CSCE 633: Machine Learning

Lecture 25: Unsupervised Learning

Texas A&M University

10-21-19

Last Time

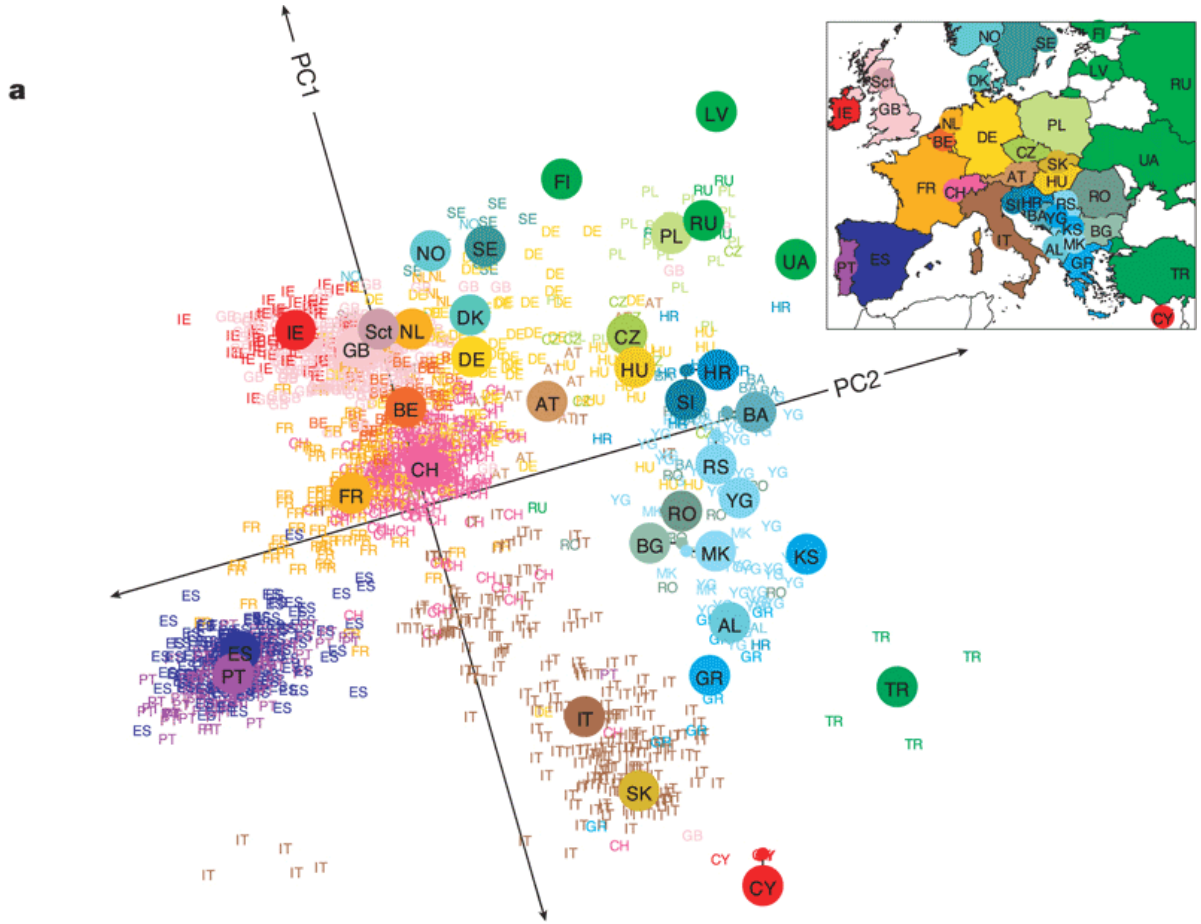
- PCA
- Clustering, K-Means
- GMM and EM

