Machine Learning per la Finanza

Docenti: Zelda Marino e Paolo Zanetti

Email: {zelda.marino;paolo.zanetti}@uniparthenope.it

Unsupervised learning

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

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

Feature Scaling

Before using many ML algorithms (including those for unsupervised learning), it is important to scale feature values so that they are comparable.

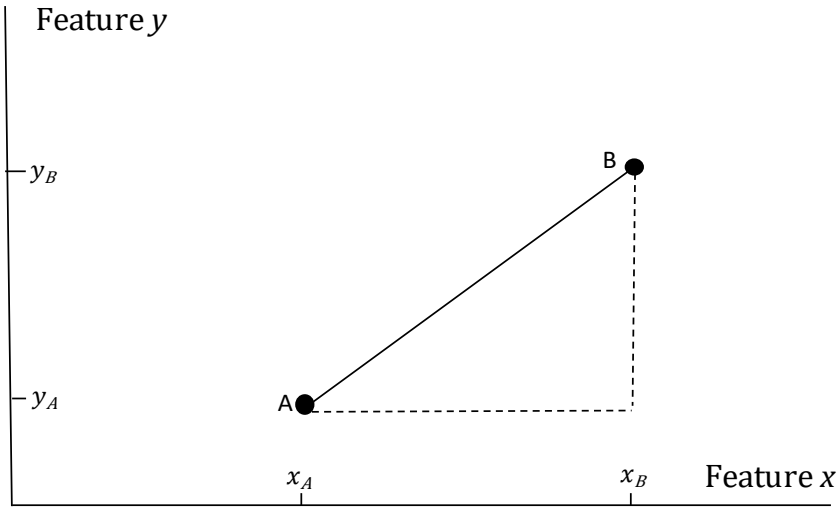
Z-score scaling: $Value \rightarrow \frac{Value - Mean}{SD}$

Min-Max scaling: $Value \rightarrow \frac{Value - Minimum}{Maximum - Minimum}$

A Distance Measure

- For clustering we need a distance measure
- The simplest distance measure is the Euclidean distance measure.

$$distance = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}$$



Distance Measure continued

- When there are m features the distance between P and Q is

$$\sqrt{\sum_{j=1}^{m} \left(v_{pj} - v_{qj}\right)^2}$$

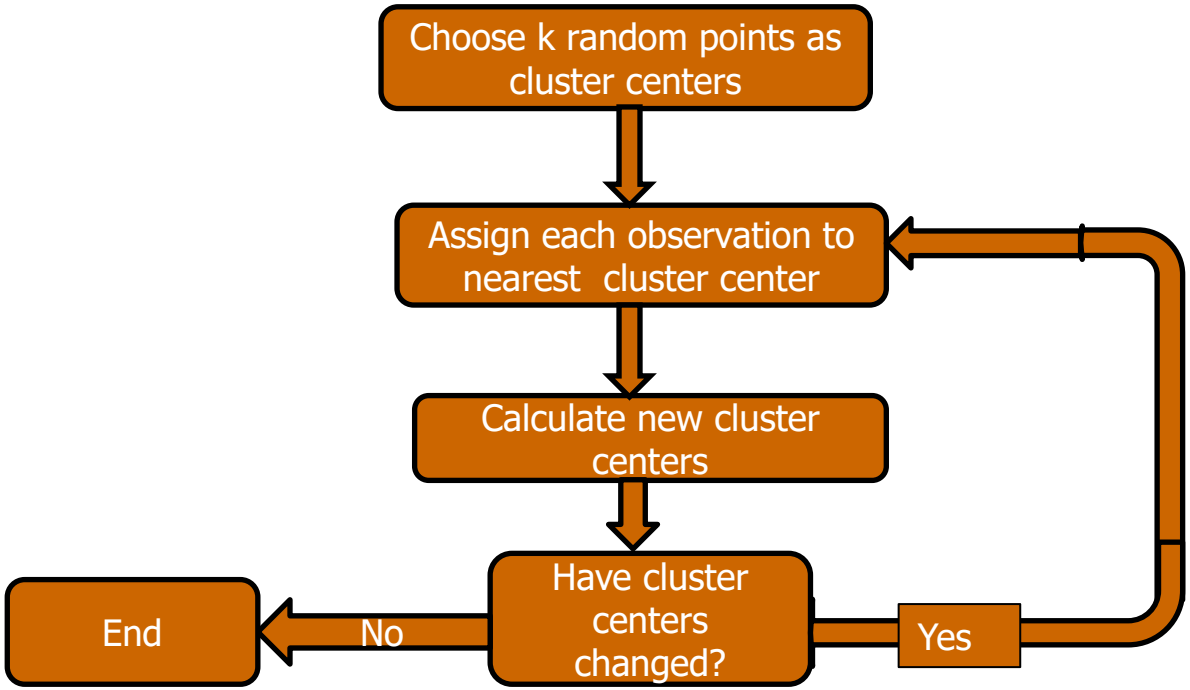
where v_{pj} and v_{qj} and the values of the j-th feature for P and Q

