Lecture 13 Clustering

What is Clustering?

- Attach label to each observation or data points in a set
- You can say this “unsupervised classification”
- Clustering is alternatively called as “grouping”
- Intuitively, if you would want to assign same label to a data points that are “close” to each other
- Thus, clustering algorithms rely on a distance metric between data points
- Sometimes, it is said that the for clustering, the distance metric is more important than the clustering algorithm

Distances: Quantitative Variables

Some examples

Identity (absolute) error

$$d_j(x_{ij}, x_{i'j}) = I(x_{ij} \neq x_{i'j})$$

Squared distance

$$d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2$$

L_q norms

$$L_{qii'} = \left[\sum_j |x_{ij} - x_{i'j}|^q\right]^{1/q}$$

Canberra distance

$$d_{ii'} = \sum_j \frac{|x_{ij} - x_{i'j}|}{|x_{ij} + x_{i'j}|}$$

Correlation

$$\rho(x_i, x_{i'}) = \frac{\sum_j (x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}_{i'})}{\sqrt{\sum_j (x_{ij} - \bar{x}_i)^2 \sum_j (x_{i'j} - \bar{x}_{i'})^2}}$$

Data point:

$x_i = [x_{i1} \ldots x_{ip}]^T$

Distances: Ordinal and Categorical Variables

- Ordinal variables can be forced to lie within (0, 1) and then a quantitative metric can be applied:

$$\frac{k-1/2}{M},\ k=1,2,\ldots,M$$

- For categorical variables, distances must be specified by user between each pair of categories.

Combining Distances

- Often weighted sum is used:

$$D(x_i, x_j) = \sum_{l=1}^{p} w_l d(x_{il}, x_{jl}), \quad \sum_{l=1}^{p} w_l = 1,\ w_l > 0.$$

K-means Overview

- An unsupervised clustering algorithm
- "*K*" stands for number of clusters, it is typically a user input to the algorithm; some criteria can be used to automatically estimate *K*
- It is an approximation to an NP-hard combinatorial optimization problem
- *K*-means algorithm is iterative in nature
- It converges, however only a local minimum is obtained
- Works only for numerical data
- Easy to implement

K-means: Setup

- $x_1, \ldots, x_N$ are data points or vectors of observations
- Each observation (vector x_i) will be assigned to one and only one cluster
- $C(i)$ denotes cluster number for the i^{th} observation
- Dissimilarity measure: Euclidean distance metric
- *K*-means minimizes within-cluster point scatter:

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(j)=k}\left\|x_i - x_j\right\|^2 = \sum_{k=1}^{K} N_k \sum_{C(i)=k}\left\|x_i - m_k\right\|^2$$

(Exercise)

where

m_k is the mean vector of the k^{th} cluster

N_k is the number of observations in k^{th} cluster

Within and Between Cluster Criteria

Let's consider total point scatter for a set of *N* data points:

$$T = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} d(x_i, x_j)$$

Distance between two points

T can be re-written as:

$$T = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\left(\sum_{C(j)=k} d(x_i, x_j) + \sum_{C(j)\neq k} d(x_i, x_j)\right)$$
$$= W(C) + B(C)$$

Where,

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(j)=k} d(x_i, x_j)$$

Within cluster scatter

$$B(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(j)\neq k} d(x_i, x_j)$$

Between cluster scatter

If *d* is square Euclidean distance, then

$$W(C) = \sum_{k=1}^{K} N_k \sum_{C(i)=k} \|x_i - m_k\|^2$$

and $$B(C) = \sum_{k=1}^{K} N_k \|m_k - m\|^2$$ Ex.

Grand mean

Minimizing *W*(*C*) is equivalent to maximizing *B*(*C*)

K-means Algorithm

- For a given cluster assignment *C* of the data points, compute the cluster means m_k:

