Algorithm to find k Clusters

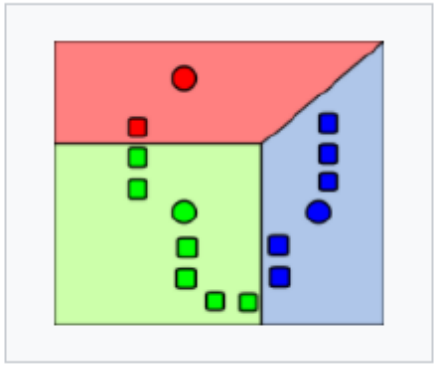


The k-Means Algorithm to find k Clusters

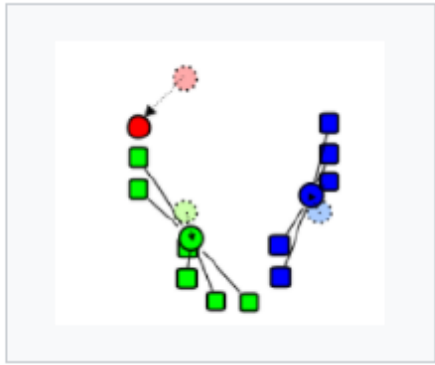
Demonstration of the standard algorithm



1. k initial "means" (in this case k=3) are randomly generated within the data domain (shown in color).



2. k clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



3. The centroid of each of the k clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

The *k*-Means Algorithm

Cluster Centers

Inertia

- A measure of the performance of the algorithm is the within cluster sum of squares also known as *inertia*;
- For any given k the objective is to minimize inertia:

$$Inertia = \sum_{i=1}^{n} d_i^2 \tag{2}$$

where d_i is the distance of observation i from its cluster center

- In practice we use the k-means algorithm with several different starting points and choose the result that has the smallest inertia

The k-Means Algorithm

Choosing k

- The elbow approach (see next slide)
- The silhouette method:
  - For each observation i calculate $a(i)$, the average distance from other observations in its cluster, and $b(i)$, the average distance from observations in the closest other cluster. The silhouette score for observation i, $s(i)$, is defined as

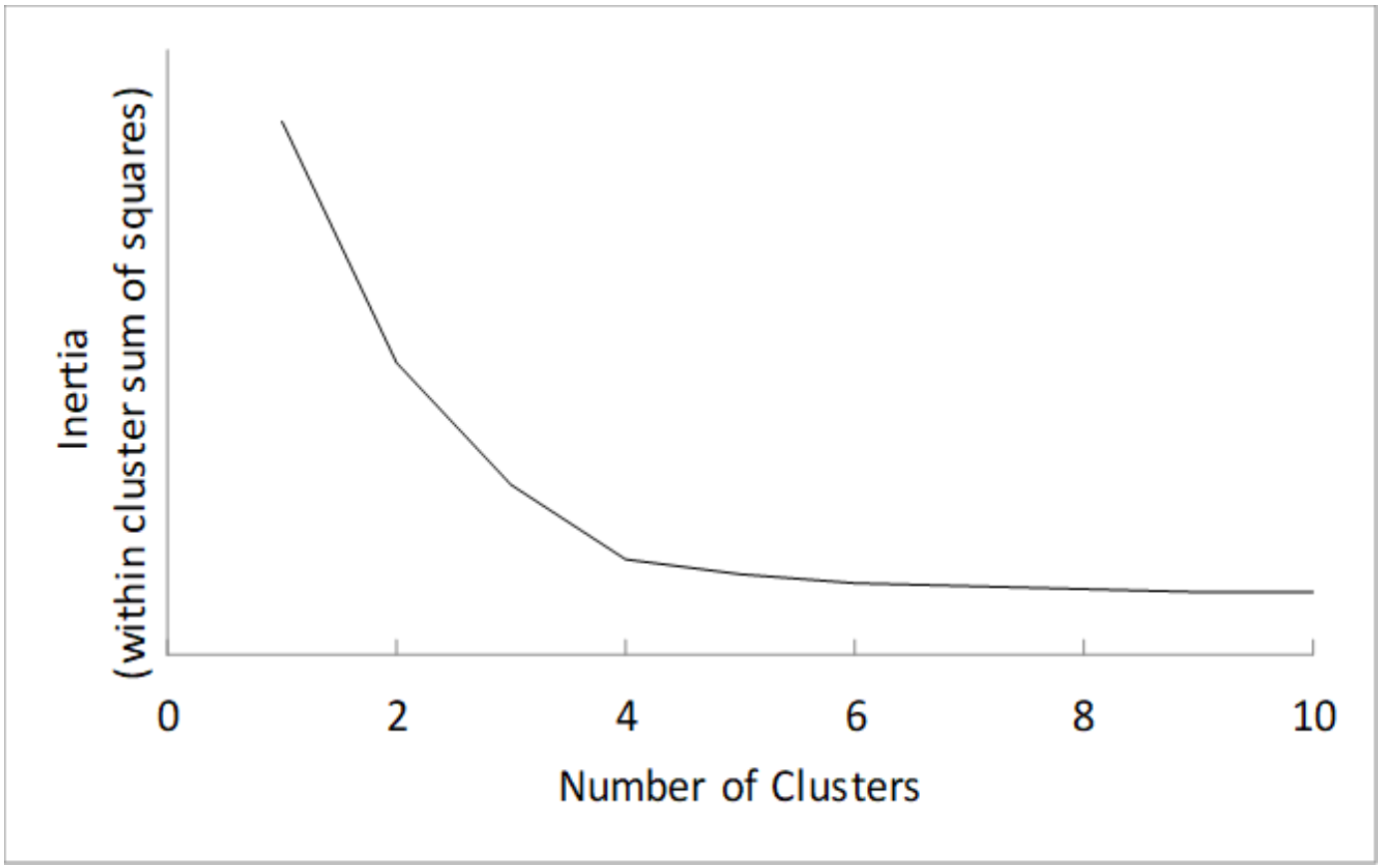
$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]} \tag{3}$$

