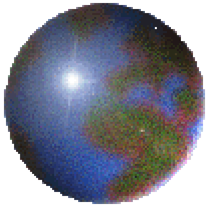


Machine Learning in Business
John C. Hull

Chapter 2
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

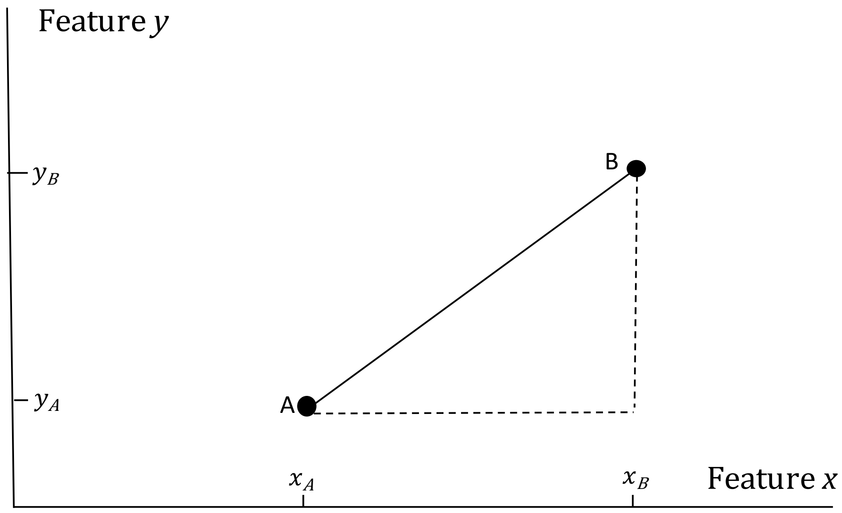
- Suppose you are a bank and have hundreds of thousands of customers and 100 features describing each one
- Unsupervised learning algorithms can be used to divide your customers into clusters so that you can anticipate their needs and communicate with them more effectively

Feature Scaling (page 24-25)

- 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 involves calculating the mean and SD from the values of each feature from the training set. Scaled feature values for all data sets are then created by subtracting the mean and dividing by the SD. The scaled feature values have a mean of zero and SD of one.
- Min-max scaling involves calculating the maximum and minimum value of each feature from the training set. Scaled feature values for all data sets are then created by subtracting the minimum and dividing by the difference between the maximum and minimum. The scaled feature values lie between zero and one.

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

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

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

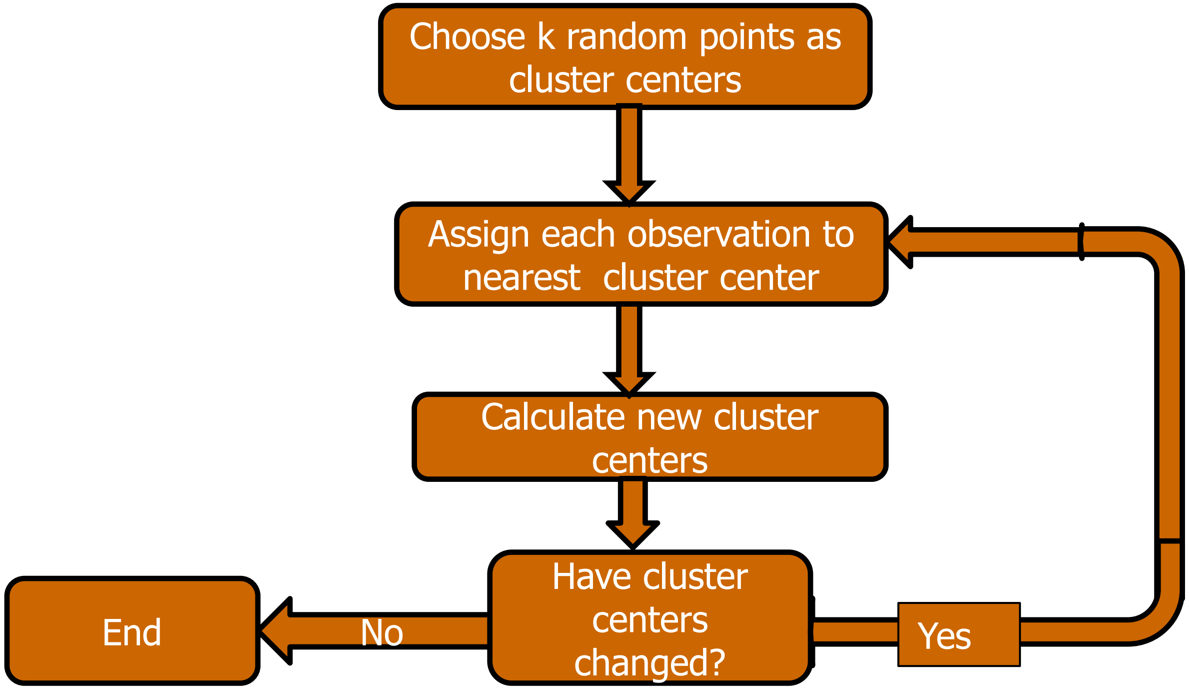
where v_{pj} and v_{qj} and the values of the *j*th feature for P and Q