Cluster Centers

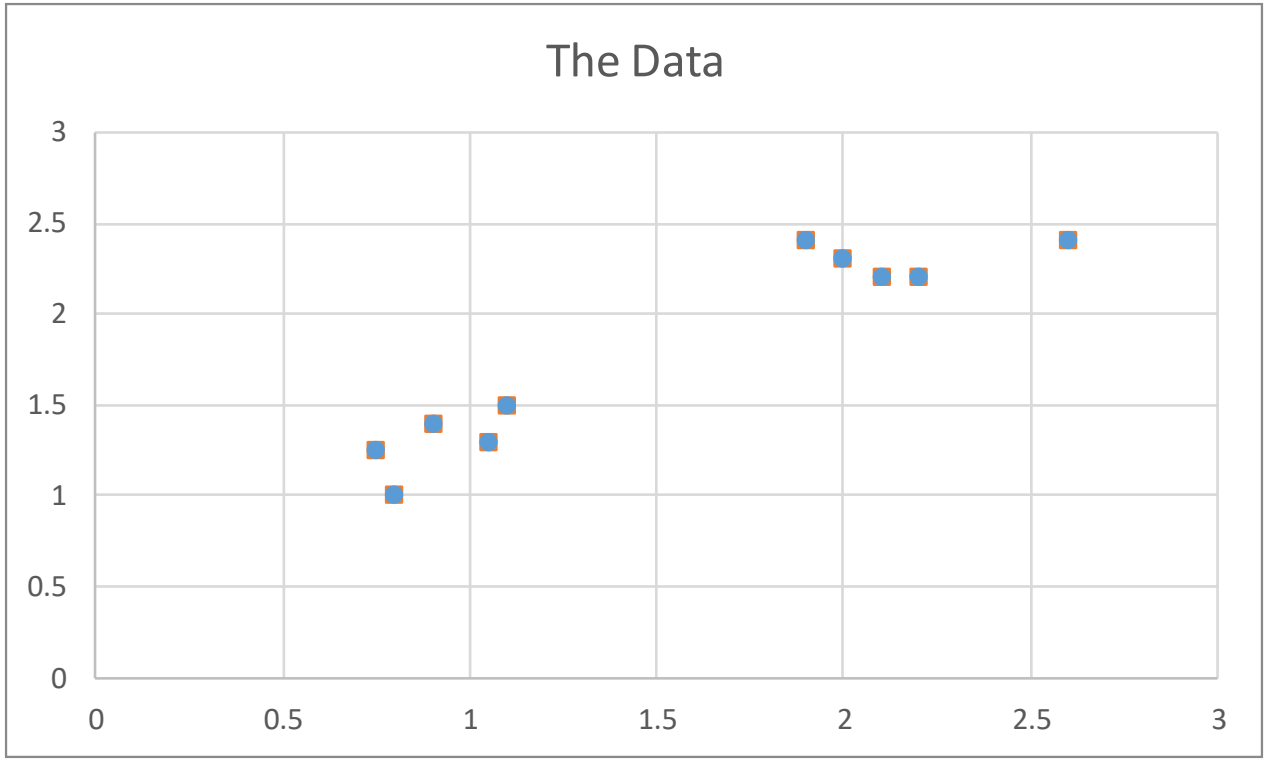
- The **center of a cluster** (sometimes called the centroid) is determined by averaging the values of each feature for all points in the cluster.

