Clustering

Adam Carter, EPCC, The University of Edinburgh
a.carter@epcc.ed.ac.uk

1 September 2020

www.archer2.ac.uk

archer2

epcc

Reusing this material

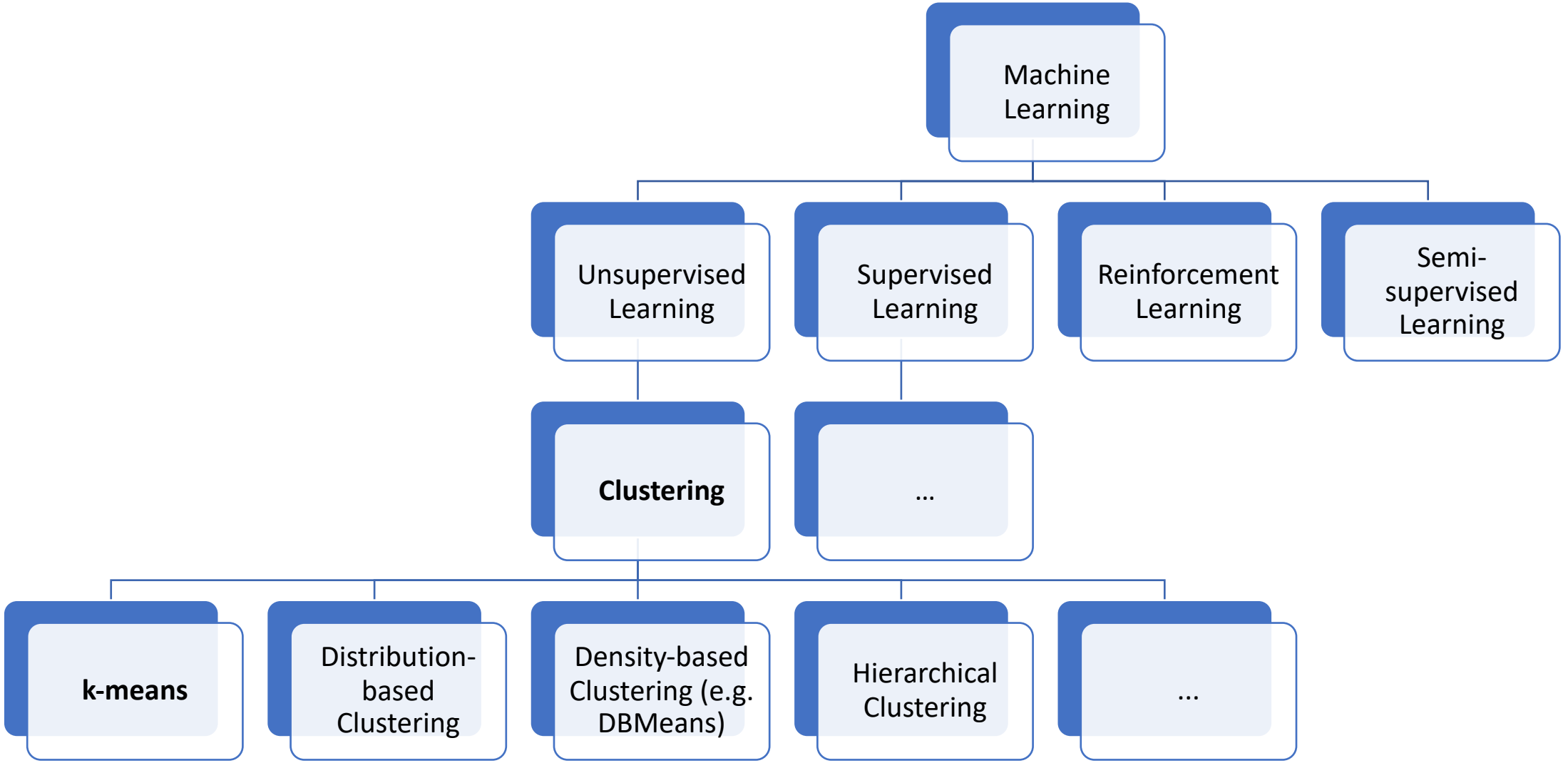
Partners

epcc

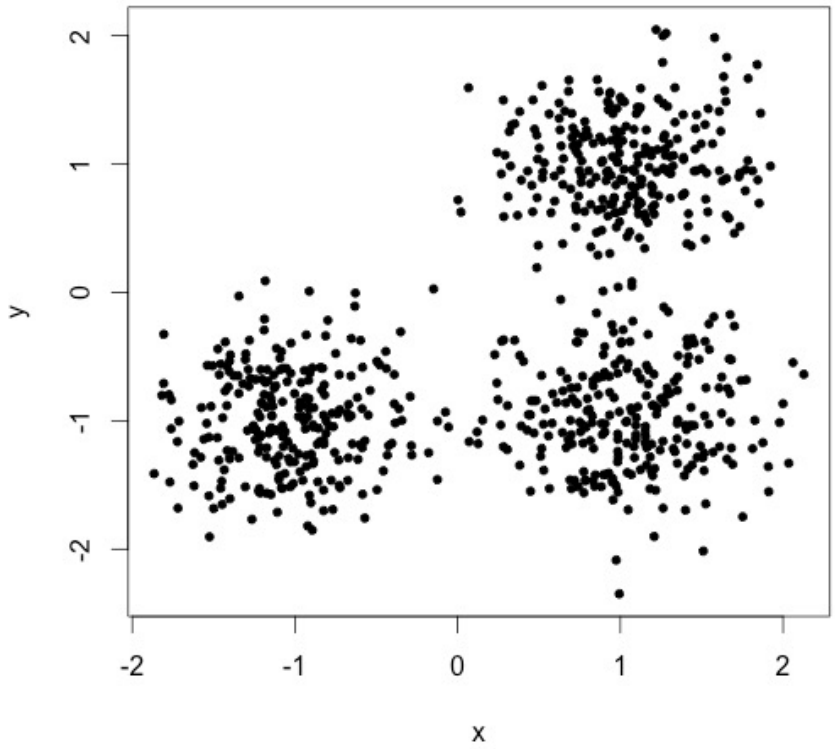
epcc

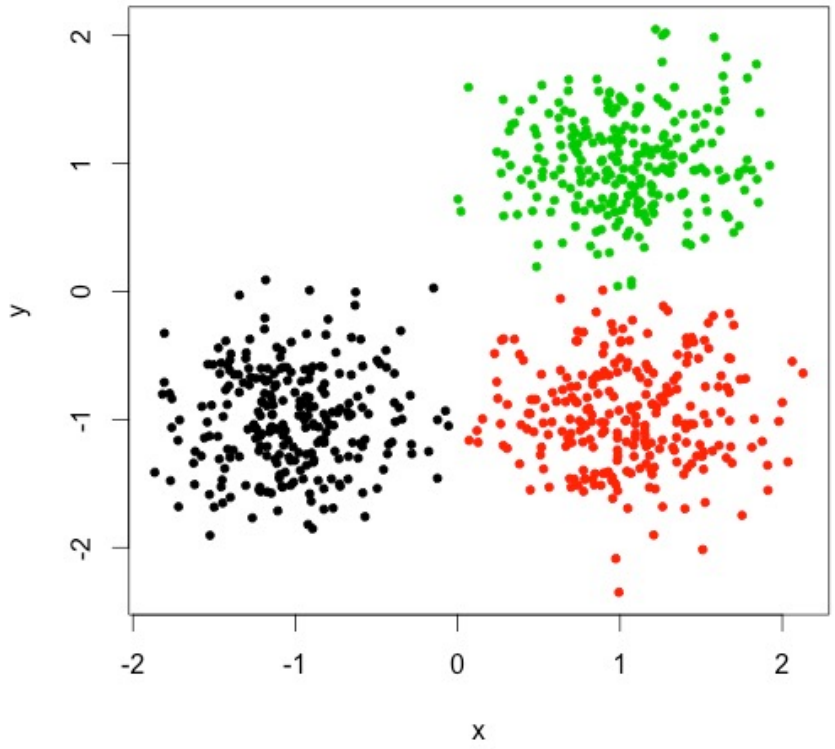
THE UNIVERSITY of EDINBURGH

Hewlett Packard Enterprise


Machine Learning
Unsupervised Learning
Supervised Learning
Reinforcement Learning
Semi-supervised Learning
Clustering
...
k-means
Distribution-based Clustering
Density-based Clustering (e.g. DBMeans)
Hierarchical Clustering
...

