CIS501 – Lecture 12

Woon Wei Lee
Fall 2013, 10:00am-11:15am,
Sundays and Wednesdays

Masdar INSTITUTE

For today:

- Administrative stuff
  - Updated presentation list
- Unsupervised learning
  - Cluster analysis
  - K-means clustering
  - “K-centres”
- Presentations
  - Chih-Hsien Chou
  - Tzu-Chun Lin

Introduction to unsupervised learning

- **Definition:**
  - Learning about data without the use of target labels.
  - “Exploratory analysis”
- **Discovery of intrinsic properties of the data**
  - Without any pre-conceived targets or objectives
  - Often more valuable the supervised methods
    - Supervised methods → “Help me to answer this question”
    - Unsupervised methods → “What is the questions which I should be asking?”
- **Examples of unsupervised learning techniques:**
  - Density estimation
  - Clustering
  - Adaptive signal processing
  - Mappings/Projections/Correlations
  - Visualizations

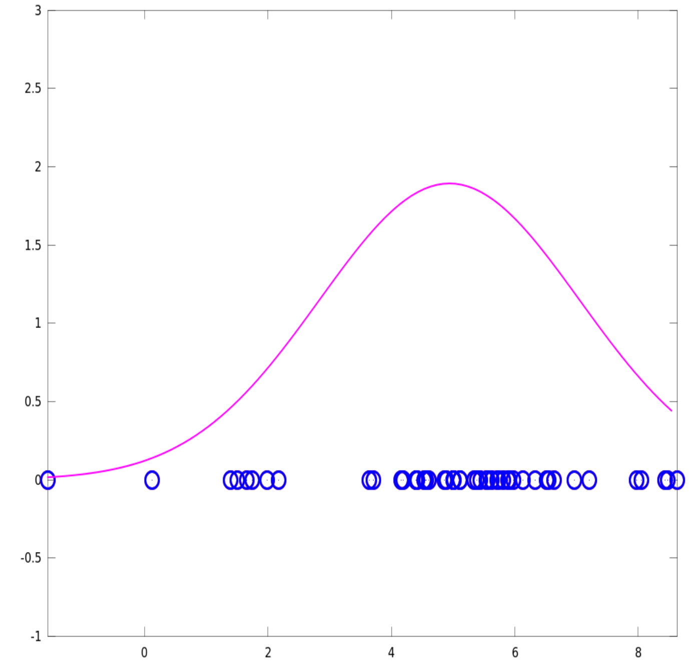
A simple density estimation problem

- **Given a set of data points, find a suitable probabilistic m**
- **Gaussian model**
  - Simply find the mean and std of the data:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x(i) \quad , \quad \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x(i)-\mu)^2$$

  - Then, the data density is:

$$p(x|\mu,\sigma) = \frac{1}{(2\pi\sigma)^{1/2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$



- **Why do we want to do this?**
  - Outlier detection
    - Having described data, anything that doesn't fit that description can be detected)
    - Understanding of nature/physics of the data
    - Ability to “generate” data points for simulations, etc.
    - Build classifiers, probabilistic models, etc.

Masdar INSTITUTE

Slightly fancier..

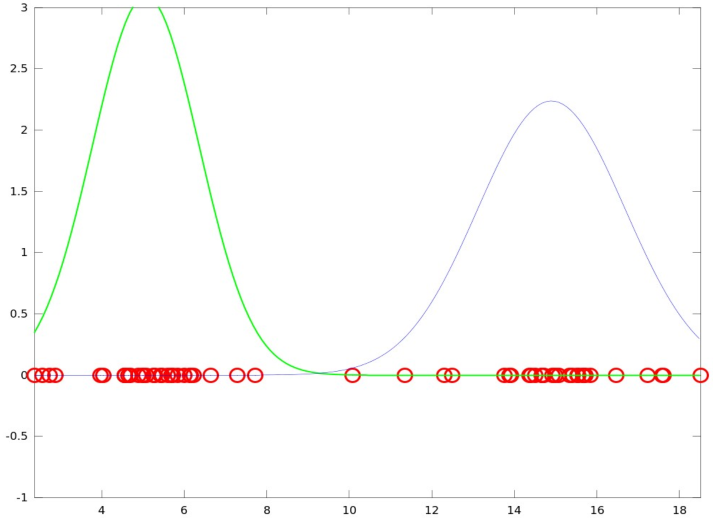
- **Next consider slightly more elaborate data collection on right →**
  - Data is now drawn from two separate probability distributions.

$$\mu_1 = \frac{1}{n}\sum_{i=1}^{n} x_1(i)\ ,\quad \sigma_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_1(i)-\mu\right)^2$$

$$\mu_2 = \frac{1}{n}\sum_{i=1}^{n} x_2(i)\ ,\quad \sigma_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_2(i)-\mu\right)^2$$

  - In which case the PDF becomes:

$$p(x|\theta) = \sum_{i=1,2}\left[\frac{1}{(2\pi\sigma_i)^{1/2}} e^{\frac{(x-\mu_i)^2}{2\sigma^2}}\right]$$



- **All this may seem familiar..**
  - Problem is.. how do we find the μ's and σ's if we don't have the class labels?