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	

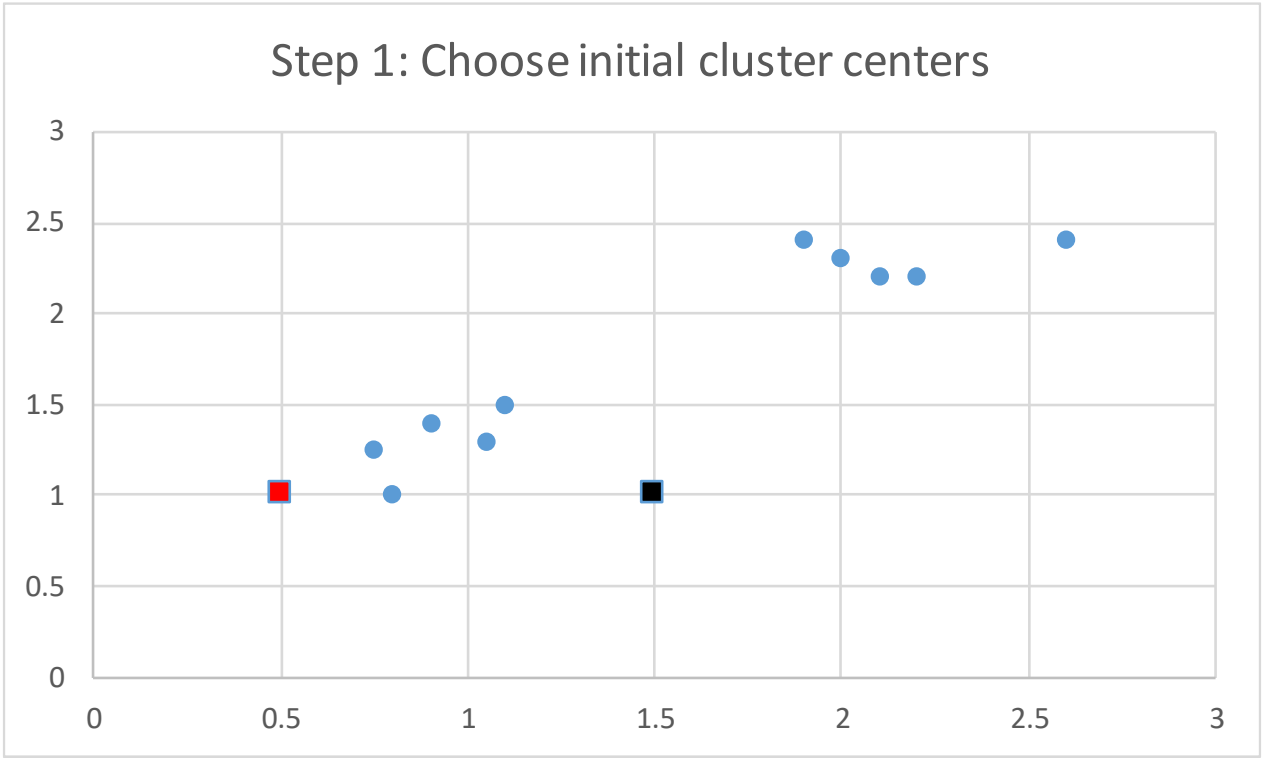
k-means algorithm to find k clusters



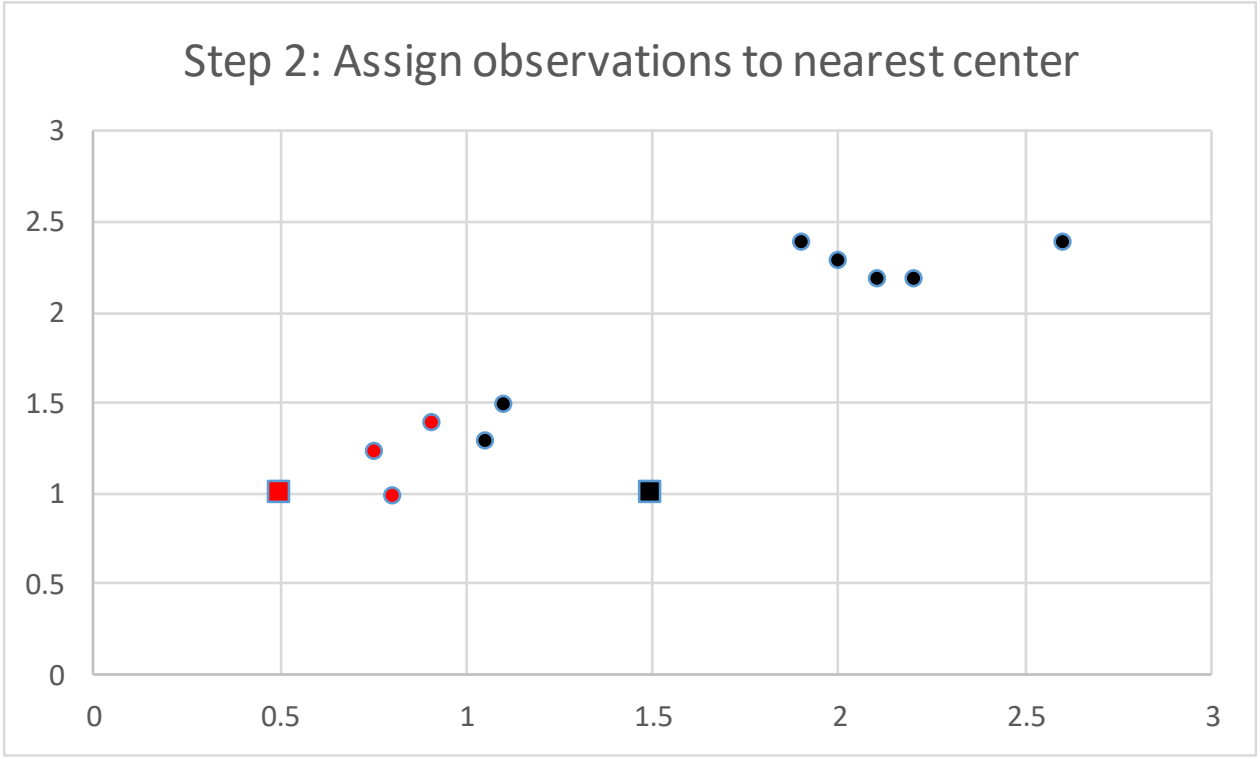
A Simple Example



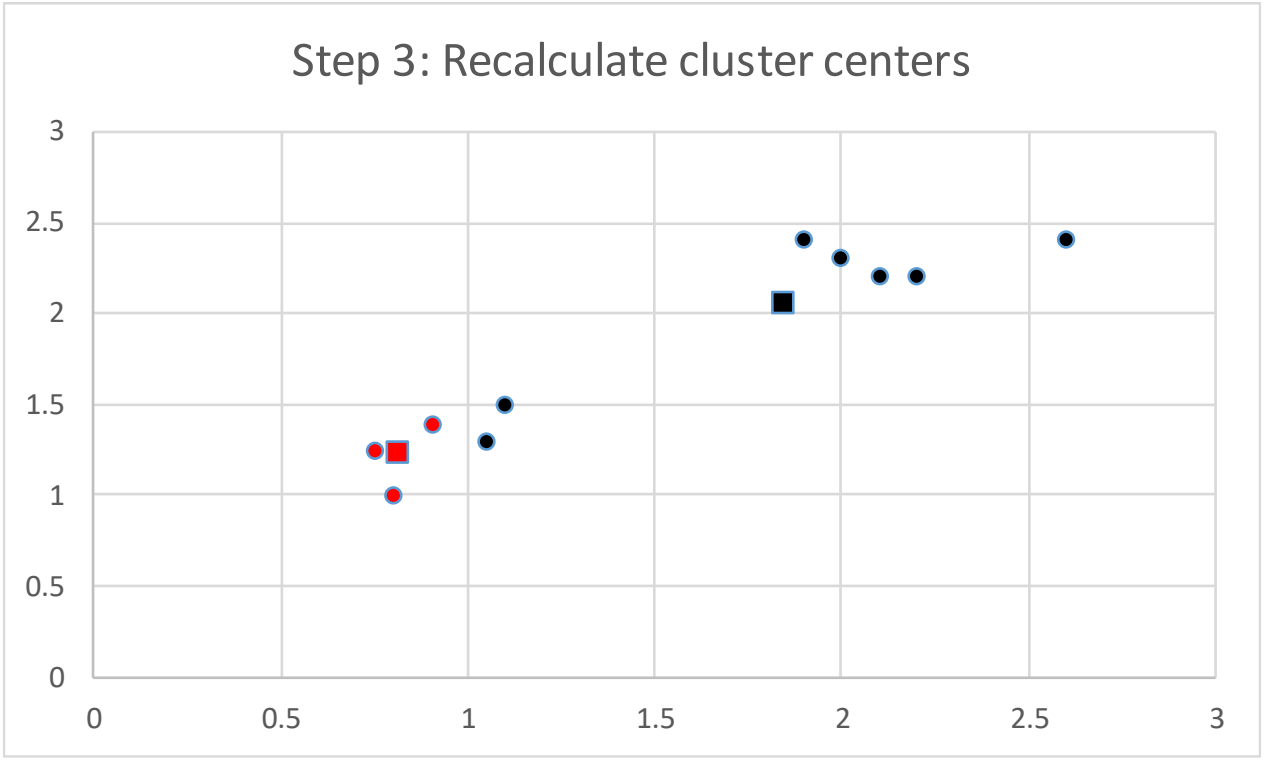
Example continued



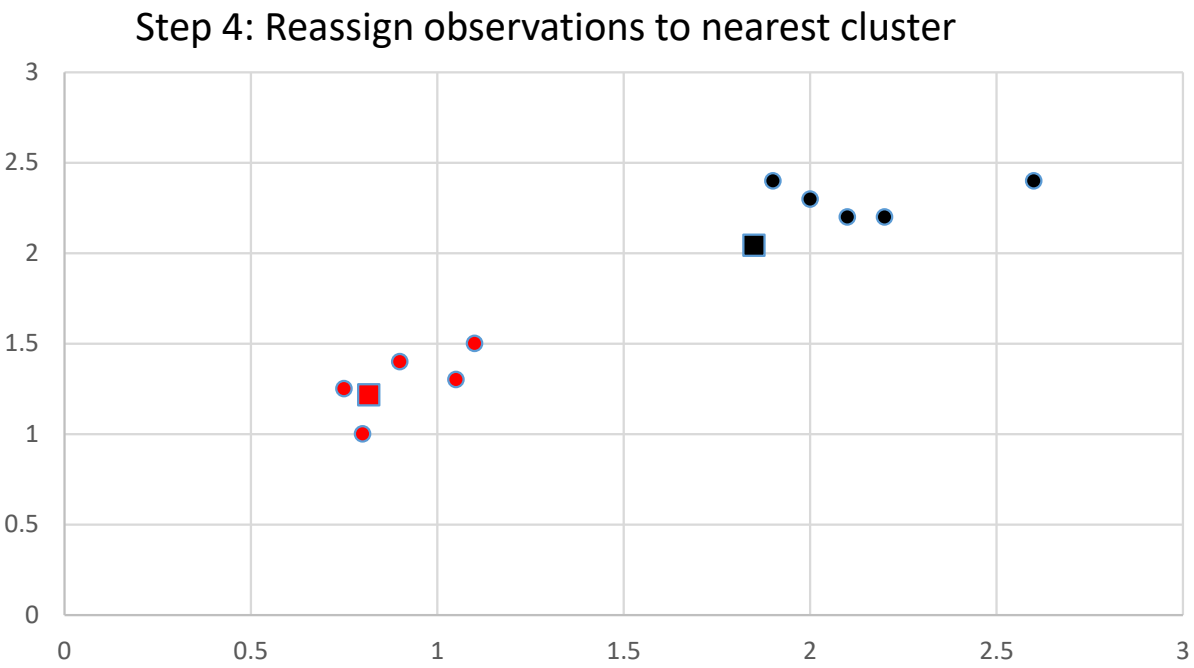
Example continued



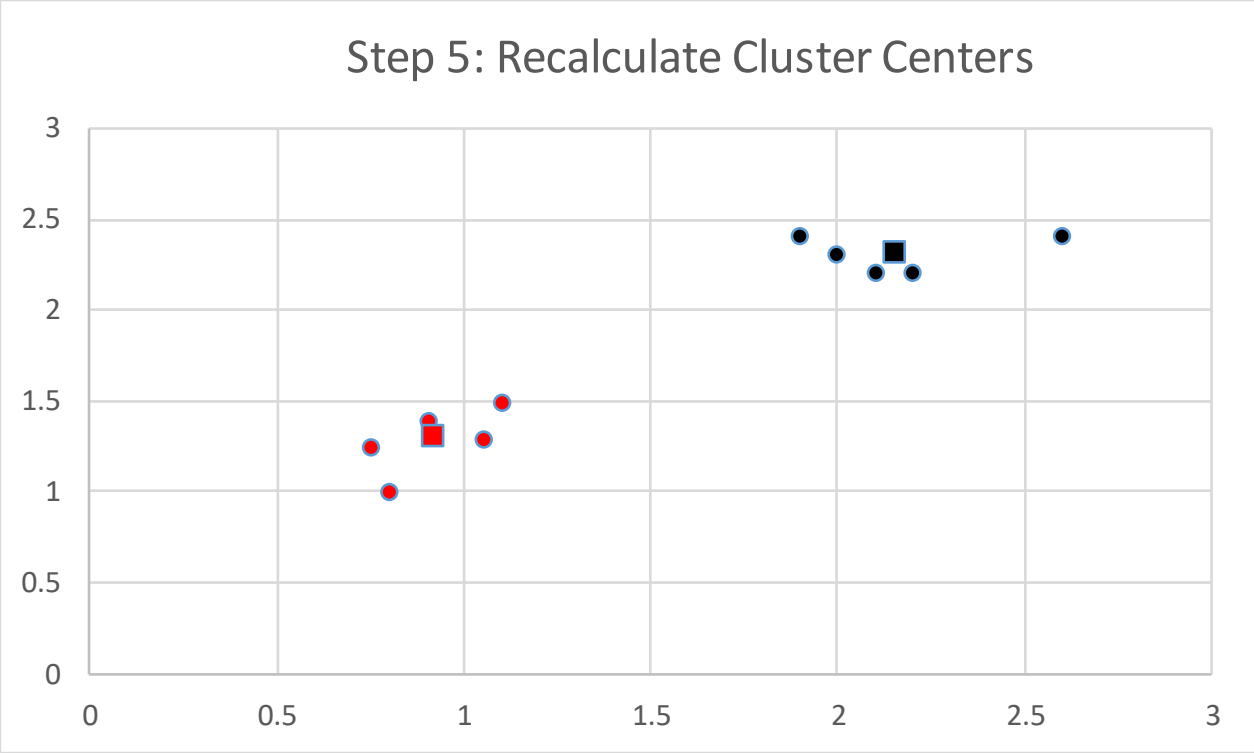