Cluster Centers (Table 2.1)

- 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	

k-means algorithm to find k clusters (Figure 2.2)


Choose k random points as cluster centers
Assign each observation to nearest cluster center
Calculate new cluster centers
Have cluster centers changed?
Yes
No
End

Inertia

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

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

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 approach (see next slide)
- The silhouette method:

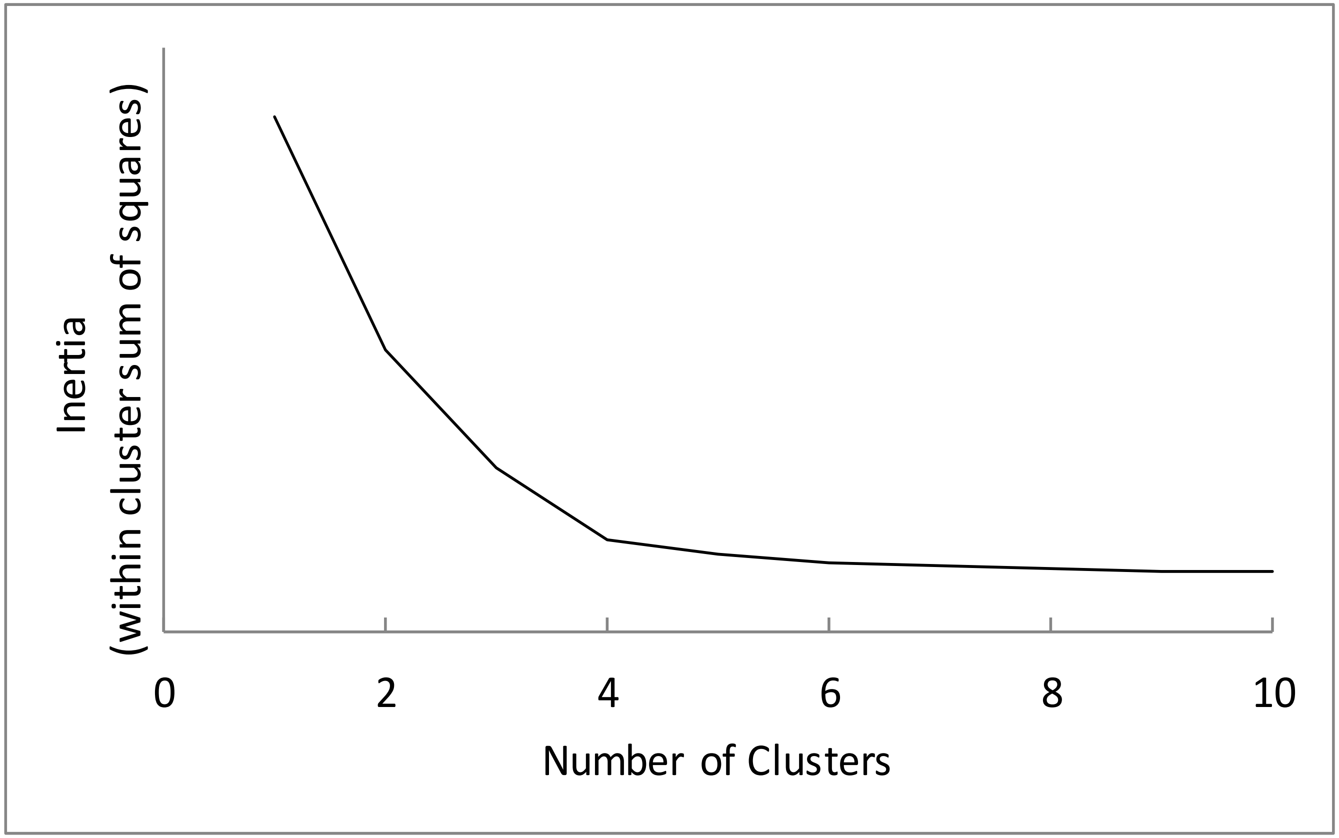
  For each observation i calculate $a(i)$, the average distance from other observations in its cluster, and $b(i)$, the average distance from observations in the closest other cluster. The silhouette score for observation i, $s(i)$, is defined as