Cont'd

- **Classic unsupervised learning problem**
  - The particular form of probability density encountered here is known as a “mixture model”
  - Trick is to assign each point to one of two component PDFs
    - without knowing labels in advance!
- **As in previous problems, an iterative solution might be attempted:**
  - Assume (random?) initial values for the μ's and σ's
  - Assign each point to one gaussian or the other
  - Recalculate parameters
  - Iterate until termination met
- **Is in fact such an algorithm, known as the “EM-Algorithm”**
  - Won't be covered here, but we will be looking at a related technique known as *k-means* clustering
  - Leads to discussion on ***clustering***

Clustering

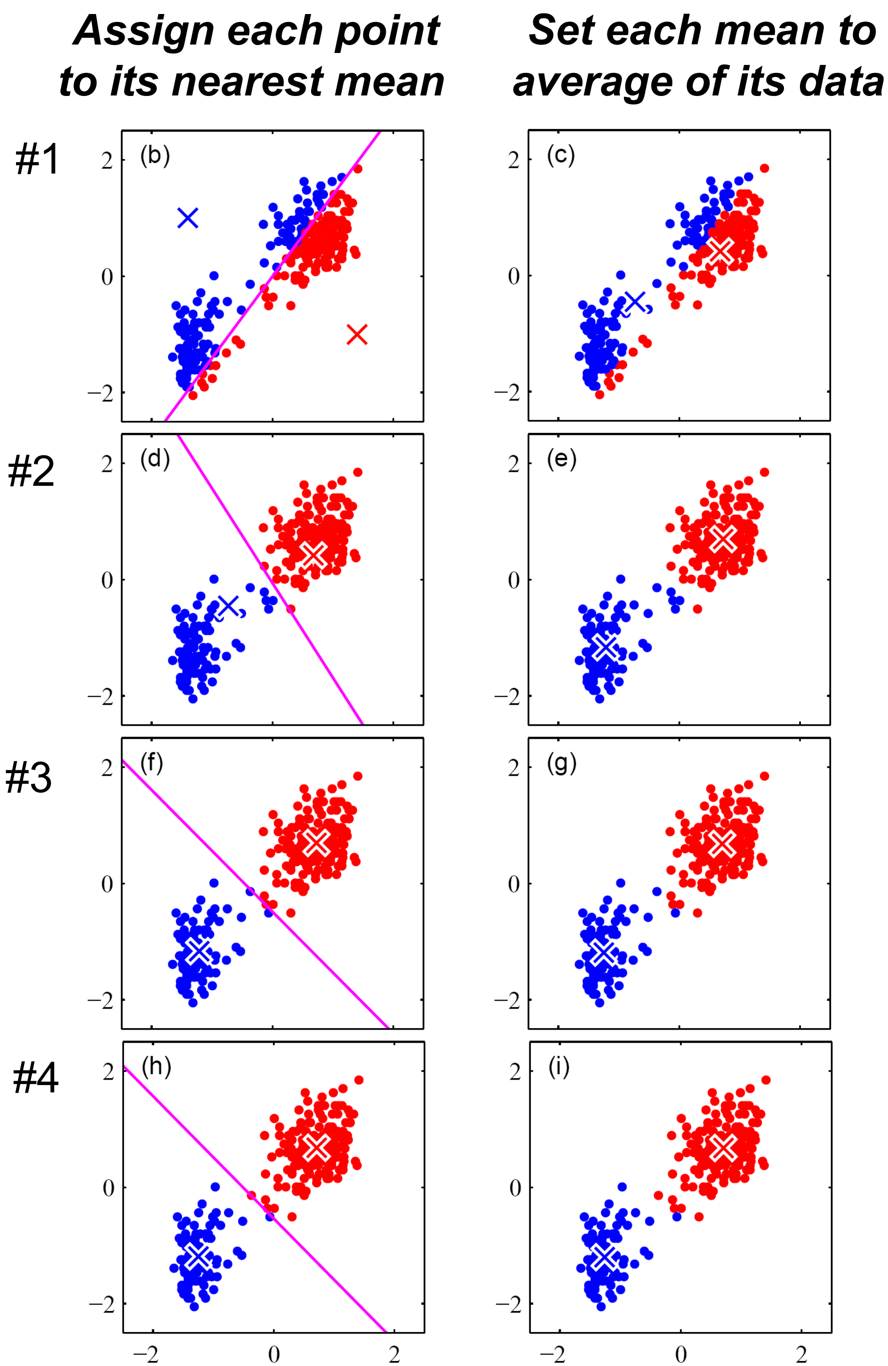
- **Definitions**
  - A.K.A. “Cluster Analysis”
  - Detection of *self-similar* groups within data sets
  - For historical reasons, clustering and unsupervised learning are often used interchangeably
    - (But, this is not accurate!)
- **The basics:**
  - Collection of techniques and algorithms for finding these groups or *clusters*
  - Objective is to partition data set such that
    - objects which are similar to each other are grouped together
    - Dissimilar objects segregated
- **Applications**
  - As a tool for understanding the underlying classes/modes/distribution of the data
  - A pre-processing step for other algorithms