Example continued



Example continued



Example continued



Inertia

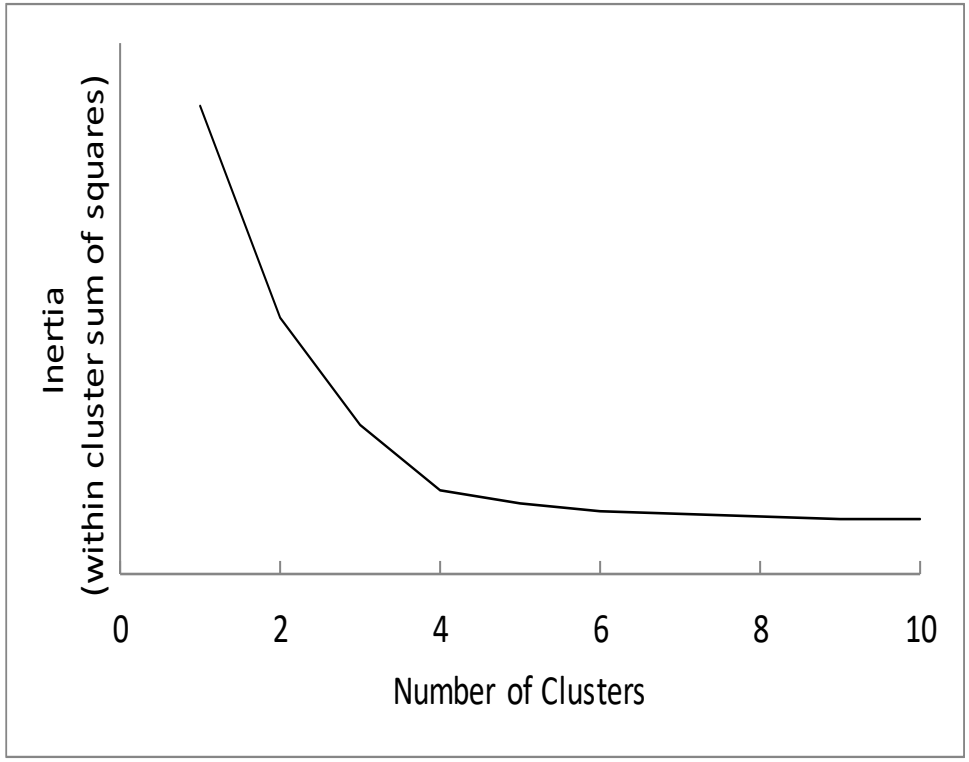
- For any given k the objective is to minimize inertia, which is defined as the within cluster sum of squares:

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

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

Choosing k: the elbow method



Choosing k

- 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

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

  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 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

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

Country Risk Case

Objective is to cluster countries according to their riskiness for foreign investment using 2019 data

Measures of Country Risk

- GDP real growth rate (IMF)
- Corruption index (Transparency international)
- Peace index (Institute for Economics and Peace)
- Legal Risk Index (Property Rights Association)

Collected data on 121 countries. Used Z-score scaling.