  - Choose the number of clusters that maximizes the average silhouette score across all observations
- Use the gap statistic which compares the within cluster sum of squares with what would be expected with random data

The k-Means Algorithm

The elbow method

The **elbow method** (In this example k=4 is suggested)



The k-Means Algorithm

The Curse of Dimensionality

- The Euclidean distance measure increases as the number of features increase.
- This is referred to as the curse of dimensionality
- Consider two observations that have values for feature j equal to x_j and y_j. An alternative distance measure that always lies between 0 and 2 is

$$d = 1 - \frac{\sum_{j=1}^{m} x_j y_j}{\sqrt{\sum_{j=1}^{m} x_j^2 \sum_{j=1}^{m} y_j^2}} \tag{4}$$

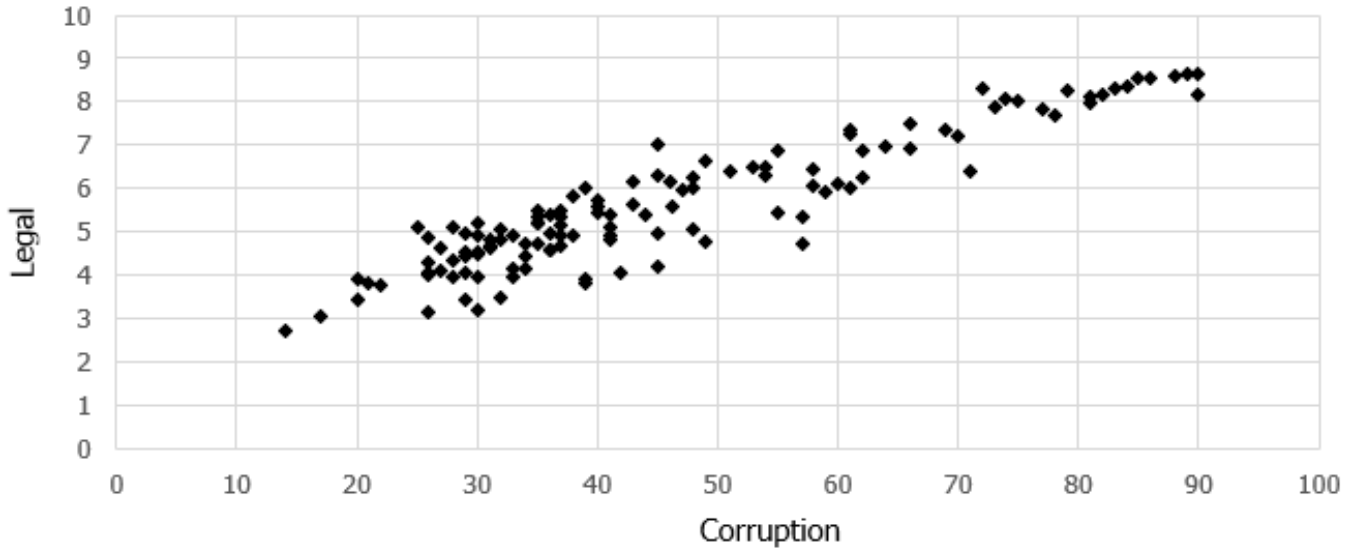
Subsection 3

Country Risk Example

Country Risk Case

- Objective is to cluster countries according to their riskiness for foreign investment
- Measures of Country Risk
- GDP growth rate (IMF)
- Corruption index (Transparency international)
- Peace index (Institute for Economics and Peace)
- Legal Risk Index (Property Rights Association)
- Collected data on 122 countries. Used Z-score scaling.