Aspects of clustering

- **Distance measure**
  - Measure of dissimilarity/similarity between objects *d(i,j)*
- **Quality measure**
  - A means of measuring the quality or “success” of a clustering operation
  - In general is some comparison of inter-class to intra-class distances
- **Algorithms - two main classes**
  - Partition-based algorithms
    - ***K*-means**
    - Mixture models
  - Hierarchical algorithms
    - Two subclasses:
      - Agglomerative (“bottom up”)
      - Divisional (“top down”)

Masdar INSTITUTE

K-means clustering



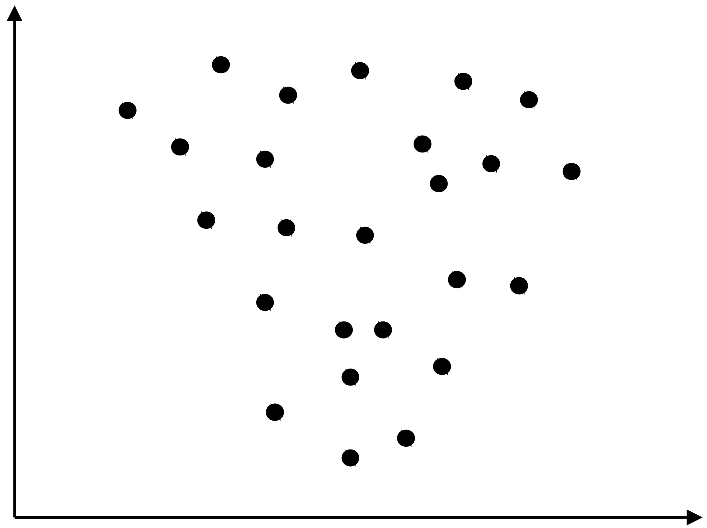
- **Is an *algorithm* for partitioning data set into different groups or clusters**
  - Choose a value of *k* (note: this sets the complexity of the model!)
  - Generate *k* initial cluster centers or *centroids*
  - Assign each point to the closest centroid
  - Recompute the location of the centroids using existing class memberships
  - Iterate until stopping criterion is met
- **Possible stopping criteria include:**
  - Cessation of membership re-assignments
  - Convergence of the centroid locations
  - Convergence of the sum squared error measure:

$$SSE = \sum_{j=1}^{k} \sum_{x \in C_j} dist(x, m_j)^2$$

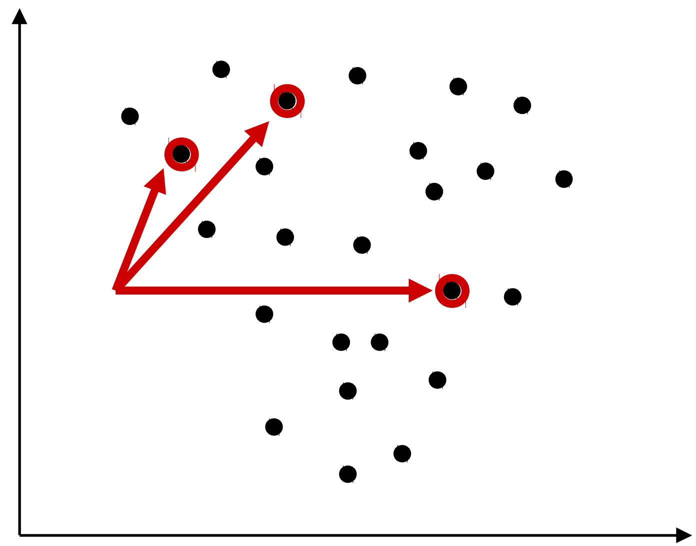
K-centers

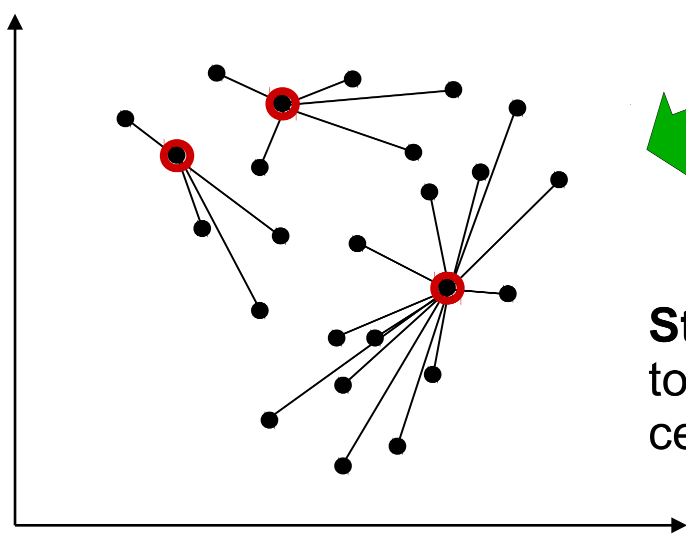
- **One of the main advantages of the k-means algorithm is that it is extremely simple**
  - Hence, highly configurable
  - A variety of customized versions exist
- **The *k-centers* algorithm allows clustering to be performed with non-vector space data**
  - All that is required is that we know the *distances* between all pairs of points
- **Useful in a variety of situations:**
  - Bioinformatics → might need to cluster DNA sequences..
  - Document or word clustering → similarity could be defined via Google searches, for e.g.
  - Clustering of signals or features via mutual information/correlation
  - Also permits the use of “kernel” methods, which allow clustering in high dimensional feature spaces

K-centers (Cont'd)



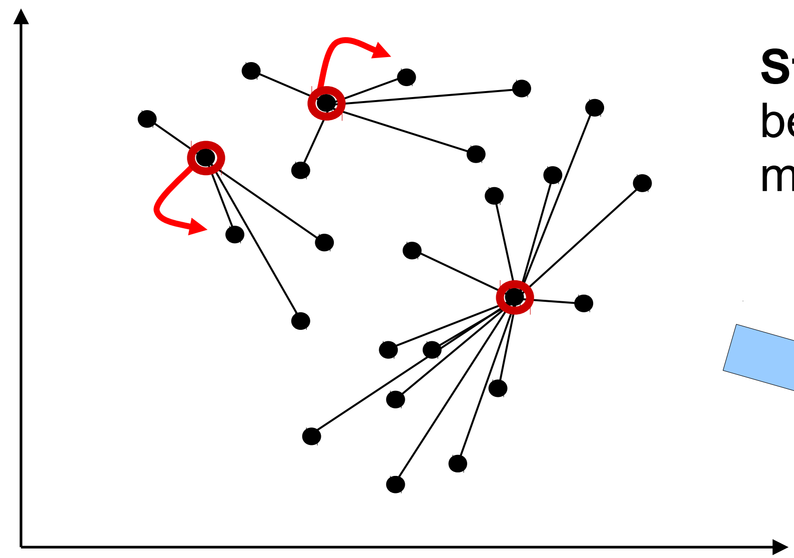
Step 1: Randomly select *k* individuals from the data set