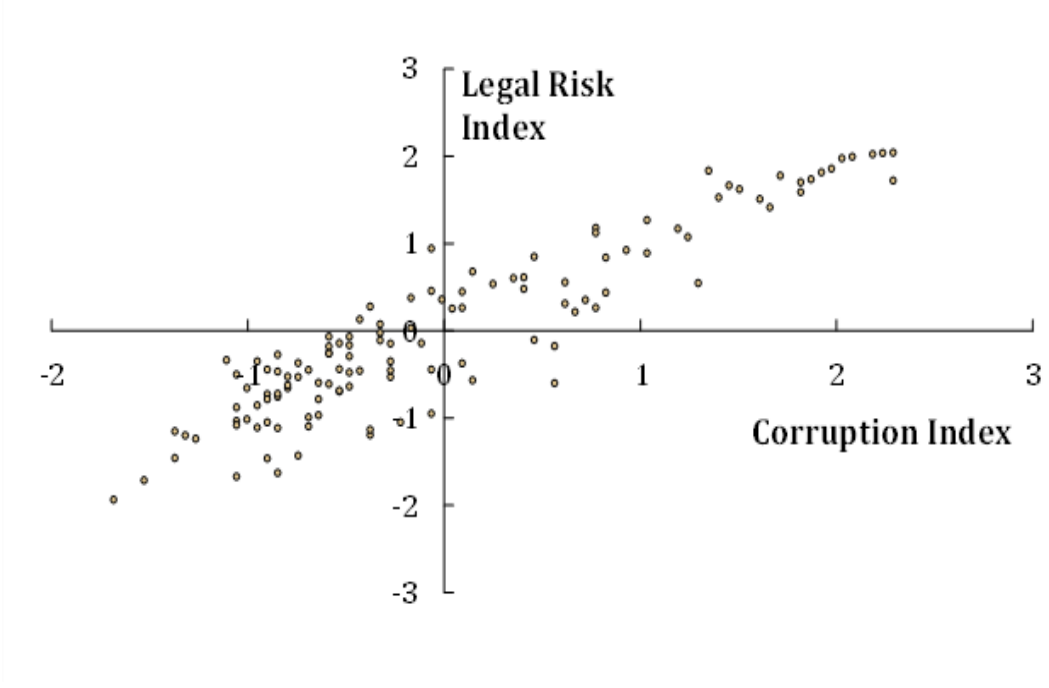
Part of Original Data

Country	Corruption Index	Peace Index	Legal Risk Index	Real GDP growth rate (% per yr)
Albania	35	1.821	4.546	2.983
Algeria	35	2.219	4.435	2.553
Argentina	45	1.989	5.087	−3.061
Armenia	42	2.294	4.812	6.000
Australia	77	1.419	8.363	1.713
Austria	77	1.291	8.089	1.605

Data after Z-score Scaling

Country	Corruption Index	Peace Index	Legal Risk Index	Real GDP growth rate (% per yr)
Albania	−0.633	−0.390	−0.878	0.127
Algeria	−0.633	0.472	−0.959	−0.041
Argentina	−0.099	−0.026	−0.484	−2.231
Armenia	−0.259	0.635	−0.685	1.304
Australia	1.612	−1.261	1.900	−0.368
Austria	1.612	−1.539	1.701	−0.411

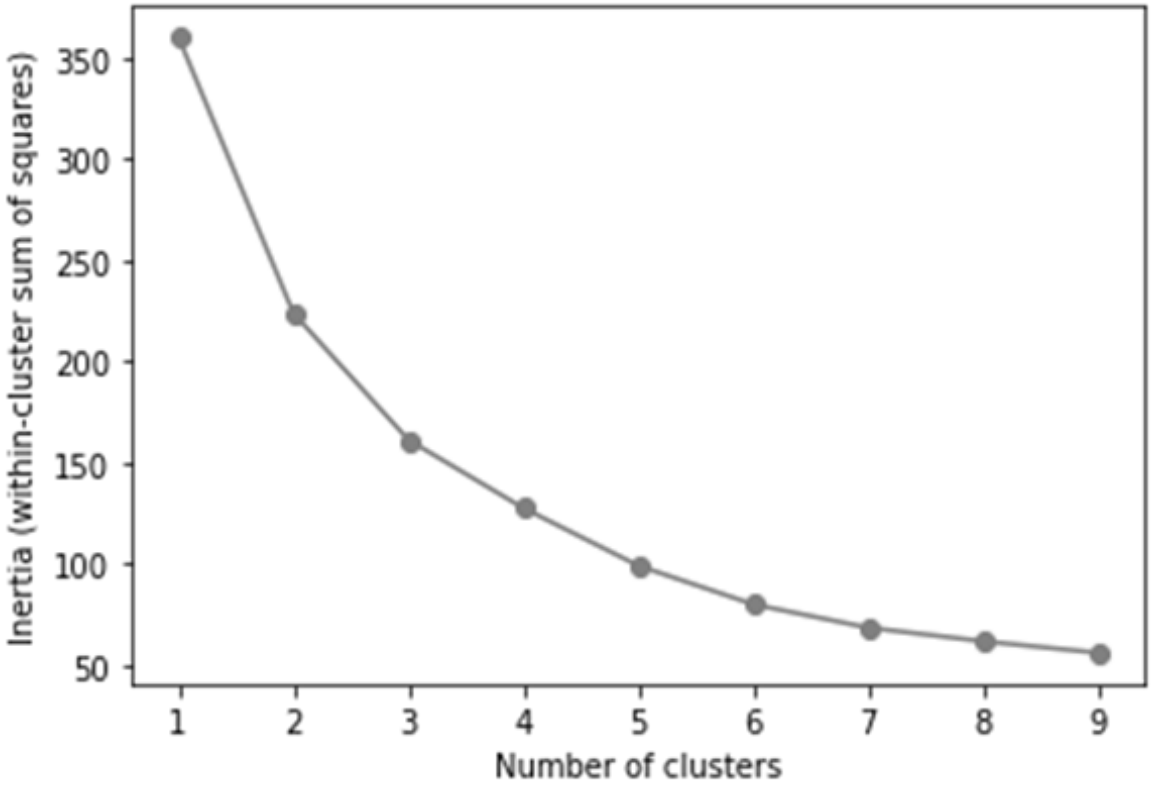
Scaled corruption and legal risk were highly correlated