Goals of this lecture

- other clustering techniques

Clustering

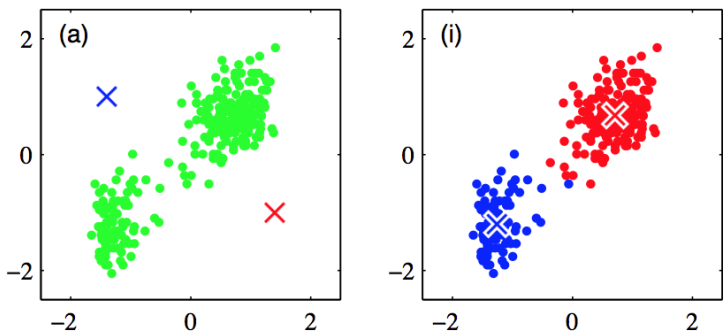
Finding patterns/structure/sub-populations in data



K-means Clustering

Representation

- Input: Data $\mathcal{D} = \{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$
- Output: Clusters $\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}$
- Decision: Cluster membership, the cluster id assigned to sample $\mathbf{x_n}$, i.e. $A(\mathbf{x_n}) \in \{1, \ldots, K\}$
- Evaluation metric: Distortion measure
  $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x_n} - \boldsymbol{\mu_k}\|_2^2$, where $r_{nk} = 1$ if $A(\mathbf{x_n}) = k$, 0 otherwise
- Intuition: Data points assigned to cluster k should be close to centroid $\boldsymbol{\mu_k}$



K-means Clustering

Evaluation metric: $\min_{r_{nk}} J = \min_{r_{nk}} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x_n} - \boldsymbol{\mu_k}\|_2^2$

Optimization:

- Step 0: Initialize $\boldsymbol{\mu_k}$ to some values
- Step 1: Assume the current value of $\boldsymbol{\mu_k}$ fixed, minimize J over r_{nk}, which leads to the following cluster assignment rule
  $r_{nk} = \begin{cases} 1, & \text{if } k = arg\min_j \|\mathbf{x_n} - \boldsymbol{\mu_j}\|_2^2 \\ 0, & \text{otherwise} \end{cases}$
- Step 2: Assume the current value of r_{nk} fixed, minimize J over $\boldsymbol{\mu_k}$, which leads to the following rule to update the prototypes of the clusters $\boldsymbol{\mu_k} = \frac{\sum_n r_{nk}\mathbf{x_n}}{\sum_n r_{nk}}$
- Step 3: Determine whether to stop or return to Step 1

K-means Clustering

Remarks

- The centroid $\boldsymbol{\mu_k}$ is the means of data points assigned to the cluster k, hence the name K-means clustering.
- The procedure terminates after a finite number of steps, as the procedure reduces J in both Step 1 and Step 2
- There is no guarantee the procedure terminates at the global optimum of J. In most cases, the algorithm stops at a local optimum, which depends on the initial values in Step 0 $\rightarrow$ random restarts to improve chances of getting closer to global optima

Expectation Maximization for GMMs

What is the E-step in GMM?

$$\mathbb{E}[\gamma_{nk}|\mathbf{x_n}, \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}] = p(\gamma_{nk} = 1|\mathbf{x_n}, \boldsymbol{\theta_k}) = \frac{p(\mathbf{x_n}|\gamma_{nk} = 1, \boldsymbol{\theta})p(\gamma_{nk} = 1|\boldsymbol{\theta_k})}{\sum_k p(\mathbf{x_n}|\gamma_{nk} = 1, \boldsymbol{\theta})p(\gamma_{nk} = 1|\boldsymbol{\theta_k})}$$

We compute the probability

$$\gamma_{nk} = p(z = k|\mathbf{x_n}, \boldsymbol{\theta}^{(t)})$$

$$\gamma_{nk} = \frac{\omega_k p(\mathbf{x_n}|\boldsymbol{\mu}_\mathbf{k}^{(t-1)}, \boldsymbol{\Sigma}_\mathbf{k}^{(t-1)})}{\sum_k \omega_k p(\mathbf{x_n}|\boldsymbol{\mu}_\mathbf{k}^{(t-1)}, \boldsymbol{\Sigma}_\mathbf{k}^{(t-1)})}$$

Expectation Maximization for GMMs

What is the M-step in GMM?