Clustering – the problem

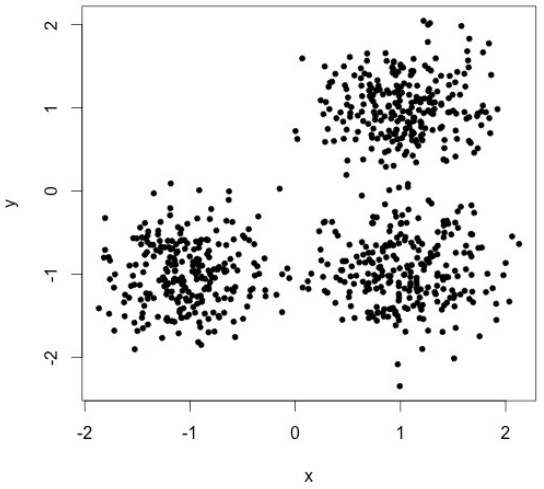




k-Means Clustering

Brief Aside: Notation (1)

- Unless specified, we'll try to stick to the following notation here:
- $\boldsymbol{x}$ A single instance
  - A single point (as in the graph below). A single member of your population. A single observation. Usually corresponds to a single *row* in a dataset. Usually, the instance has many features or variables. The values of each feature are the values of this vector. So, if there are n features, then $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ or $\boldsymbol{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})$



- $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ A set of m observations
  - Usually corresponds to a dataset with *m* rows and *n* columns

Brief Aside: Notation (2)

- $\boldsymbol{x}$ A single instance
- $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ A set of m observations

- $\boldsymbol{x}_a, \boldsymbol{x}_b$ Two observations a and b
- $\|\boldsymbol{x}_b - \boldsymbol{x}_a\|$ The "distance" between the two points a and b

  - Usually (*c.f.* "square on the hypotenuse"):

$$\|\boldsymbol{x}_b - \boldsymbol{x}_a\| = \sqrt[2]{(x_b^{(1)} - x_a^{(1)})^2 + (x_b^{(2)} - x_a^{(2)})^2 + \cdots + (x_b^{(n)} - x_a^{(n)})^2}$$

  - Note that $\|\boldsymbol{x}_b - \boldsymbol{x}_a\| = \|\boldsymbol{x}_a - \boldsymbol{x}_b\|$