Therefore analysis based on

- GDP growth rate
- Peace index
- Legal risk index

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



Silhouette scores (suggests k=3)

Number of clusters	Average silhouette score
2	0.351
3	0.360
4	0.340
5	0.344
6	0.348
7	0.355
8	0.355
9	0.332

Cluster centers (scaled values)

	Peace index	Legal index	GDP
High risk	1.35	−0.83	−1.11
Moderate risk	0.22	−0.55	0.60
Low risk	−0.85	1.02	−0.24

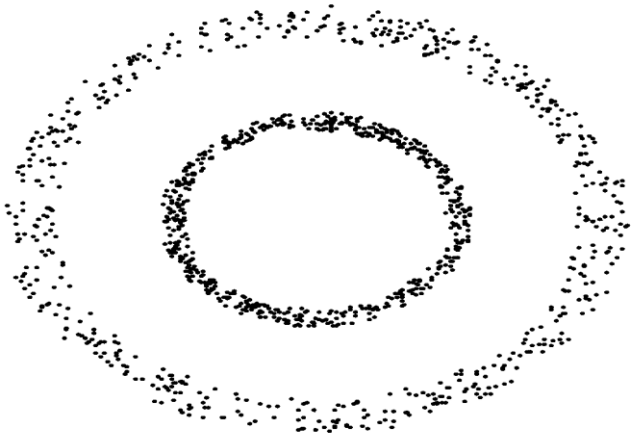
Hierarchical Clustering

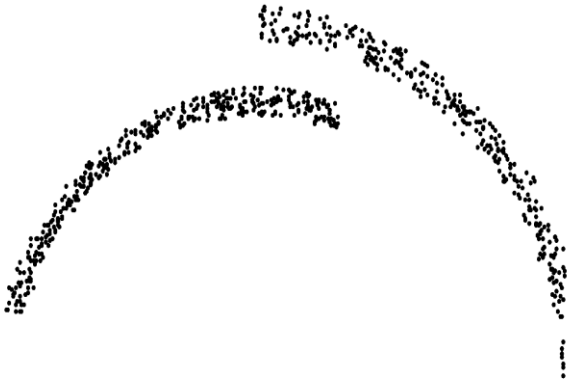
- Start with each observation in its own cluster
- Combine the two closest clusters
- Continue until all observations have been combined into a single cluster
- Can be implemented in Python with AgglomerativeClustering.
- Measures of closeness of clusters:
  - Average Euclidean distance between points 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





Distribution-based Clustering

- Assumes that observations come from a mixture of distributions and uses statistical procedures to separate the distributions

Principal Components Analysis

- This is another approach to reducing the number of variables
- PCA replaces a set of *n* variables by *n* factors so that:
  - Any observation on the original variables is a linear combination of the *n* factors
  - The *n* factors are uncorrelated
  - The quantity of a particular factor in a particular observation is the factor score
  - The importance of a particular factor is measured by the standard deviation of its factor score across observations
- The idea is to find a few variables that account for a high percentage of the variance in the observations

Example: Daily interest rate changes

Maturity	PC1	PC2	PC3
1yr	0.083	−0.242	0.685
2yr	0.210	−0.465	0.376
3yr	0.286	−0.467	0.006
5yr	0.386	−0.315	−0.332
7yr	0.430	−0.099	−0.349
10yr	0.428	0.119	−0.153
20yr	0.426	0.394	0.172
30yr	0.411	0.478	0.323

Interest rate changes

- SD of factor scores

PC1	PC2	PC3
11.54	3.55	1.78

- The fraction of the variance accounted for by first factor is

$$= \frac{11.54^2}{11.54^2 + 3.55^2 + 1.78^2 + \cdots}$$

  or about 87.3%.

- The first two factors account for about 95.6% of the variance

Application to Country Risk case when all 4 features are used

	PC1	PC2	PC3	PC4
Corruption index	0.602	−0.015	0.328	0.728
Peace index	−0.524	0.201	0.825	0.065
Legal risk index	0.594	0.022	0.425	−0.683
GDP Growth rate	0.103	−0.979	0.174	−0.013

	PC1	PC2	PC3	PC4
SD of factor scores	1.600	1.001	0.614	0.243
% of variance	64%	25%	9%	2%