BLG 454E Learning From Data

FALL 2022-2023

Assoc. Prof. Yusuf Yaslan

Clustering

Lecture Notes from Alpaydın 2010 Introduction to Machine Learning 2e © The MIT Press (V1.0) AND
Hertzmann and Fleet 2010 Machine Learning and Data Mining Lecture Notes, CSC 411/D11,
Computer Science Department, University of Toronto AND
C. Bishop 2003 Mixture of Gaussians and EM Part I and II, BCS Summer School, Exeter AND
D. Kobak 2021 Lecture Notes on Clustering and EM, Introduction to Machine Learning. University of Tübingen

Unsupervised Learning

- The collection of unlabelled data
- Discover the groups/clusters or patterns within the data
- Dimensionality Reduction and Clustering are two important examples

Classes versus Clusters

Supervised X=$\{\boldsymbol{x^t}, \boldsymbol{r^t}\}$

- Classes $C_i\ i = 1 \dots K$
- $p(x) = \sum_{i=1}^{k} p(x|C_i)p(C_i)$
- $\theta = \{p(C_i), \mu_i, \Sigma_i\}_{i=1}^{k}$
- $p(C_i) = \frac{\sum_{t=1}^{N} r_i^t}{N}$
- $m_i = \frac{\sum_{t=1}^{N} x^t\ r_i^t}{\sum_{t=1}^{N} r_i^t}$
- $S_i = \frac{\sum_{t=1}^{N} r_i^t (x^t - m_i)(x^t - m_i)}{\sum_{t=1}^{N} r_i^t}$

Unsupervised X=$\{\boldsymbol{x^t}\}$

- Clusters $G_i\ i = 1 \dots K$
- $p(x) = \sum_{i=1}^{k} p(x|G_i)p(G_i)$
- $\theta = \{p(G_i), \mu_i, \Sigma_i\}_{i=1}^{k}$
- Labels $\boldsymbol{r^t}$?