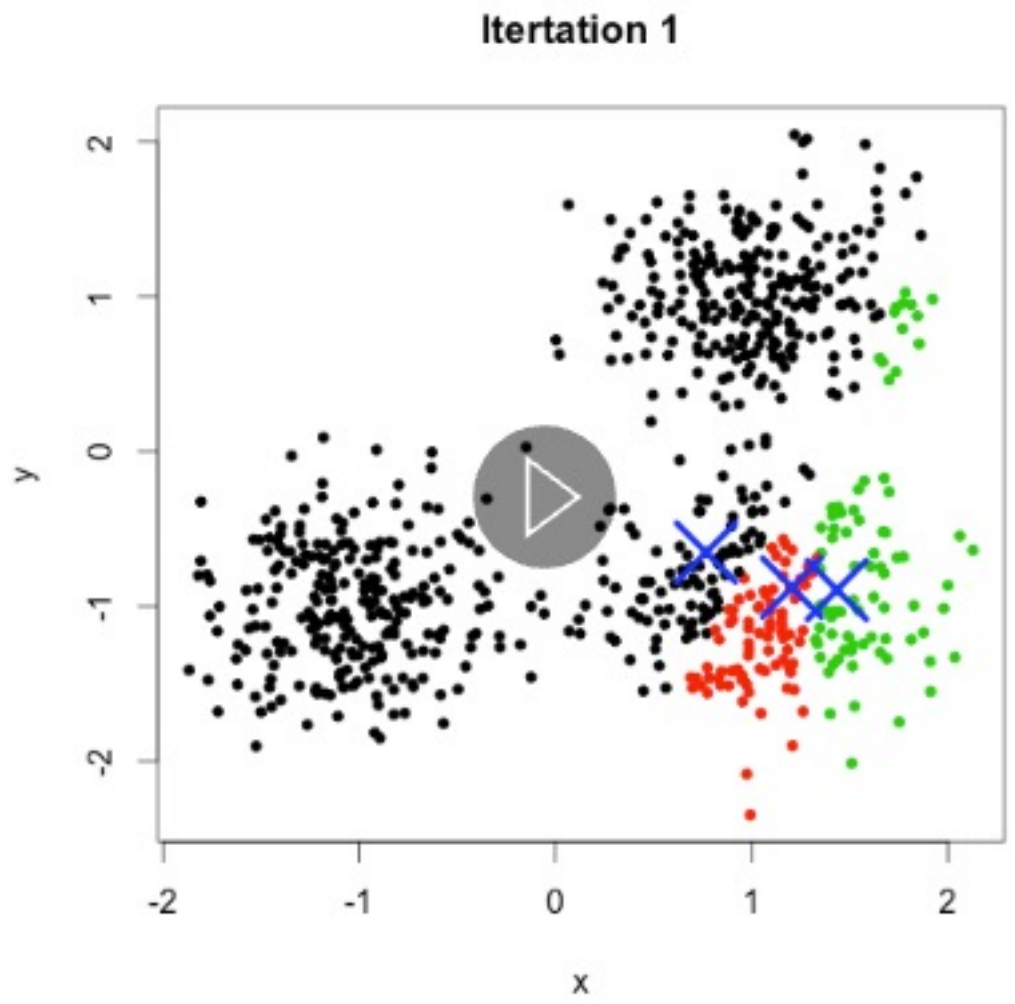
k-means clustering

- Know in advance that there are k clusters
- Goal:
  - Given observation vectors: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$
  - Group them in k distinct sets: $S_1, S_2, \ldots, S_k$
  - Minimise the within-cluster sum of squares:
  - $\sum_{c=1}^{k} \sum_{\boldsymbol{x} \in S_c} \|\boldsymbol{x} - \boldsymbol{\mu}_c\|^2$
  - where $\boldsymbol{\mu}_c$ is the mean of the points in the set S_c
  -