Country Risk Case



Corruption and legal risk were highly correlated therefore analysis based on

- GDP growth rate
- Peace index
- Legal risk index

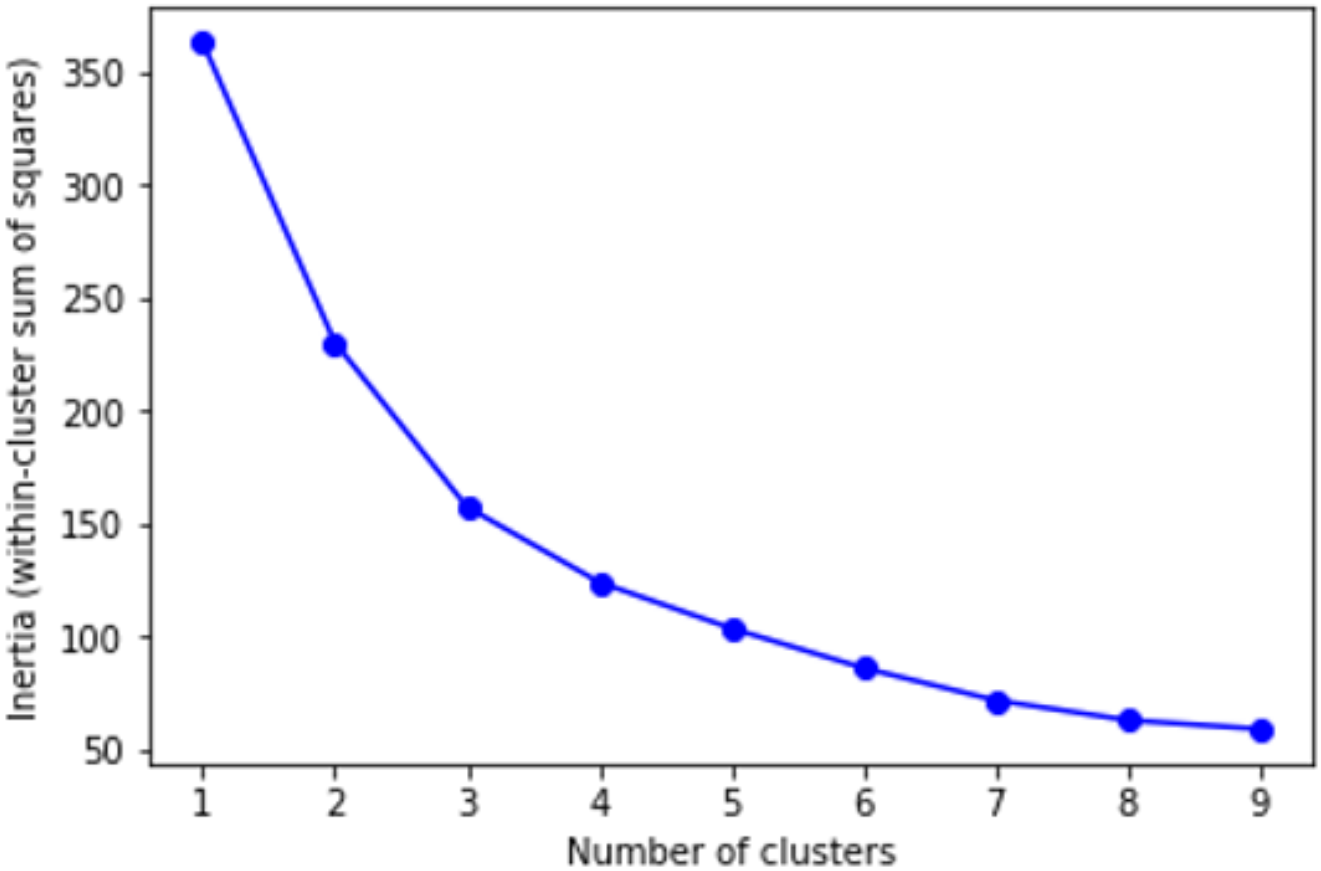
Country Risk Case

How the total within-cluster sum of squares declines as k increases when k-means algorithm is used

Number of clusters	Average silhouette score
2	0.363
3	0.388
4	0.370
5	0.309
6	0.303
7	0.315
8	0.321
9	0.292
10	0.305