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

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



The Curse of Dimensionality (page 31)

- 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

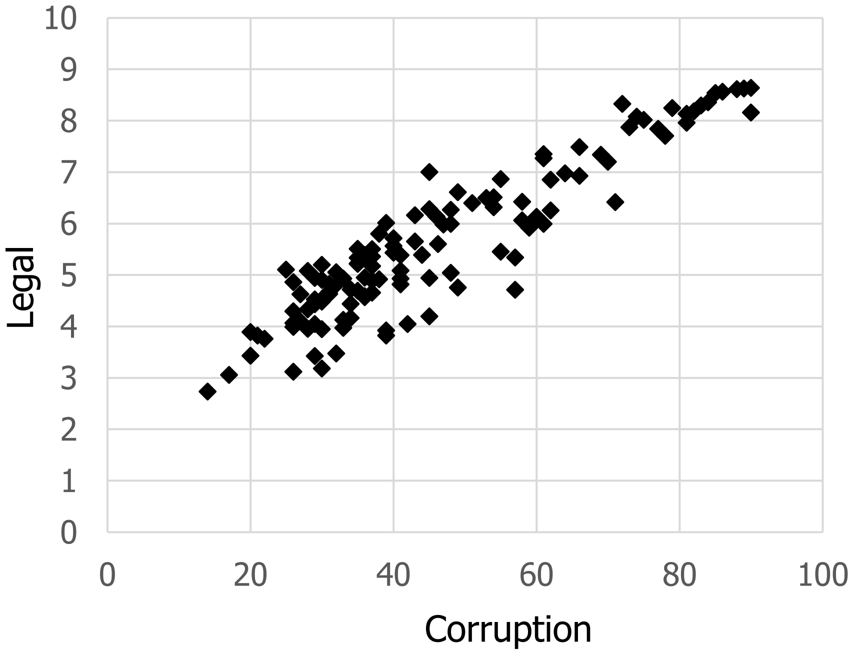
Objective is to cluster countries according to their riskiness for foreign investment

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.