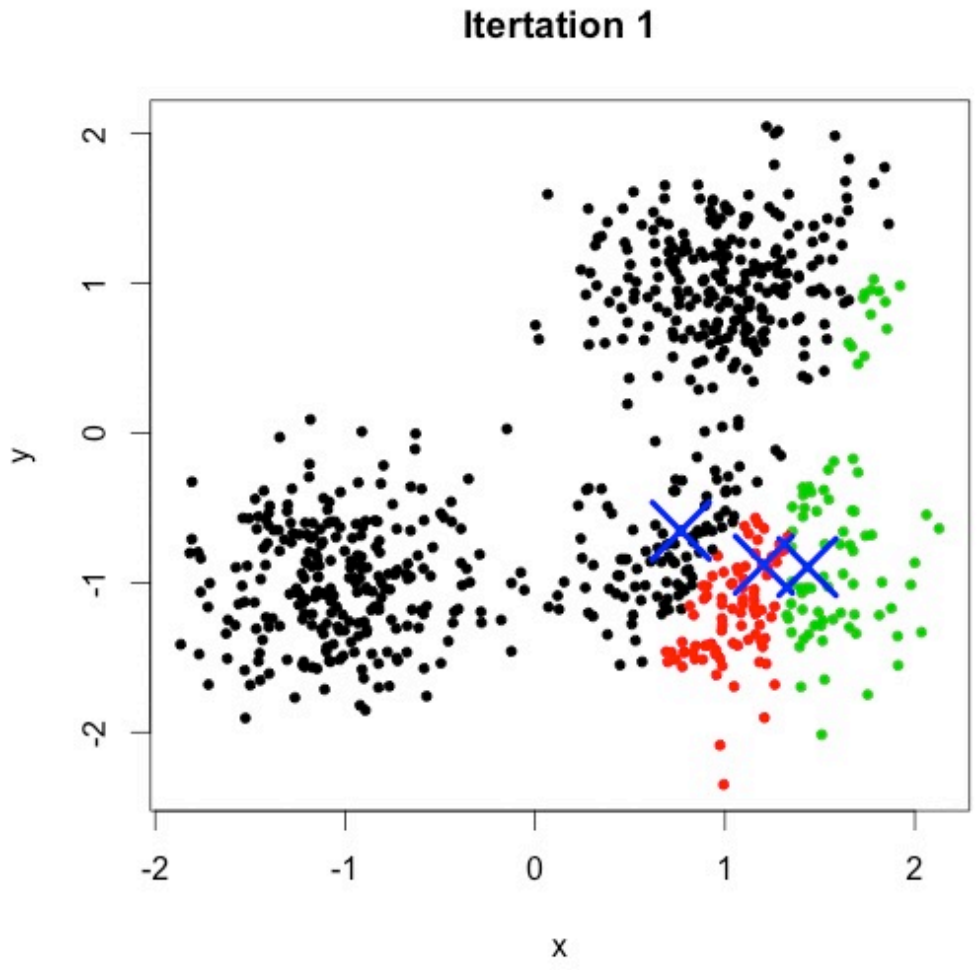
k-means algorithm

- Define each cluster by the centroid (mean) of the values in the cluster
- Algorithm:
  - Choose k random centroids (possibly k observations)
  - Repeat:
    - Assign each observation to nearest centroid
    - Update centroids to be mean for those observations assigned to it.
  - Until there is no, or very little, change in the centroids

k-means in action



k-means in action

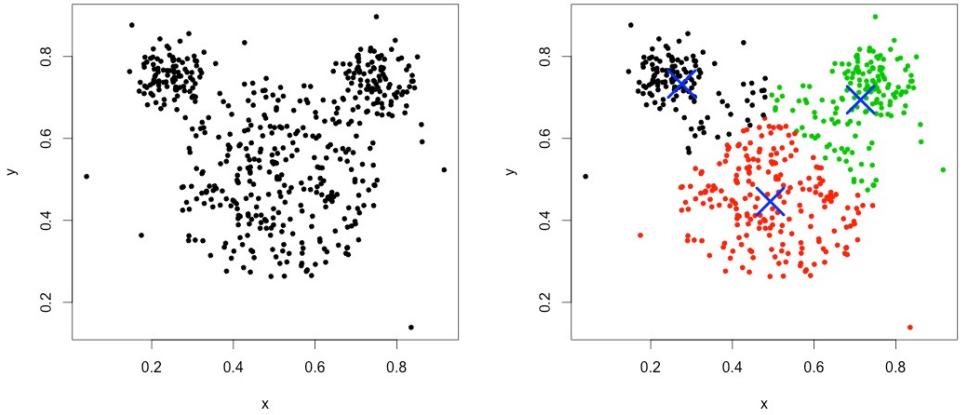


k-means in practice

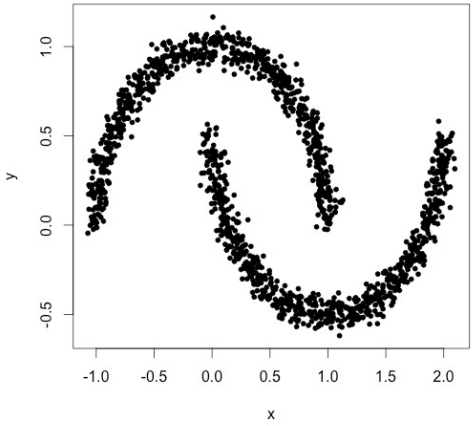
- Must normalise the features so that the distances are not biased to a particular dimension
- Need to think carefully about the features you wish to include as algorithm will give each feature equal weight
  - Unlike certain supervised learning approaches like linear regression where the weight may be zero, or Naïve Bayes where a meaningless feature will have no real impact
- Can be hard to interpret
  - Sometimes the clusters seem meaningless
- Need to choose k
  - Sometimes you know there are k processes generating the data
  - Trial and error
  - Look for 'knee' in plot of cost against k