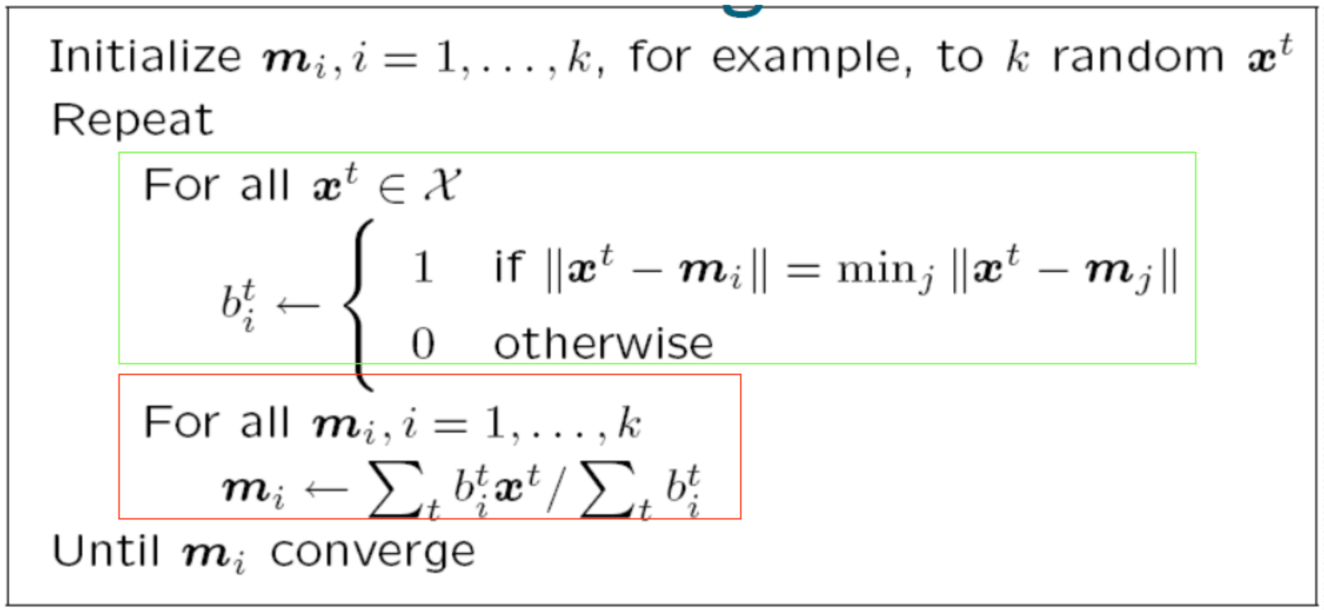
K-Means Clustering

- Given N input data vectors {x$_i$}$_{i=1}^{N}$, we wish to label each vector as belonging to one of K clusters.
- Each data point belongs to a single cluster (mutually exclusive)
- We will estimate a **center** for each cluster.
- The full objective function/``total reconstruction error''
  - $E\left(\{m_i\}_{i=1}^{k} \middle| X\right) = \sum_t \sum_i b_i^t \|x^t - m_i\|^2$

where $b_i^t = \begin{cases} 1 & if \|x^t - m_i\| = min_j \left\|x^t - m_j\right\| \\ 0 & otherwise \end{cases}$

Cannot be optimized in closed form