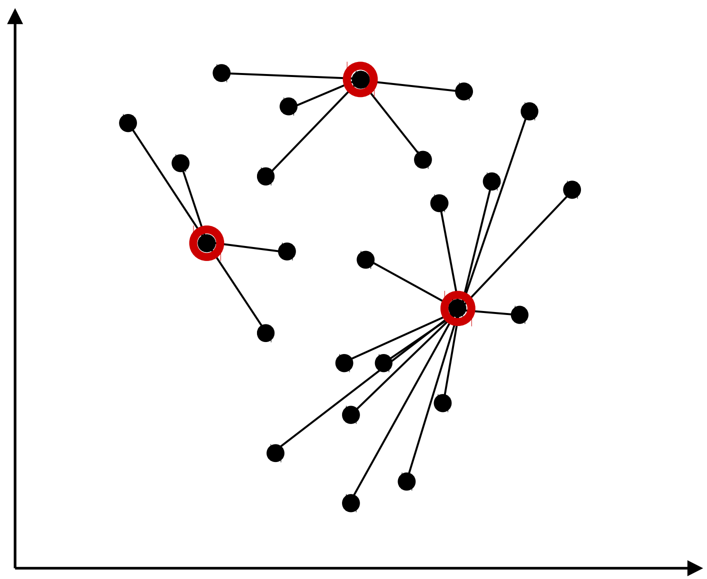


Step 2: Assign remaining points to closest amongst the existing centroids

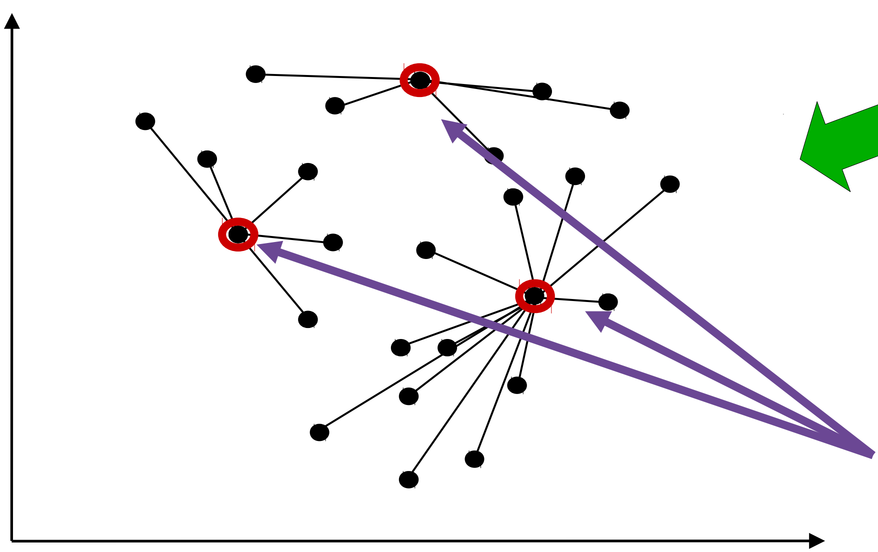
K-centers (Cont'd)



Step 3: For each cluster, pick best new centroid (based on minimum distance)



Step 2*: Assign remaining points to closest amongst the existing centroids



Convergence: Terminate iteration when stopping criterion met

Clustering quality measures

- **Why?**
  - A means of validation – does clustering work at all?
    - Difficult to tell with high dimensional data!
  - Model order selection..
  - Clustering algorithms are often stochastic – can repeat and choose best outcome
  - Allows direct optimization of cluster partitions
- **Dunn index**

$$DI(c) = \min_{i,j \in c : i \neq j} \left\{ \frac{\delta(A_i, A_j)}{\max_{k \in c} \left(\Delta(A_k) \right)} \right\}$$

(Large *DI* is good!)

  - $\delta(A_i, A_j)$ is the distance between the two closest points in clusters *i* and *j*
  - $\triangle(A_i)$ is the cluster “diameter”: i.e. the distance between the two furthest points in cluster *i*.

Quality measures (cont'd)

- **Davies-Bouldin Index**

$$DB(c) = \frac{1}{c} \sum_{i \in c} \max_{i \neq j} \left\{ \frac{\Delta(A_i) + \Delta(A_j)}{\delta(A_i, A_j)} \right\}$$ (Small *DB* is good!)

  - $\triangle(A_i)$ and $\delta(A_i, A_j)$ have the same meanings as in previous formula

- **C-index**

$$C = \frac{S - S_{min}}{S_{max} - S_{min}}$$ (Small C is good!)

  - *S* – sum of distances between all pairs of objects which are in the same cluster(s)
  - S_{min} – sum of the *n* smallest distances between all pairs of objects
  - S_{max} – sum of the *n* biggest distances between all pairs of objects

(cont'd)

- **“External” quality metrics**
  - An alternative approach can be applied if we do in fact have labels for the data (but chose not to use it during the clustering)
  - In this case, can use any of supervised measures, such as GINI impurity, Information Gain, etc..
- **Model selection**
  - Evaluation using the quality measure mentioned here
    - e.g. by evaluating each value of *k* and finding the “kink” in the metric curve (shown on right)
- **Reliable clustering**
  - Clustering algorithms like k-means (*et al*) are heuristics and may not be globally optimal.
    - By repeating clustering operations multiple times and selecting the best options we can obtain more robust clusters
  - Also possible to use optimization algorithms like GA and Particle Swarm to directly optimize these quality metrics