Course notes: Cluster Analysis

365 DataScience

Cluster analysis

Cluster analysis

Cluster analysis is a multivariate statistical technique that groups observations on the basis some of their features or variables that they are described by.

The goal of clustering is to maximize the similarity of observations within a cluster and maximize the dissimilarity between clusters.

365 DataScience

Euclidean distance

The most intuitive way to measure the distance between them is by drawing a straight line from one to the other. That's also known as Euclidean distance.

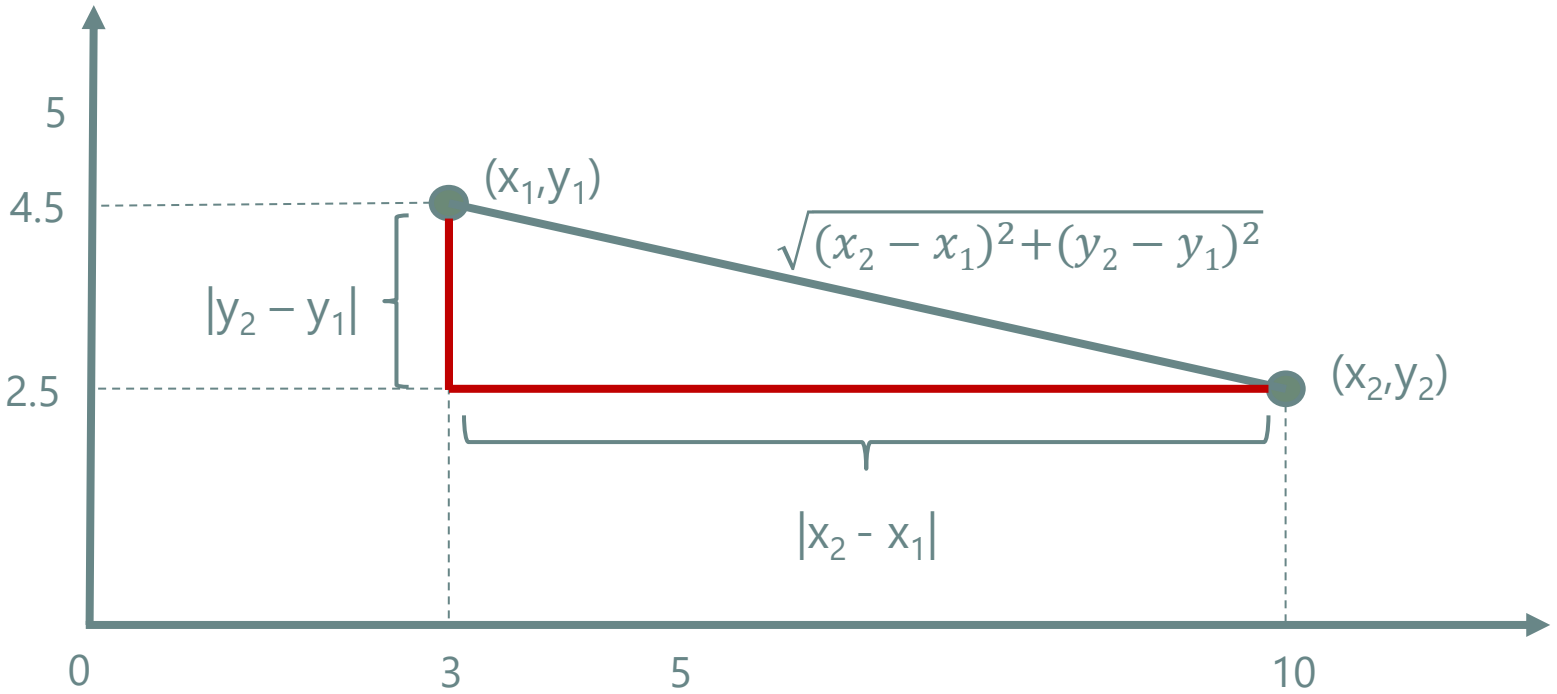
2D space: $d(A,B) = d(B,A) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

3D space: $d(A,B) = d(B,A) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$

If the coordinates of A are $(a_1,a_2,\ldots,a_n)$ and of B are $(b_1,b_2,\ldots b_n)$

N-dim space:

$$d(A,B) = d(B,A) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - bn)^2}$$



365 DataScience

K-means clustering

1. Choose number of clusters

Number – K, chosen by the person performing the clustering

2. Specify the number of seeds

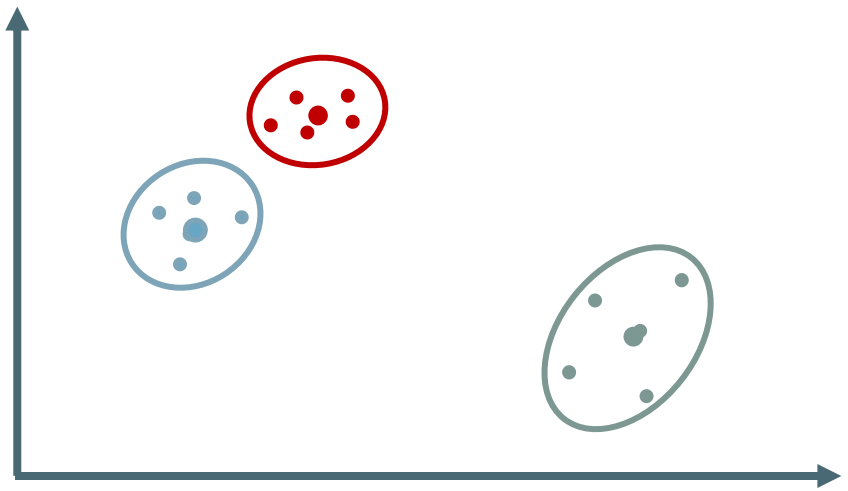
A seed is a starting centroid (can be chosen at random, with an algorithm or according to some prior knowledge)