Maximize the auxiliary function

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \sum_n \sum_k p(z = k|\mathbf{x_n}, \boldsymbol{\theta}^{(t)}) \log p(\mathbf{x_n}, z = k|\boldsymbol{\theta})$$

$$= \sum_n \sum_k \gamma_{nk} \log p(\mathbf{x_n}, z = k|\boldsymbol{\theta})$$

$$= \sum_n \sum_k \gamma_{nk} \log(p(z = k)p(\mathbf{x_n}|z = k))$$

$$= \sum_n \sum_k \gamma_{nk} [\log \omega_k + \log \mathcal{N}(\mathbf{x_n}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})]$$

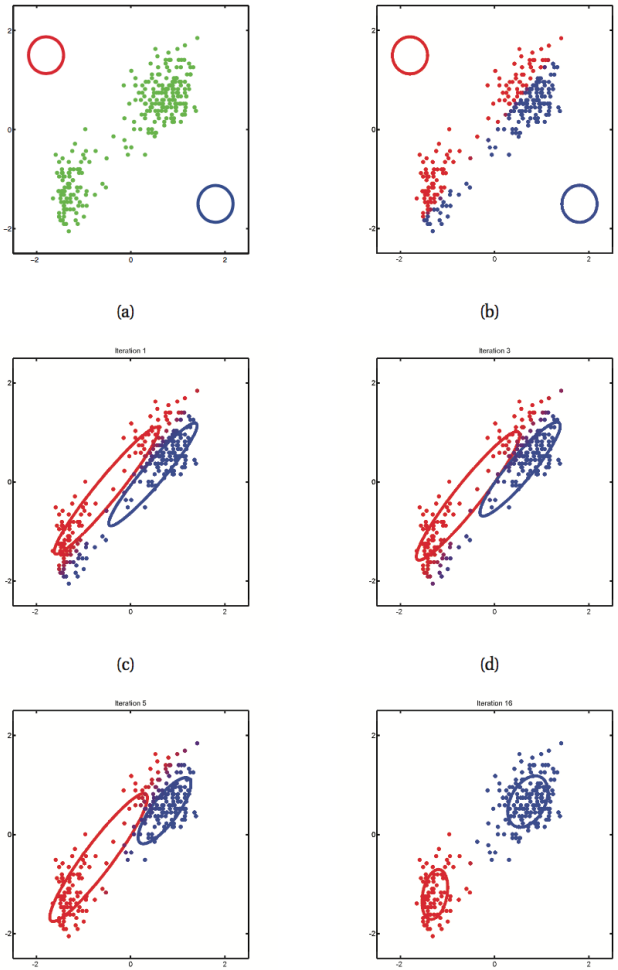
$$\Rightarrow \omega_k = \frac{\sum_n \gamma_{nk}}{N}$$

$$\Rightarrow \boldsymbol{\mu_k} = \frac{1}{\sum_n \gamma_{nk}} \sum_n \gamma_{nk} \mathbf{x_n}$$