Pseudocode for k-means



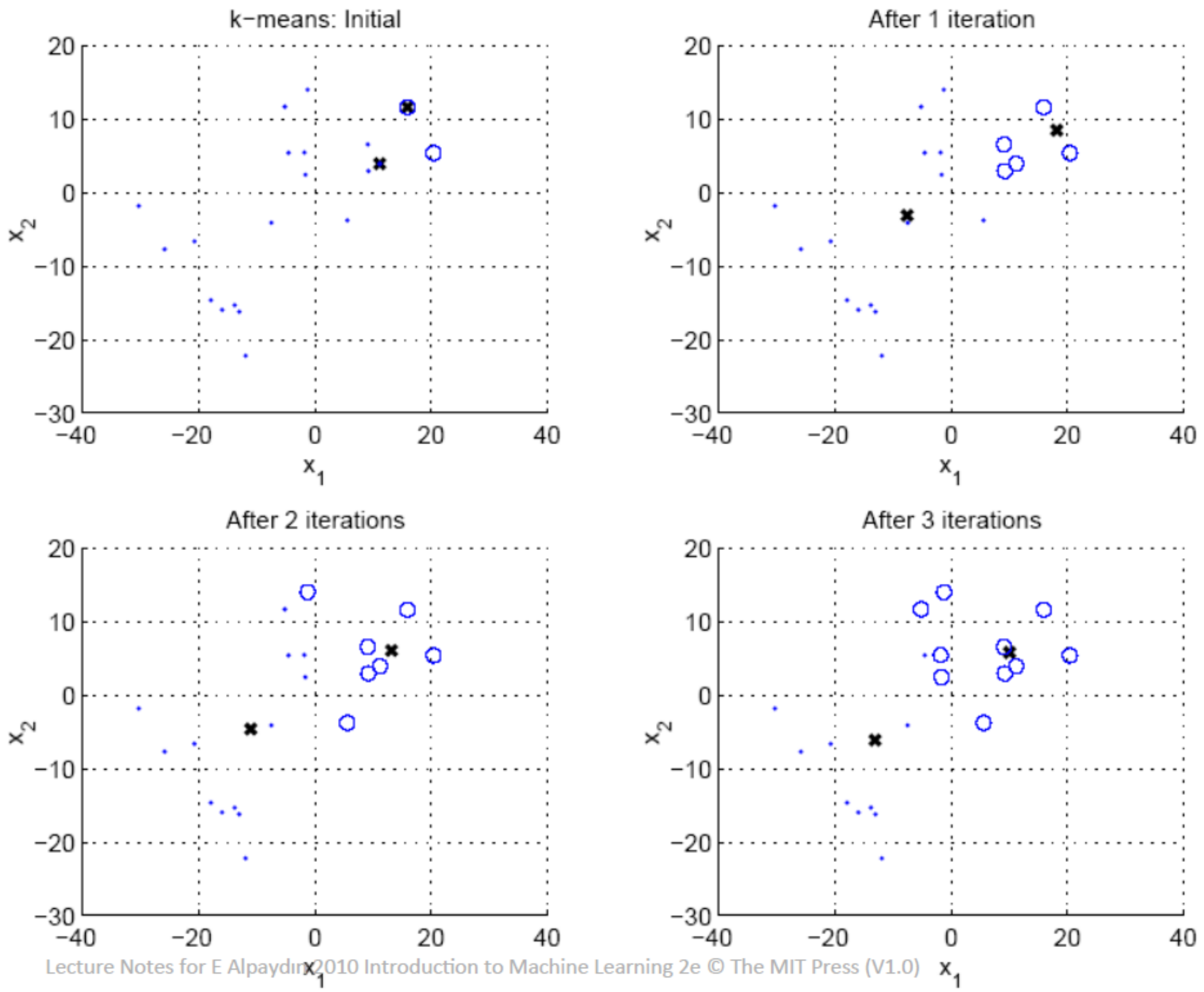
```
Initialize $\boldsymbol{m}_i, i = 1, \dots, k$, for example, to $k$ random $\boldsymbol{x}^t$
Repeat
    For all $\boldsymbol{x}^t \in \mathcal{X}$
        $b_i^t \leftarrow \begin{cases} 1 & \text{if } \|\boldsymbol{x}^t - \boldsymbol{m}_i\| = \min_j \|\boldsymbol{x}^t - \boldsymbol{m}_j\| \\ 0 & \text{otherwise} \end{cases}$
    For all $\boldsymbol{m}_i, i = 1, \dots, k$
        $\boldsymbol{m}_i \leftarrow \sum_t b_i^t \boldsymbol{x}^t / \sum_t b_i^t$
Until $\boldsymbol{m}_i$ converge
```