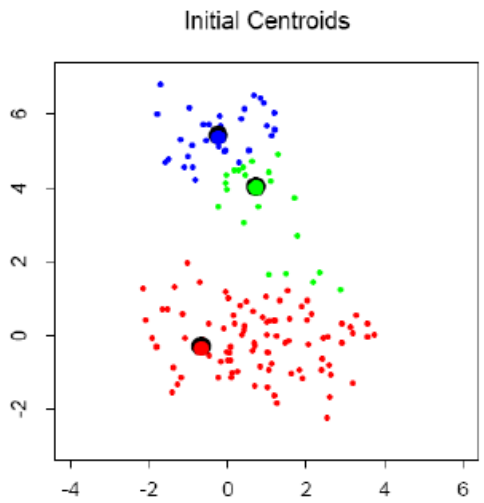
$$m_k = \frac{\sum_{i:C(i)=k} x_i}{N_k},\ k = 1, \ldots, K.$$

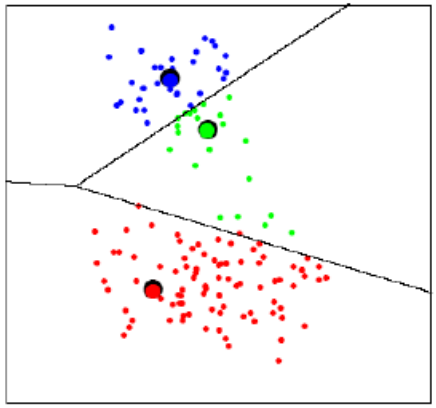
- For a current set of cluster means, assign each observation as:

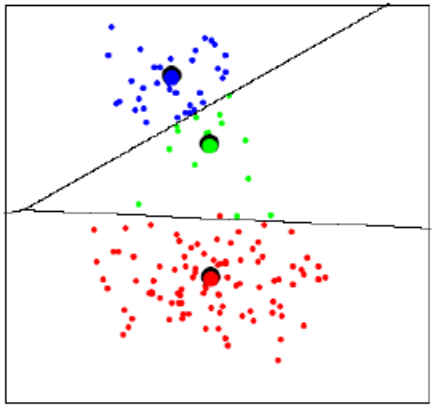
$$C(i) = \arg\min_{1 \le k \le K} \left\| x_i - m_k \right\|^2,\ i = 1, \ldots, N$$

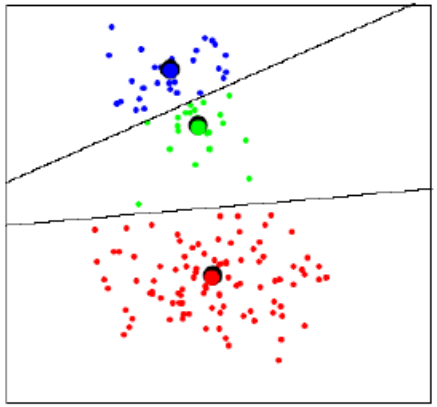
- Iterate above two steps until convergence

K-means clustering example

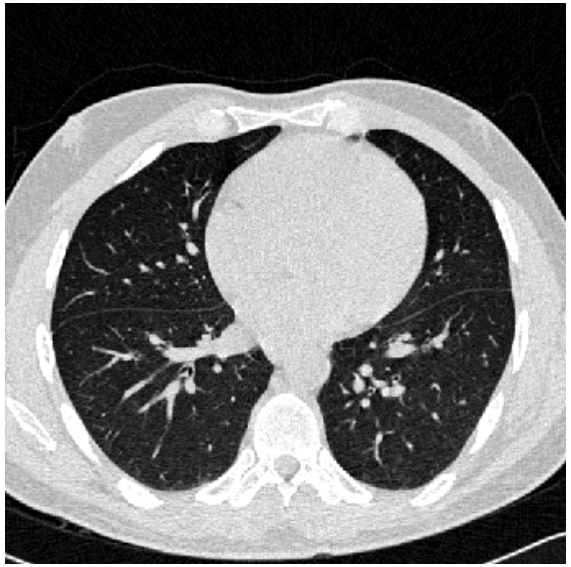








K-means Image Segmentation



An image (*I*)

Three-cluster image (*J*) on gray values of *I*

Matlab code:

```
I = double(imread('…'));

J = reshape(kmeans(I(:),3),size(I));
```

Note that *K*-means result is "noisy"

K-means: summary

- Algorithmically, very simple to implement
- *K*-means converges, but it finds a local minimum of the cost function
- Works only for numerical observations
- *K* is a user input; alternatively BIC (Bayesian information criterion) or MDL (minimum description length) can be used to estimate *K*
- Outliers can considerable trouble to *K*-means