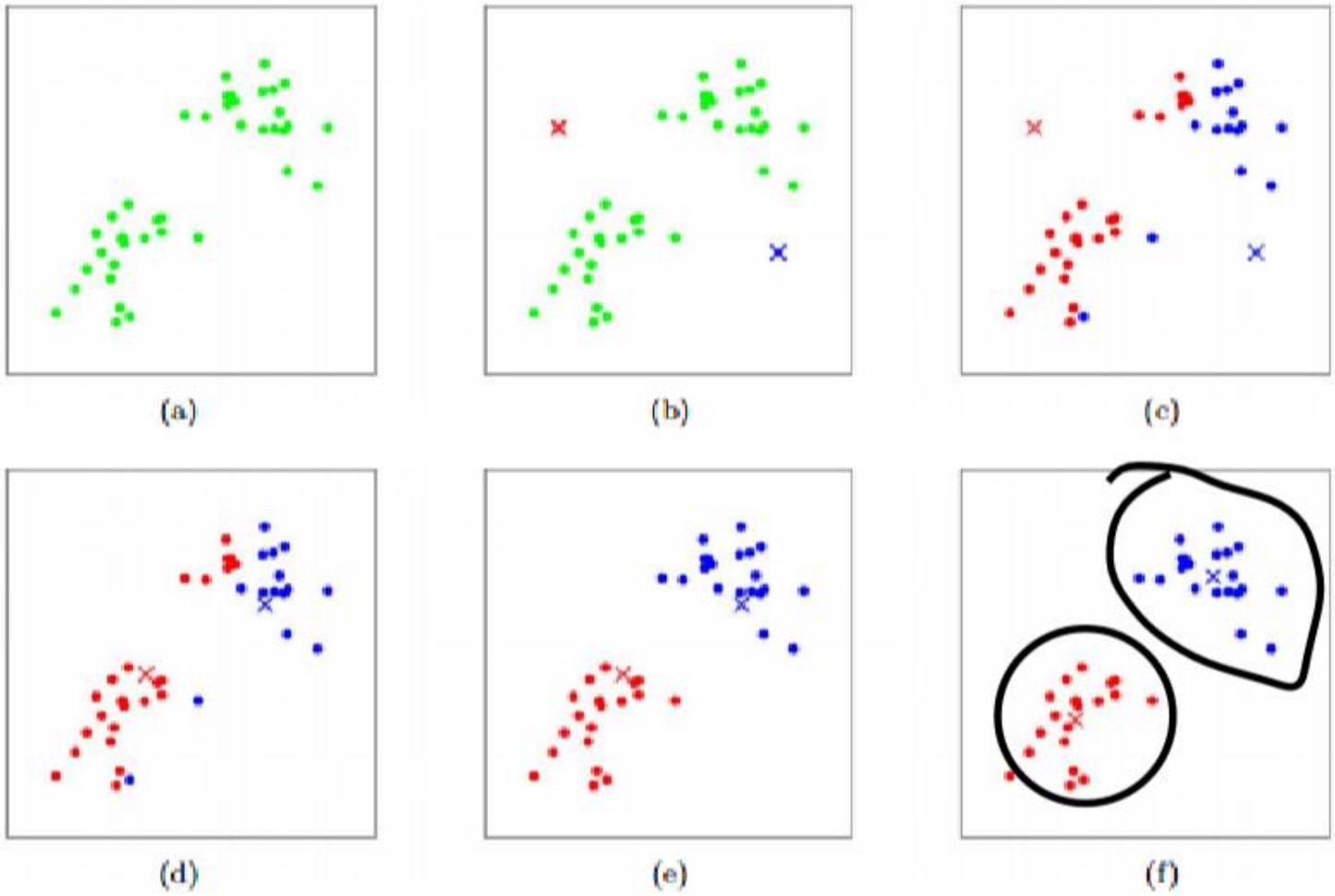
K-Means

BY MG ANALYTICS

K- Means

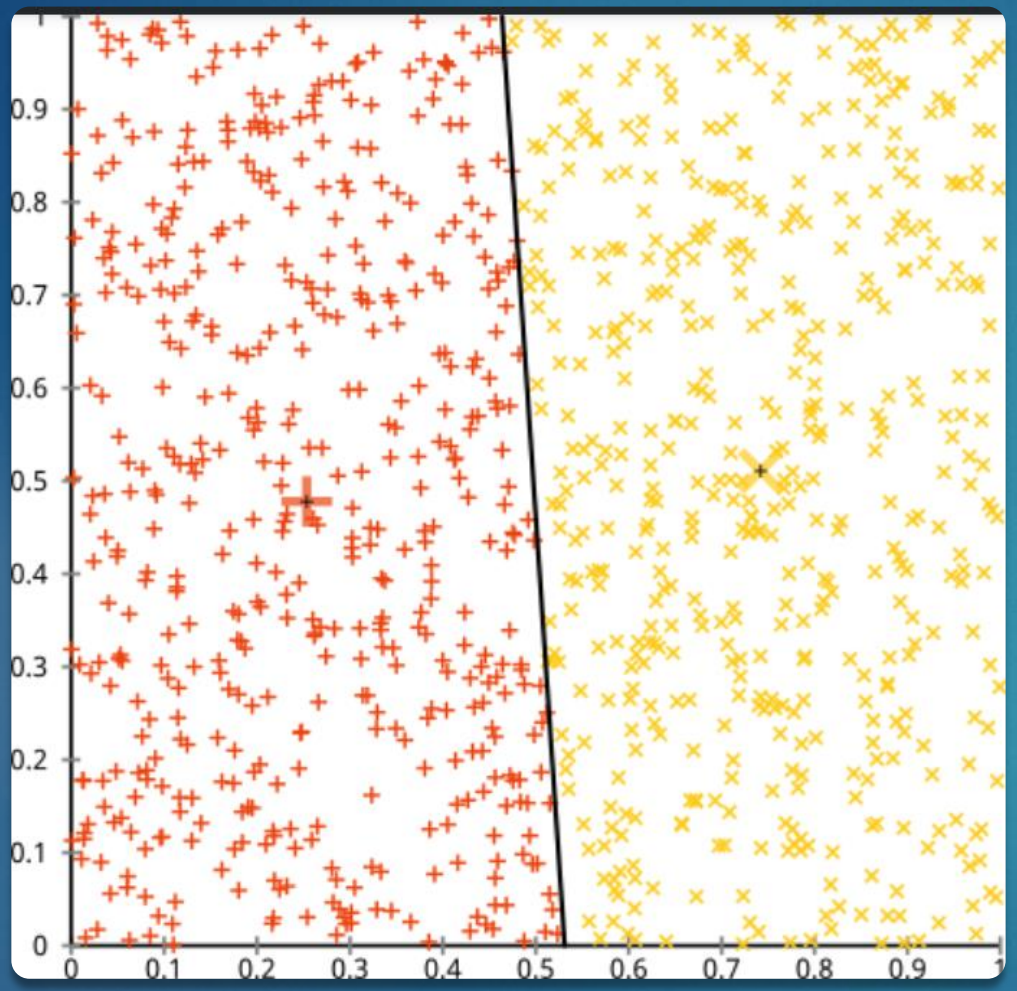
- K i.e. the numbers of clusters to be created.
- The mechanism randomly initializes K random centroids in feature space.
- At time of initialization the K points are not at actual centroids of data.
- The Data points are assigned to the nearest K points.
- The centroids are moved so that the are at the center of the current designated clusters.

