UNIVERSITY OF LEEDS

Class: Machine Learning

Unsupervised learning: k-means

Instructor: Matteo Leonetti

Learning outcomes

UNIVERSITY OF LEEDS

- Apply the k-means algorithm to a dataset

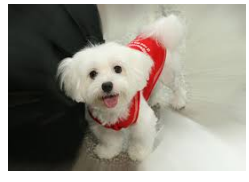
Clustering

UNIVERSITY OF LEEDS









?

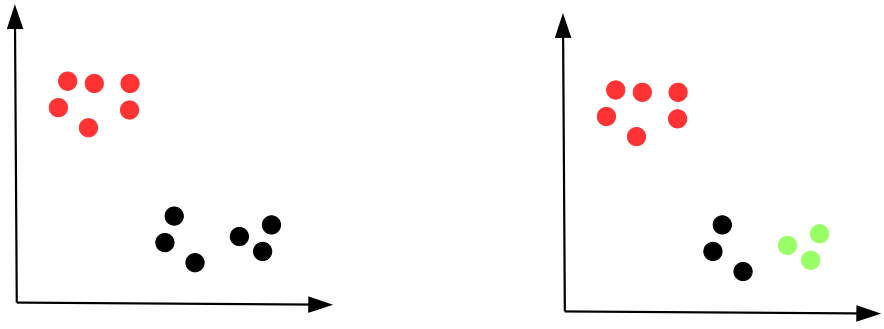
We go now back to the first lecture, when we discussed *clustering*.

Clustering is the problem of grouping elements into *clusters*, because of their proximity in some sense.

This is an *unsupervised* problem, so there are no labels, and therefore no correct or incorrect clustering. It is up to the design to decide whether a certain method and parameters give an acceptable result.

Scale

UNIVERSITY OF LEEDS



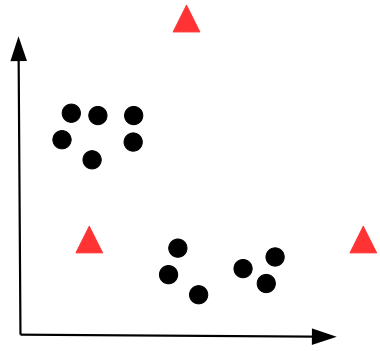
How many clusters do we have?

Depending on the granularity with which you look at a problem, the number of cluster may change.

In this example, considering either two or three clusters seems equally reasonable.

K-means

UNIVERSITY OF LEEDS



1. Choose the number of clusters (in the example: k=3)

2. Place k centroids randomly (the triangles)

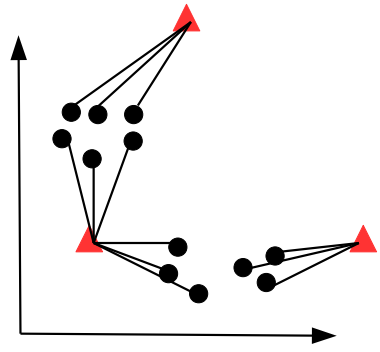
5

We look now at the simplest and most popular clustering algorithm: k-means.

The algorithm begins by choosing the number of clusters we are going to find k, and placing k cluster centres (or centroids) randomly in the feature space.

K-means

UNIVERSITY OF LEEDS

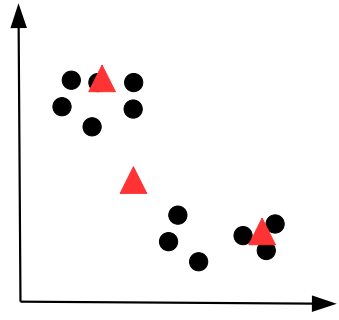


3. Identify the closest centroid to each point

4. Compute the new centroids for the clusters

K-means

UNIVERSITY OF LEEDS

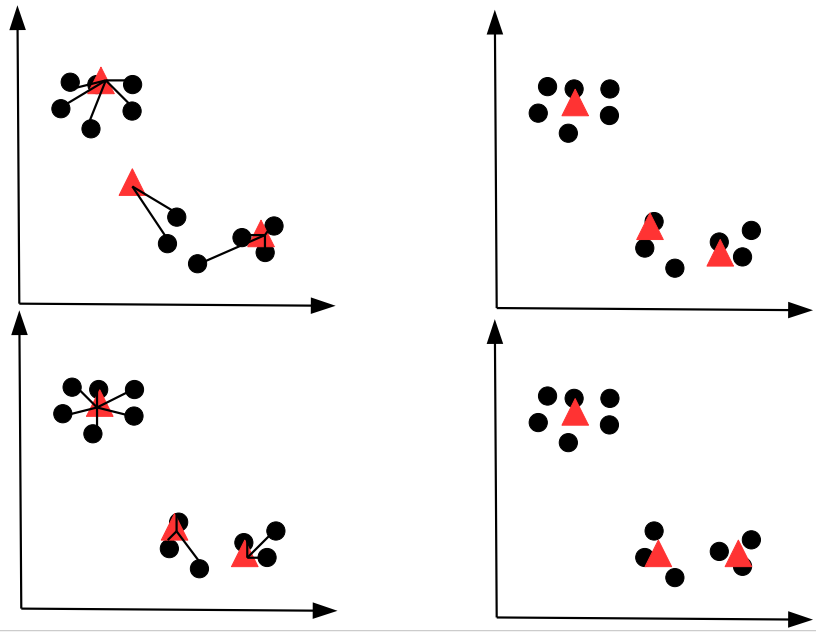


5. Repeat until the centroids do not move

Cluster new points with the closest cluster centre (centroid)

K-means

UNIVERSITY OF LEEDS



K-means - characteristics

UNIVERSITY OF LEEDS

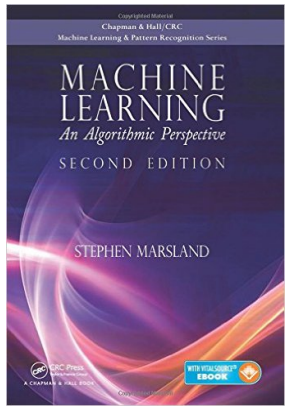
Very easy to implement 

You have to choose the number of clusters 

Subject to local minima (clusters depend on initial positions of the centroids) 

Yet, a popular first thing to try!

UNIVERSITY OF LEEDS



Chapter 14 (intro)
Chapter 14.1