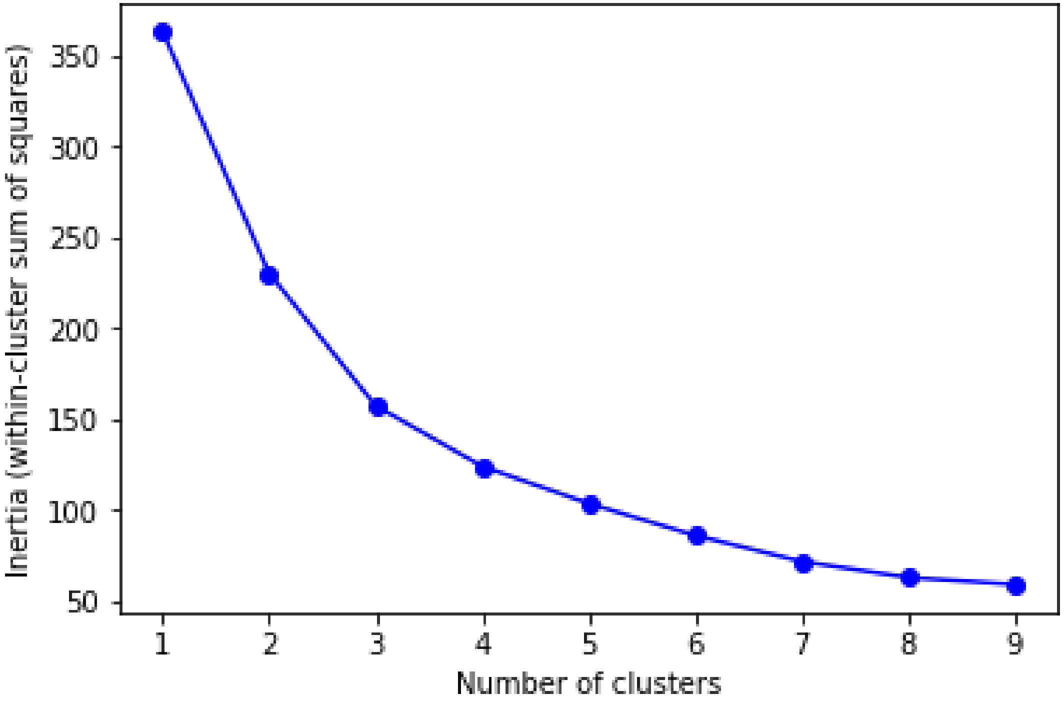
Corruption and legal risk were highly correlated



Therefore analysis based on

- GDP growth rate
- Peace index
- Legal risk index

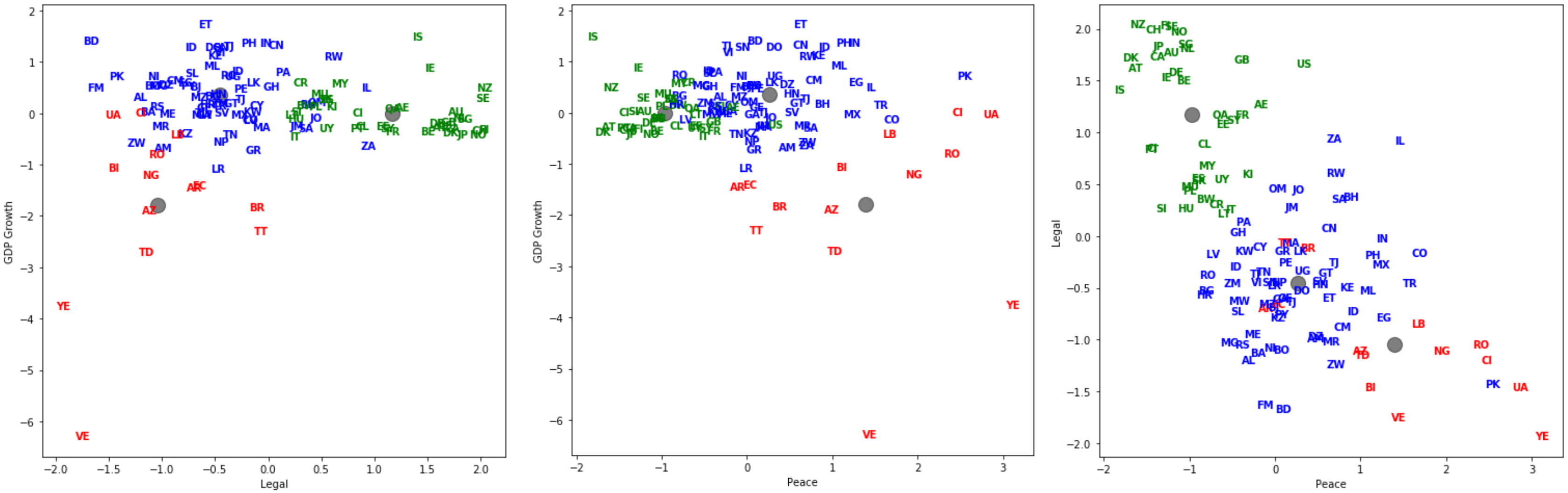
How the total within-cluster sum of squares declines as k increases when k-means algorithm is used (Figure 2.5)