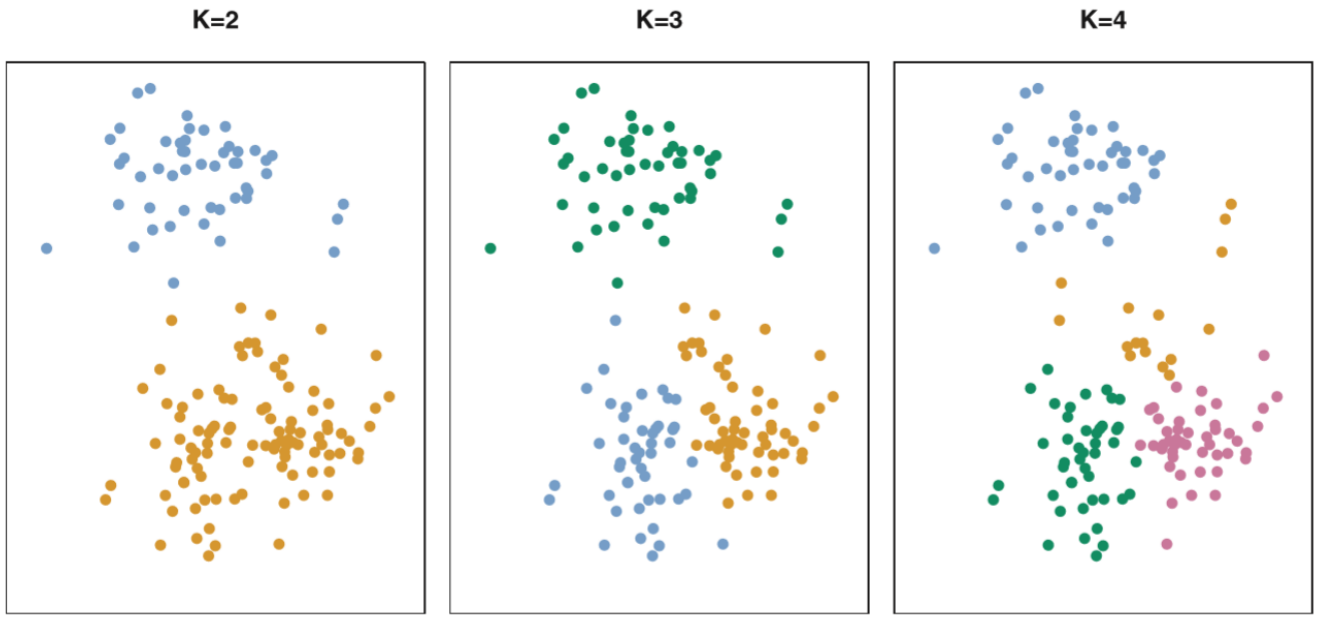
$$\Rightarrow \boldsymbol{\Sigma_k} = \frac{1}{\sum_n \gamma_{nk}} \sum_n \gamma_{nk} (\mathbf{x_n} - \boldsymbol{\mu_k})(\mathbf{x_n} - \boldsymbol{\mu_k})^T$$