MG ANALYTICS

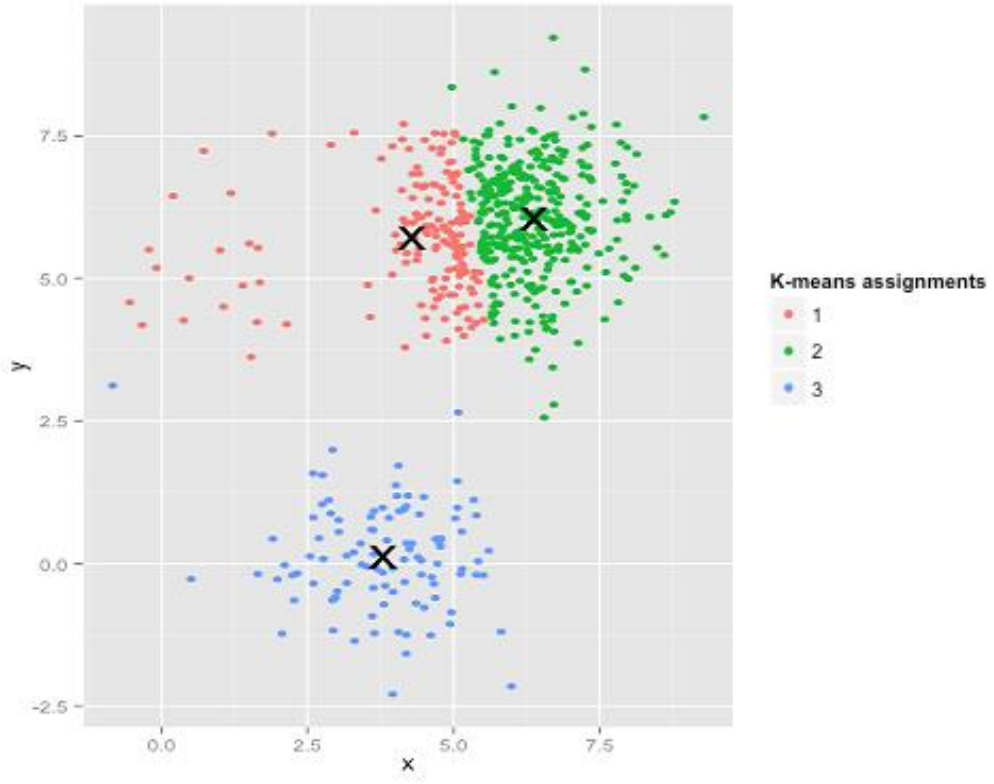

(a)
(b)
(c)
(d)
(e)
(f)

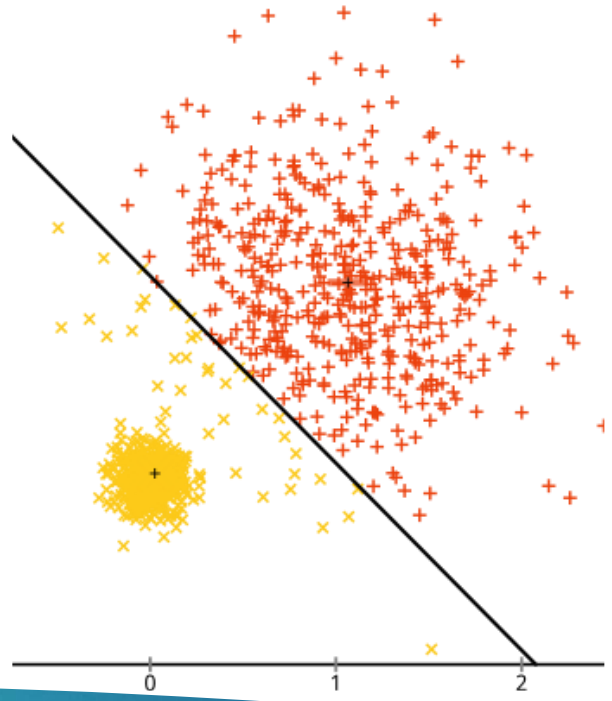
Assumptions

- k-means assumes the variance of the distribution of each attribute (variable) is spherical;
- all variables have the same variance;
- the prior probability for all k clusters is the same, i.e., each cluster has roughly equal number of observations;