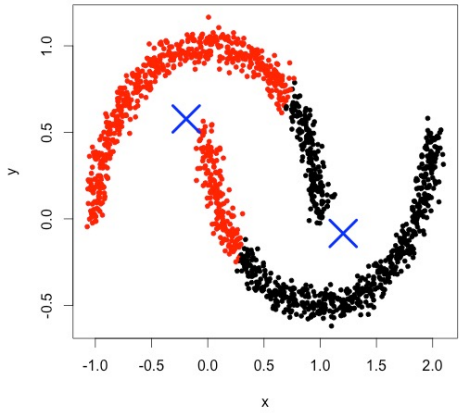
Limitations of k-means

- Clusters assumed to the the same size



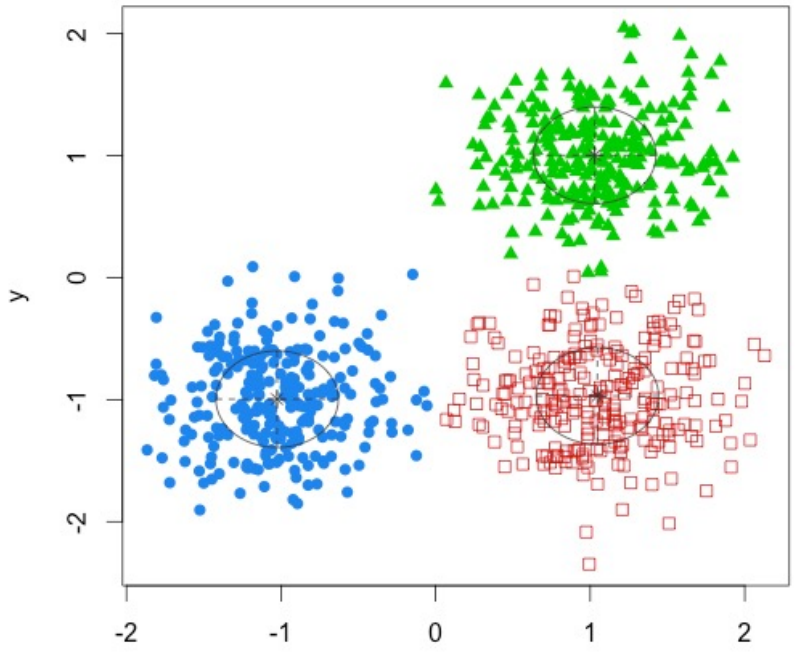
- Clusters on density not so good

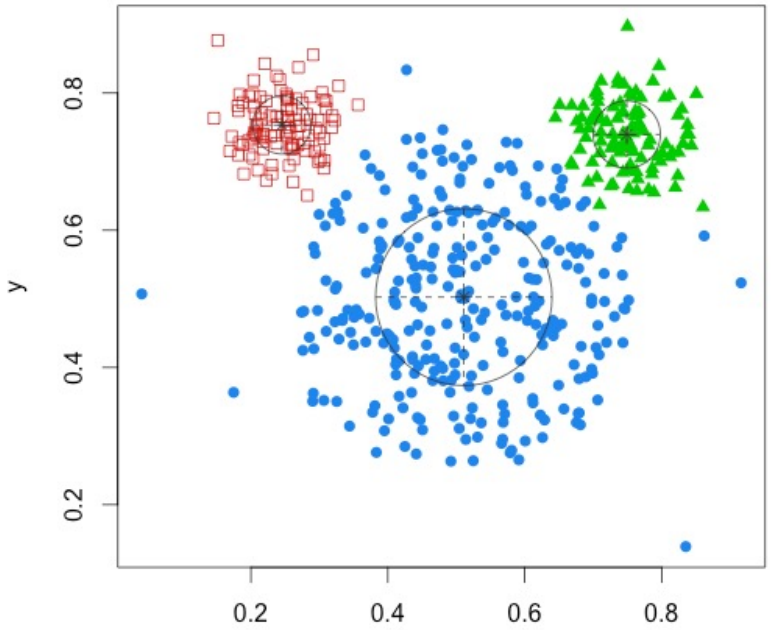


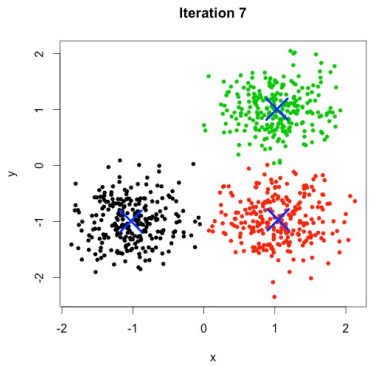


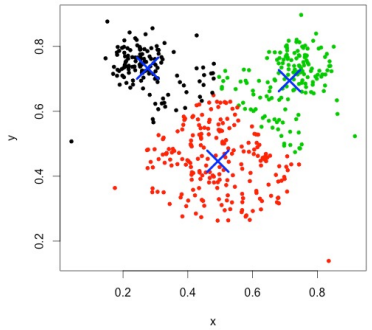
Some Other Clustering Techniques