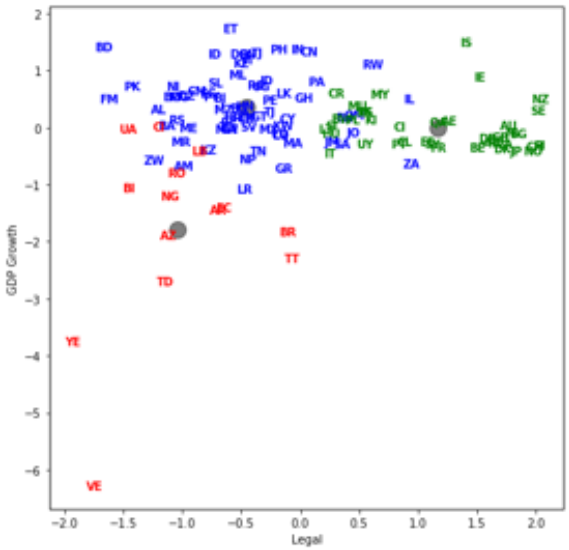
Country Risk Case

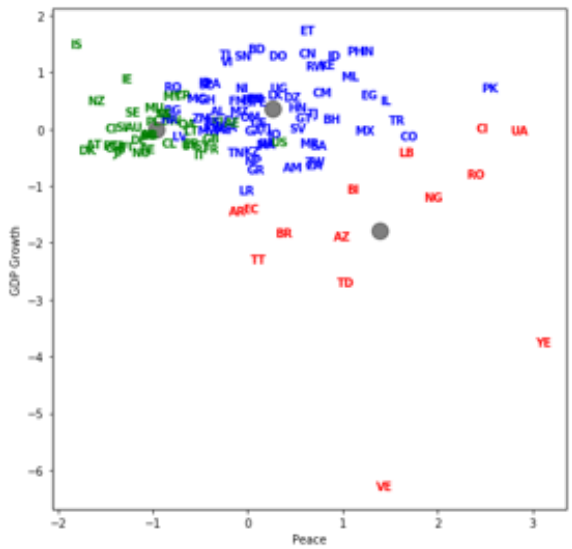
Silhouette scores suggest k=3

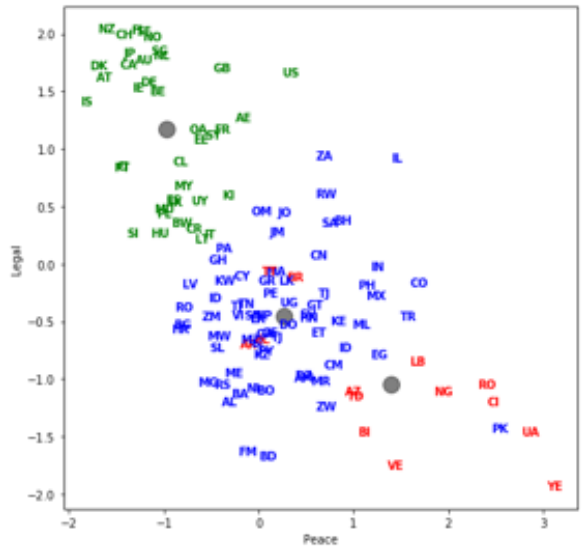


Country Risk Case

The three-cluster results: Green = Low country risk, Blue = Medium country risk, Red = High country risk







Country Risk Case

Cluster centers (scaled values). Note that high values for the peace index are bad whereas high values for the legal risk index are good.

	Peace index	Legal index	GDP
High risk	1.39	−1.04	−1.79
Moderate risk	0.27	−0.45	0.36
Low risk	−0.97	1.17	0.00

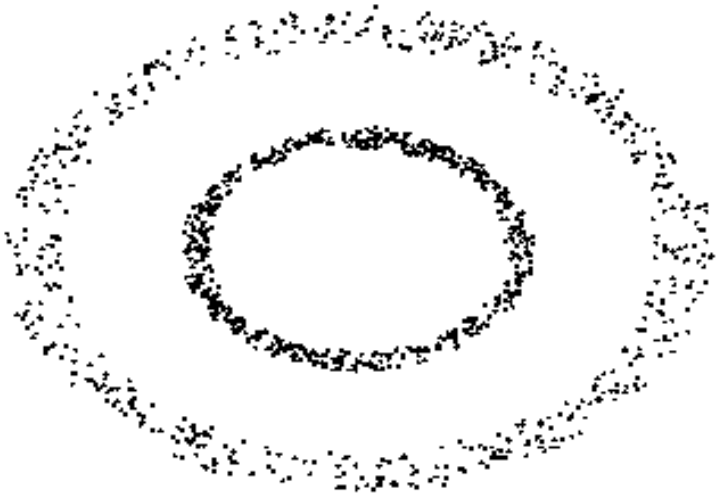
Hierarchical Clustering

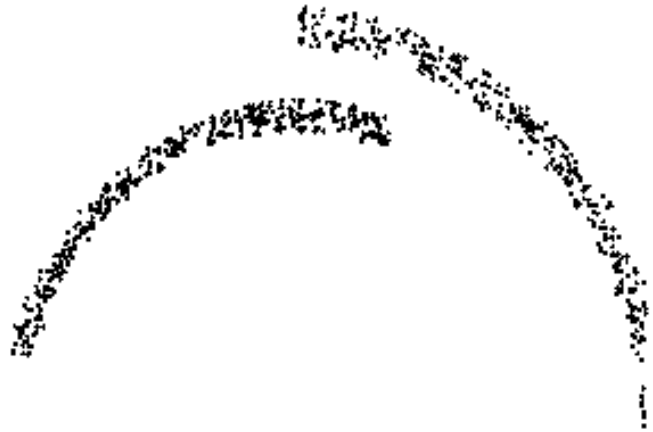
- Start with each observation in its own cluster
- Combine the two closest clusters
- Continue until all observations have been combined into a single cluster
- Can be implemented in Python with AgglomerativeClustering.
- Measures of closeness of clusters:
- Average Euclidean distance between points in clusters
- Maximum distance between points in clusters
- Minimum distance between points in clusters
- Increase in inertia (a version of Ward's method)

Density-based clustering

- Forms clusters based on the closeness of individual observations
- Unlike k-means the algorithm, it is not based on cluster centers.
- We might initially choose 8 observations that are close. After that we add an observation to the cluster if it is close to at least 5 other observations in the cluster, and repeat.

Density-based Clustering Examples





Bibliography

- John C. Hull, *Machine Learning in Business: An Introduction to the World of Data Science*, Amazon, 2019.
- Paul Wilmott, *Machine Learning: An Applied Mathematics Introduction*, Panda Ohana Publishing, 2019.