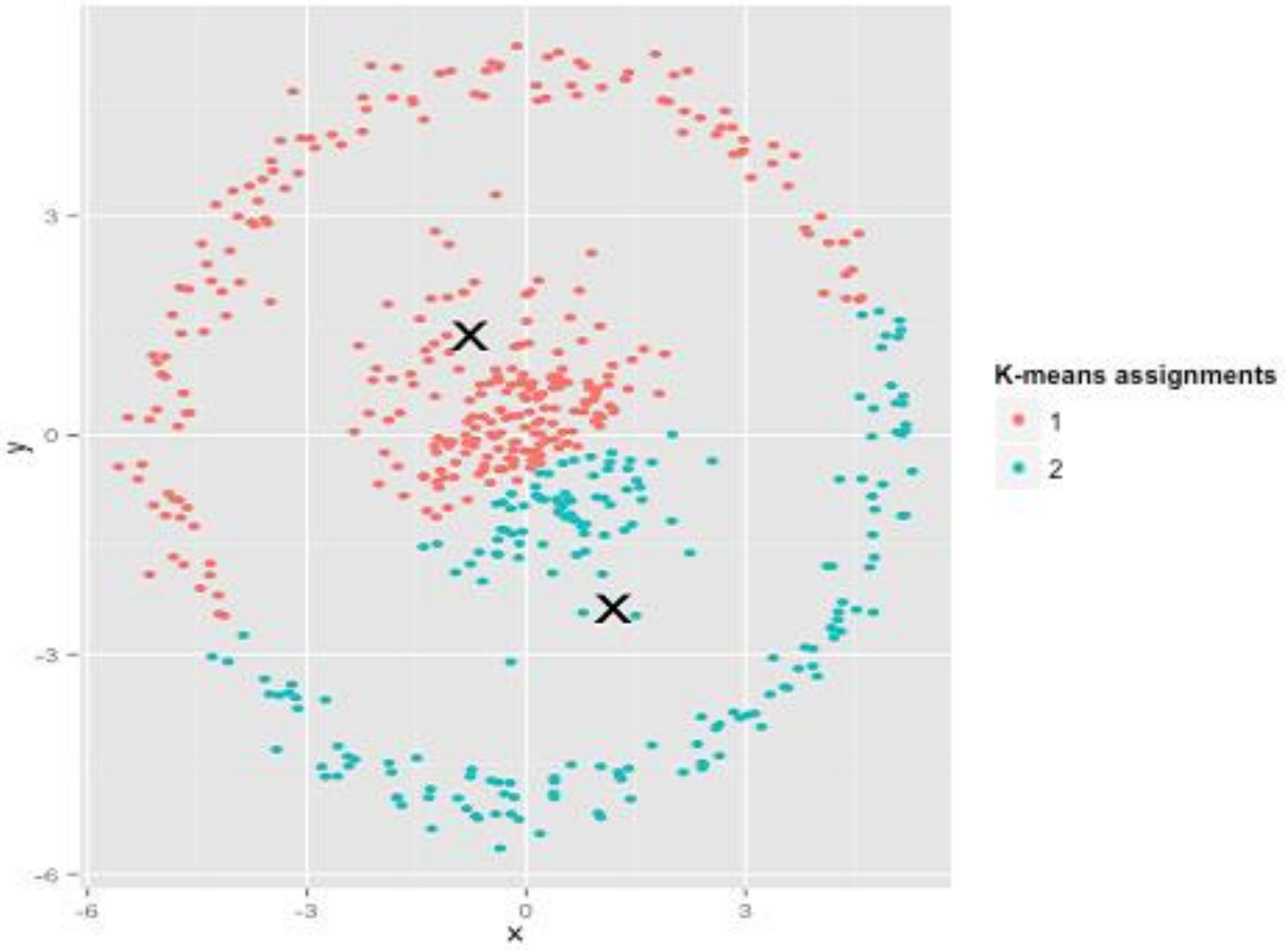


Non clustered Data





Clusters are expected to be of same size



Spherical Clusters

Pros:

- Simple
- Flexible
- Suitable for large dataset
- Detects spherical clusters very well

Cons:

Sensitive to initial centroids

Sensitive to outliers

Always creating spherical clusters

Not applicable to categorical data