Distribution-based clustering in practice



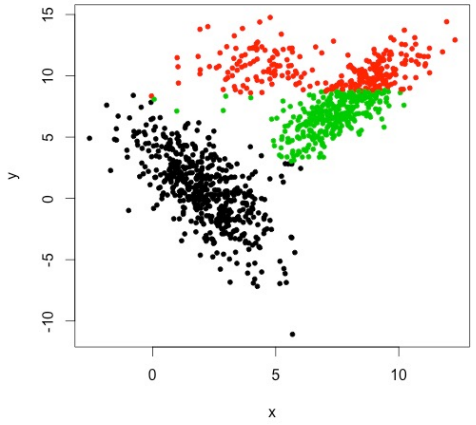


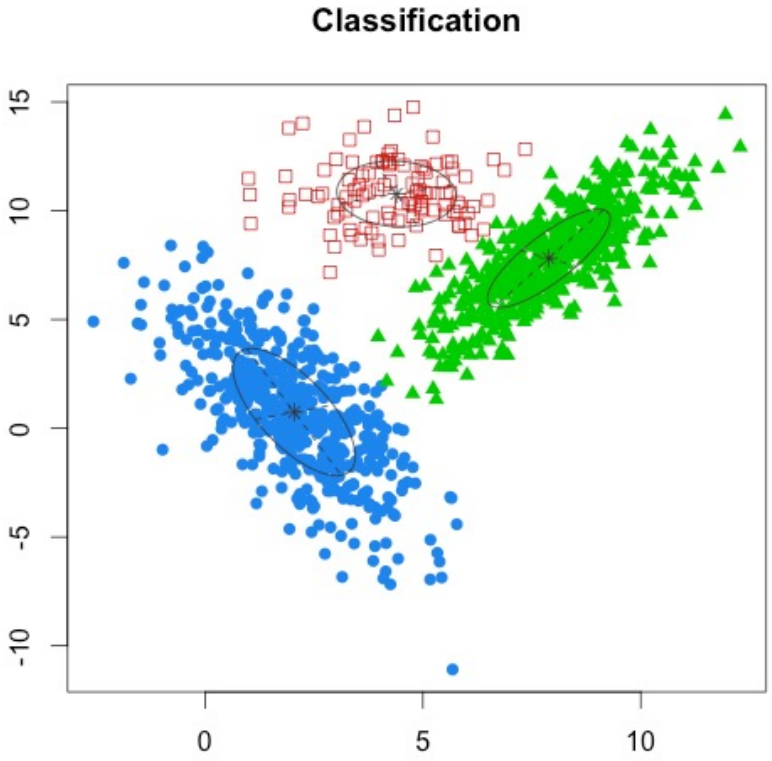




Distribution-based clustering

- Handles covariance of features
  - No need to normalise data





Limitations of distribution-based clustering

- Bad for density-based clusters that don't match distribution model