Expectation Maximization for GMMs

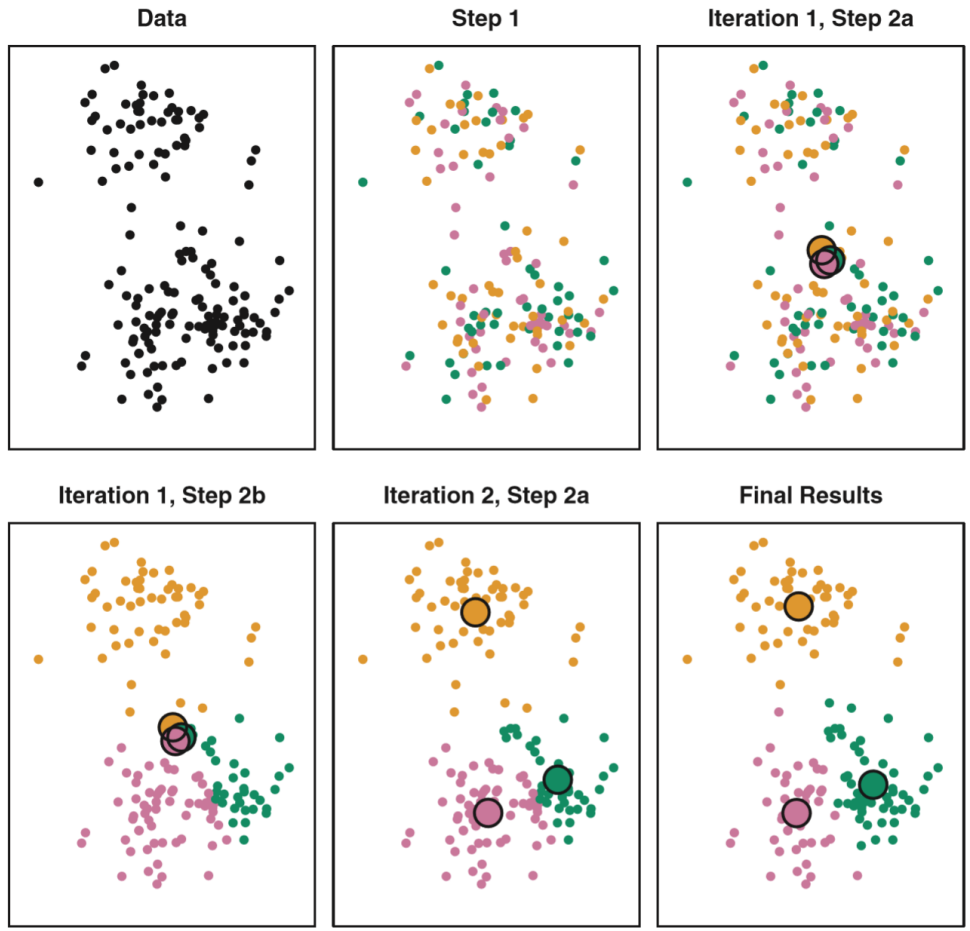
Example of EM implementation



Limitations



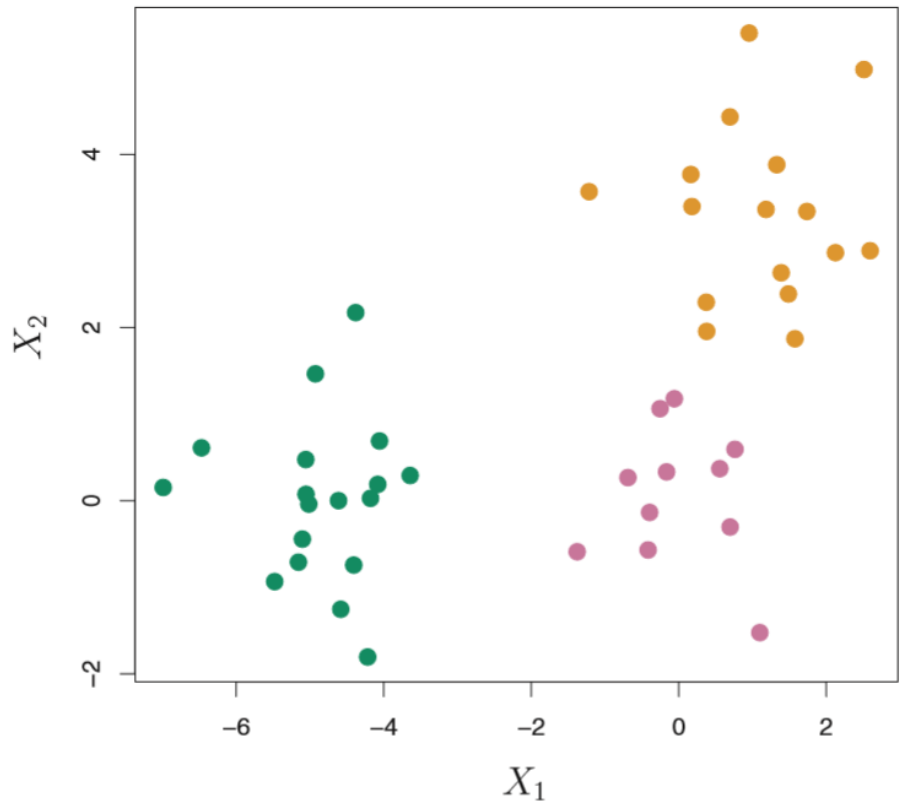
Limitations



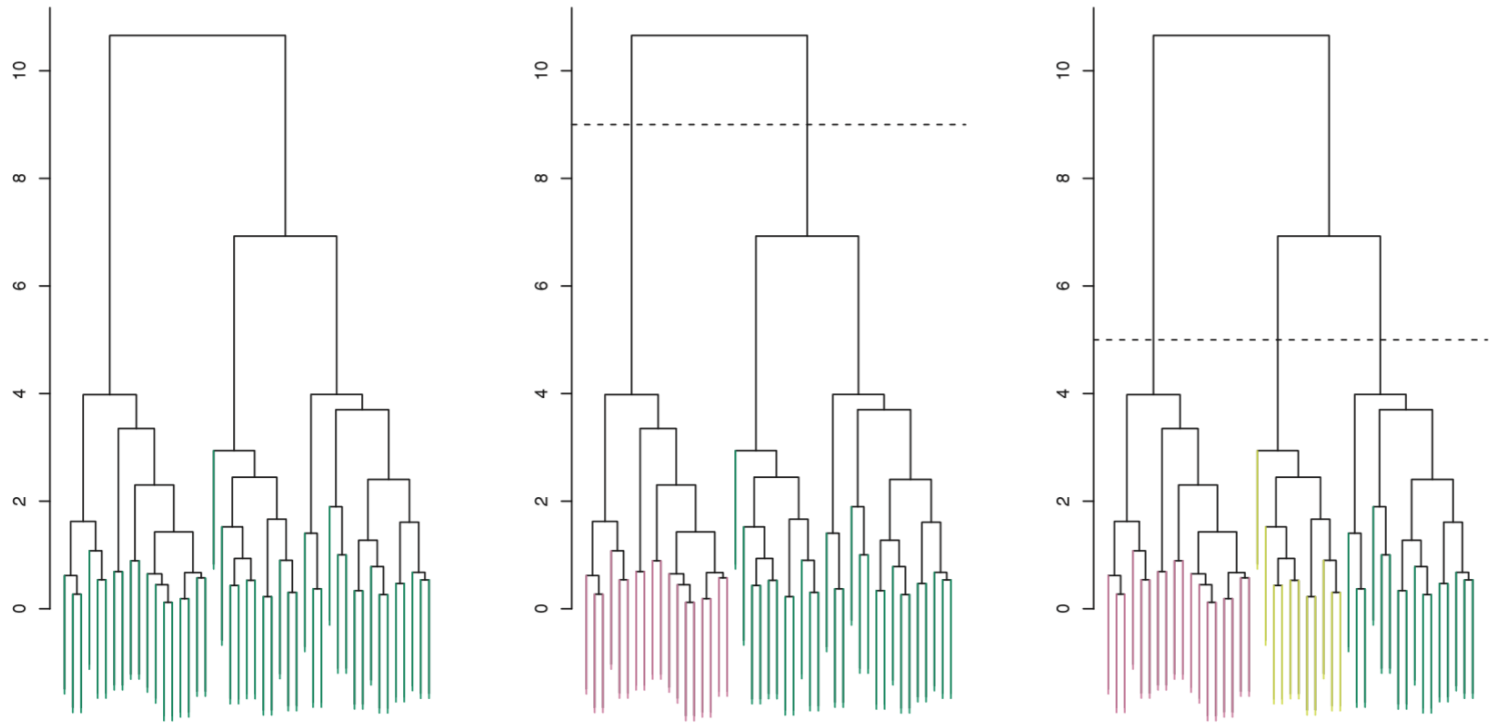
Limitations

- Could we build bottom up?
- Could we cluster without pre-defining the number of clusters?
- can do an agglomerative clustering (most common version of hierarchical clustering).

How many clusters?