3. Assign each point to a centroid

Based on proximity (measured by Euclidian distance)

4. Adjust the centroids

Repeat 2. and 3. until there is you can no longer find a better clustering solution



365 DataScience

K-means clustering - pros and cons

PROS

- Simple to implement
  (so many people can use it)
- Computationally efficient
  (it takes considerably less time than any hierarchical clustering model)
- Widely used
  (popular, therefore, in demand)
- Always yields a result
  (also a con as it may be deceiving)

CONS

- We need to pick K
  (often, we don't know how many clusters we need)
- Sensitive to initialization
  (but we can use methods such as kmeans++ to determine the seeds)
- Sensitive to outliers
  (by far the biggest downside of k-means)
- Produces spherical solutions
  (thus, not as generalizable)

365 DataScience

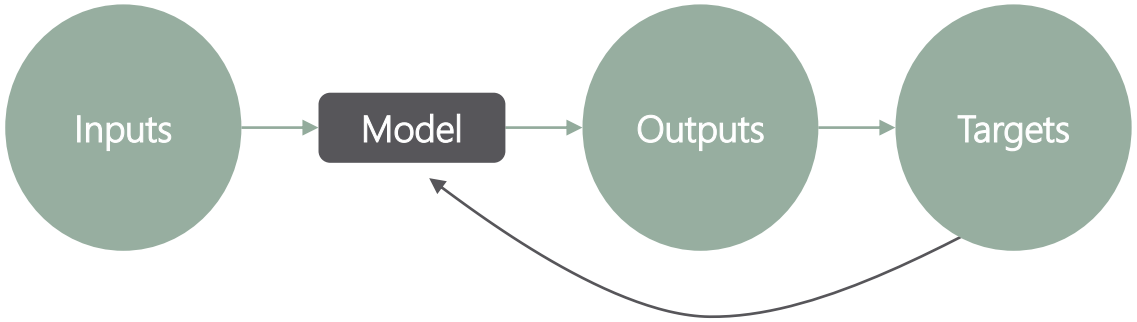
Classification vs Clustering

Classification is a typical example of **supervised learning**.

It is used whenever we have input data and the desired correct outcomes (targets). We train our data to find the patterns in the inputs that lead to the targets.

With classification we essentially need to know the correct class of each of the observations in our data, in order to apply the algorithm.

A logistic regression is a typical example of classification.

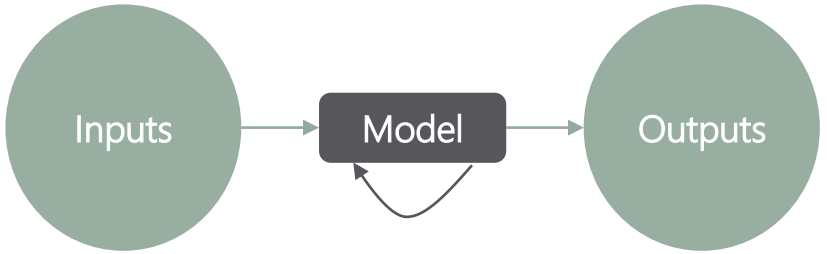


We use the targets (correct values) to adjust the model to get better outputs.

Cluster analysis is a typical example of **unsupervised learning**.

It is used whenever we have input data but have no clue what the correct outcomes are.

Clustering is about grouping data points together based on similarities among them and difference from others.



There is no feedback loop, therefore, the model simply finds the outputs it deems best.

365 DataScience

Types of clustering

With flat methods there is no hierarchy, but rather the number of clusters are chosen prior to clustering.

Flat methods have been developed because hierarchical clustering is much slower and computationally expensive.

Nowadays, flat methods are preferred because of the volume of data we typically try to cluster.

Historically, lustering was developed first. An example hierarchical cof clustering with hierarchy is taxonomy of the animal kingdom.

It is superior to flat clustering in the fact that it explores (contains) all solutions.

With divisive clustering we start from a situation where all observations are in the same cluster, e.g. from the dinosaurs. Then we split this big cluster into 2 smaller ones. Then we continue with 3, 4, 5, and so on, until each observation is its separate cluster.

To find the best split, **we must explore all possibilities at each step**.

When it comes to agglomerative clustering, the approach is bottom up. We start from different dog and cat breeds, cluster them into dogs and cats respectively, and then we continue pairing up species, until we reach the animal cluster.

To find the combination of observations into a cluster, **we must explore all possibilities at each step.**

365 DataScience