Silhouette scores (suggest k=3)

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

The three-cluster results



Green = Low country risk

Blue = Medium country risk

Red = High country risk

Cluster centers (scaled values) Table 2.5

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

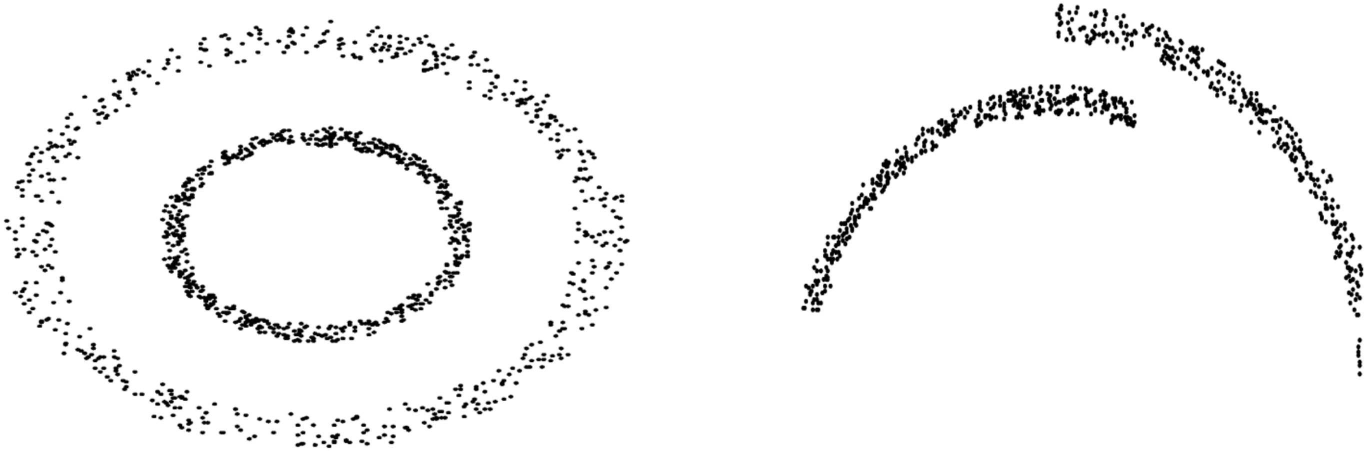
Hierarchical Clustering (page 37)

- 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

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
1yr	0.216	0.501	0.627	0.487	0.122	0.237	-0.011	-0.034
2yr	0.331	0.429	0.129	-0.354	-0.212	-0.674	0.100	0.236
3yr	0.372	0.267	-0.157	-0.414	-0.096	0.311	-0.413	-0.564
4yr	0.392	0.110	-0.256	-0.174	-0.019	0.551	0.416	0.512
5yr	0.404	-0.019	-0.355	0.269	0.595	-0.278	0.316	-0.327
7yr	0.394	-0.194	-0.195	0.336	0.007	-0.100	-0.685	0.422
10yr	0.376	-0.371	0.068	0.305	-0.684	-0.039	0.278	-0.279
30yr	0.305	-0.554	0.575	-0.398	0.331	0.022	-0.007	0.032

Interest rate changes *continued*

- SD of factor scores

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
17.55	4.77	2.08	1.29	0.91	0.73	0.56	0.53

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

$$= \frac{17.55^2}{17.55^2 + 4.77^2 + 2.08^2 + \cdots}$$

or about 90%.

- The first two factors account for over 97% of the variance

Application to Country Risk Case (Tables 2.11 and 2.12)

	PC1	PC2	PC3	PC4
Corruption index	0.594	0.154	−0.292	−0.733
Peace index	-0.530	0.041	−0.842	−0.086
Legal risk index	0.585	0.136	−0.431	0.674
GDP Growth rate	0.152	−0.978	−0.141	−0.026

	PC1	PC2	PC3	PC4
SD of factor scores	1.600	0.988	0.625	0.270
% of variance	64%	24%	10%	2%