Figure from Alpaydin 2010. Chapter 7, Introduction to Machine Learning

Initialization

- Random labelling (not recommended)
- Mean of all data + small random vectors to get k initial means
- Random data points as centers
- Principal components and take the means of these groups
- **Multiple restart**

Visualisation of k-means



Lecture Notes for E Alpaydın 2010 Introduction to Machine Learning 2e © The MIT Press (V1.0)

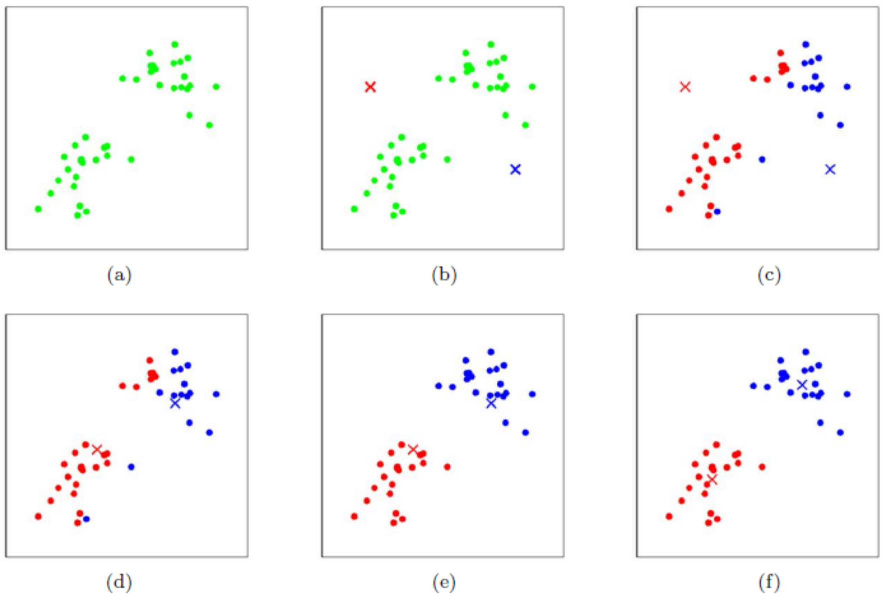
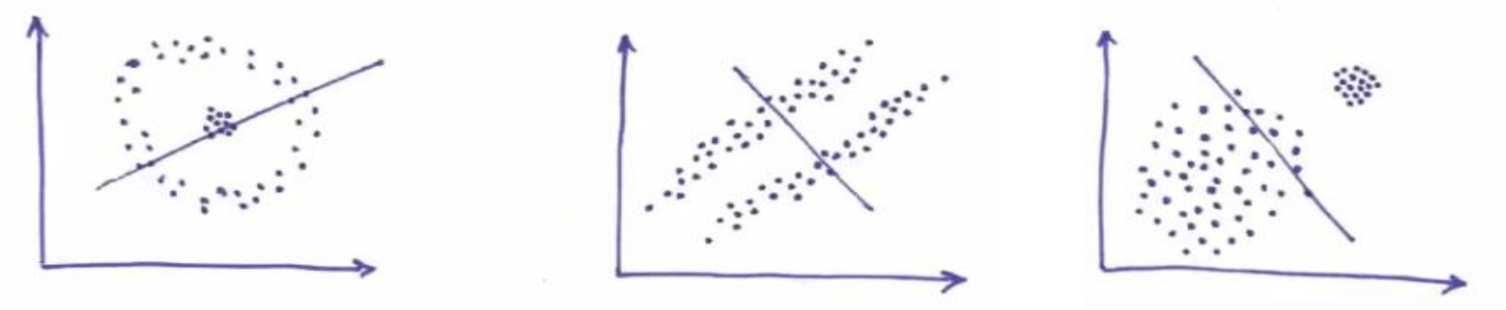

(a) (b) (c) (d) (e) (f)

Figure 1: K-means algorithm. Training examples are shown as dots, and cluster centroids are shown as crosses. (a) Original dataset. (b) Random initial cluster centroids (in this instance, not chosen to be equal to two training examples). (c-f) Illustration of running two iterations of k-means. In each iteration, we assign each training example to the closest cluster centroid (shown by "painting" the training examples the same color as the cluster centroid to which is assigned); then we move each cluster centroid to the mean of the points assigned to it. (Best viewed in color.) Images courtesy Michael Jordan.

CS229 Lecture notes, Andrew Ng, 2003

Drawbacks of k-means

- It can stuck in local minima
- It does not work on non-spherical clusters



inspired from Lecture Notes, D. Kobak, Clustering and EM, Uni of Tübingen

Hierarchical Clustering

- Cluster based on similarities/distances
- Distance measure between instances $\boldsymbol{x}^r$ and $\boldsymbol{x}^s$

Minkowski (L_p) (Euclidean for p = 2)

$$d_m\left(\mathbf{x}^r, \mathbf{x}^s\right) = \left[\sum_{j=1}^{d}\left(x_j^r - x_j^s\right)^p\right]^{1/p}$$

City-block distance

$$d_{cb}\left(\mathbf{x}^r, \mathbf{x}^s\right) = \sum_{j=1}^{d}\left|x_j^r - x_j^s\right|$$

Agglomerative Clustering

- Start with *N* groups each with one instance and merge two closest groups at each iteration
- Distance between two groups G$_i$ and G$_j$:
  - Single-link:

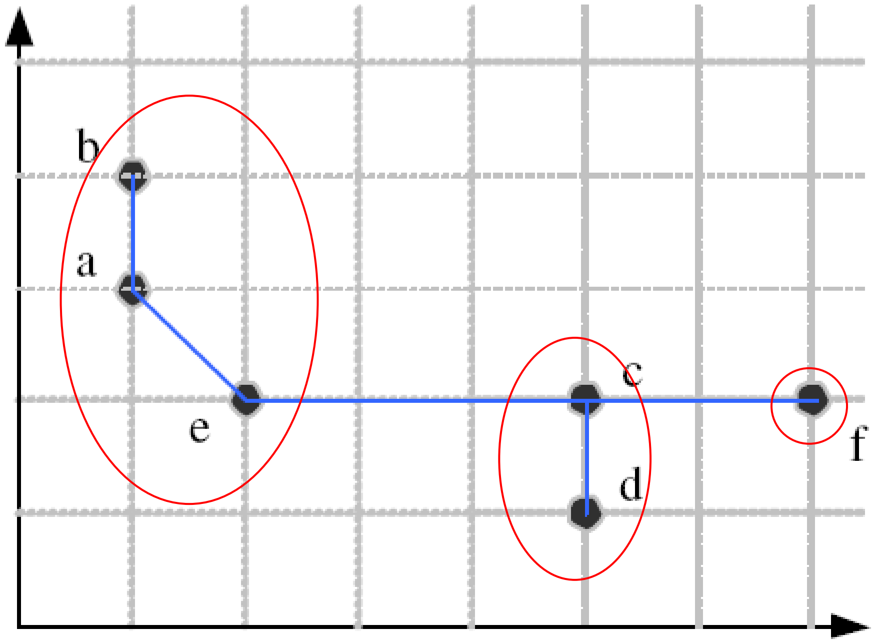
$$d(G_i, G_j) = \min_{\mathbf{x}^r \in \mathcal{G}_i, \mathbf{x}^s \in \mathcal{G}_j} d(\mathbf{x}^r, \mathbf{x}^s)$$

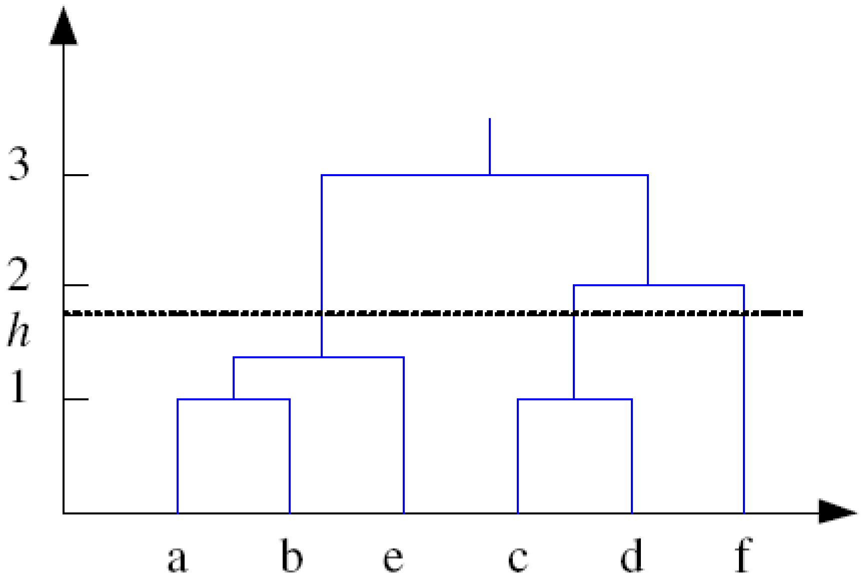
  - Complete-link:

$$d(G_i, G_j) = \max_{\mathbf{x}^r \in \mathcal{G}_i, \mathbf{x}^s \in \mathcal{G}_j} d(\mathbf{x}^r, \mathbf{x}^s)$$

  - Average-link, centroid

Example: Single-Link Clustering





Dendrogram

Choosing *k*

- Defined by the application, e.g., image quantization
- Plot data (after PCA) and check for clusters
- Incremental (leader-cluster) algorithm: Add one at a time until "elbow" (reconstruction error/log likelihood/intergroup distances)
- Manually check for meaning