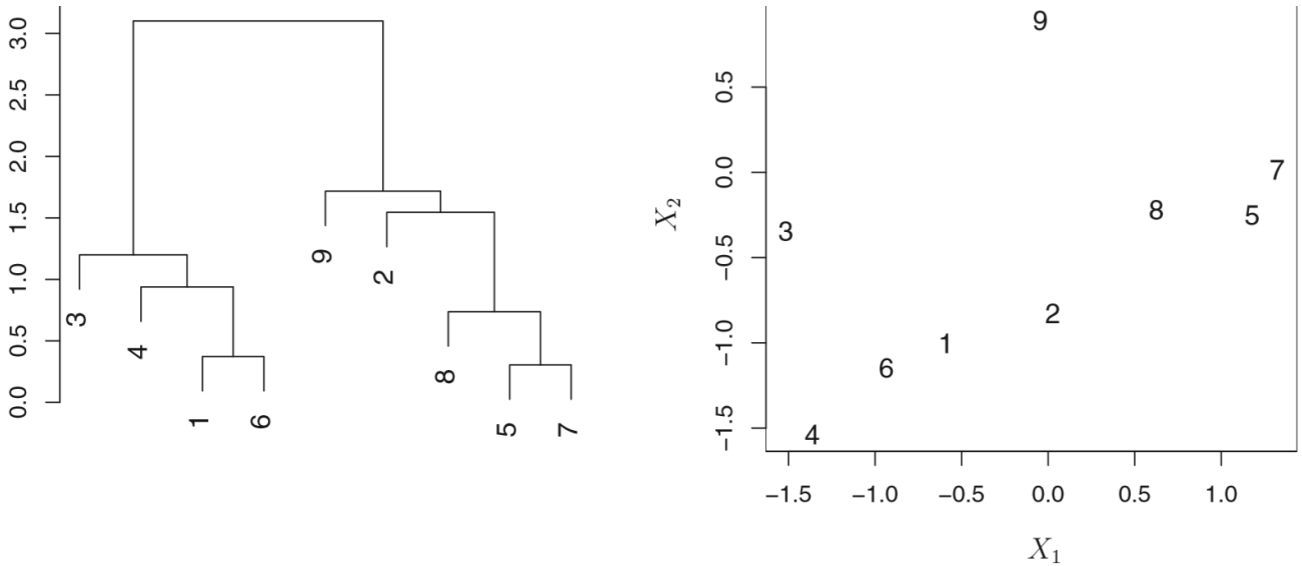
Tree-based Interpretability



Fusing

- How do you determine similarity?
- at what level do you fuse together to determine clusters?
- When you choose a level to fuse what tradeoff with similarity are you potentially making?

Tree-based Fusing



Is 9 actually that close to 2?

Interpretting

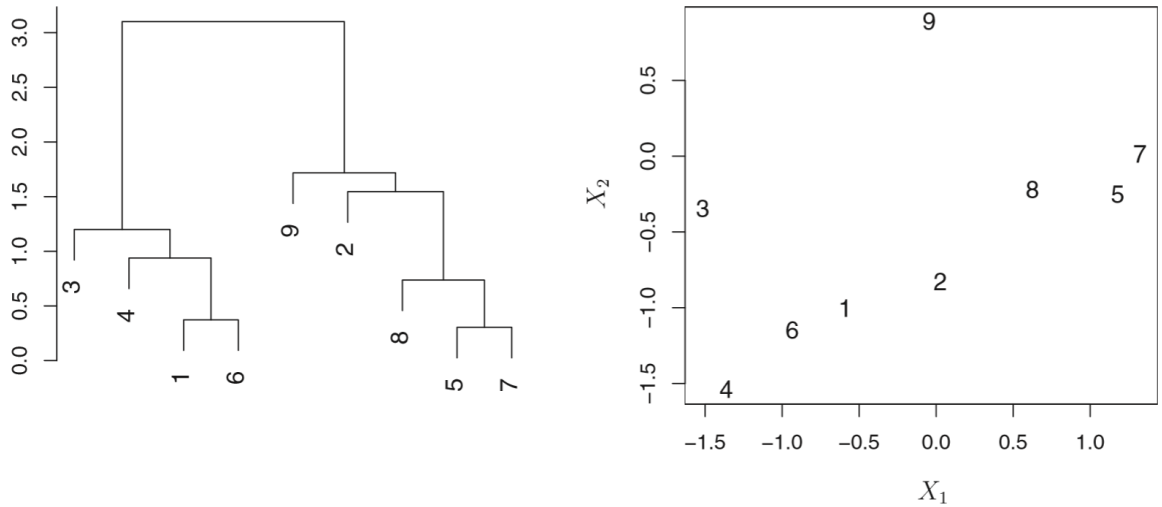


FIGURE 10.10. *An illustration of how to properly interpret a dendrogram with nine observations in two-dimensional space.* Left: *a dendrogram generated using Euclidean distance and complete linkage. Observations* 5 *and* 7 *are quite similar to each other, as are observations* 1 *and* 6. *However, observation* 9 *is* no more similar to *observation* 2 *than it is to observations* 8, 5, *and* 7, *even though observations* 9 *and* 2 *are close together in terms of horizontal distance. This is because observations* 2, 8, 5, *and* 7 *all fuse with observation* 9 *at the same height, approximately* 1.8. Right: *the raw data used to generate the dendrogram can be used to confirm that indeed, observation* 9 *is no more similar to observation* 2 *than it is to observations* 8, 5, *and* 7.

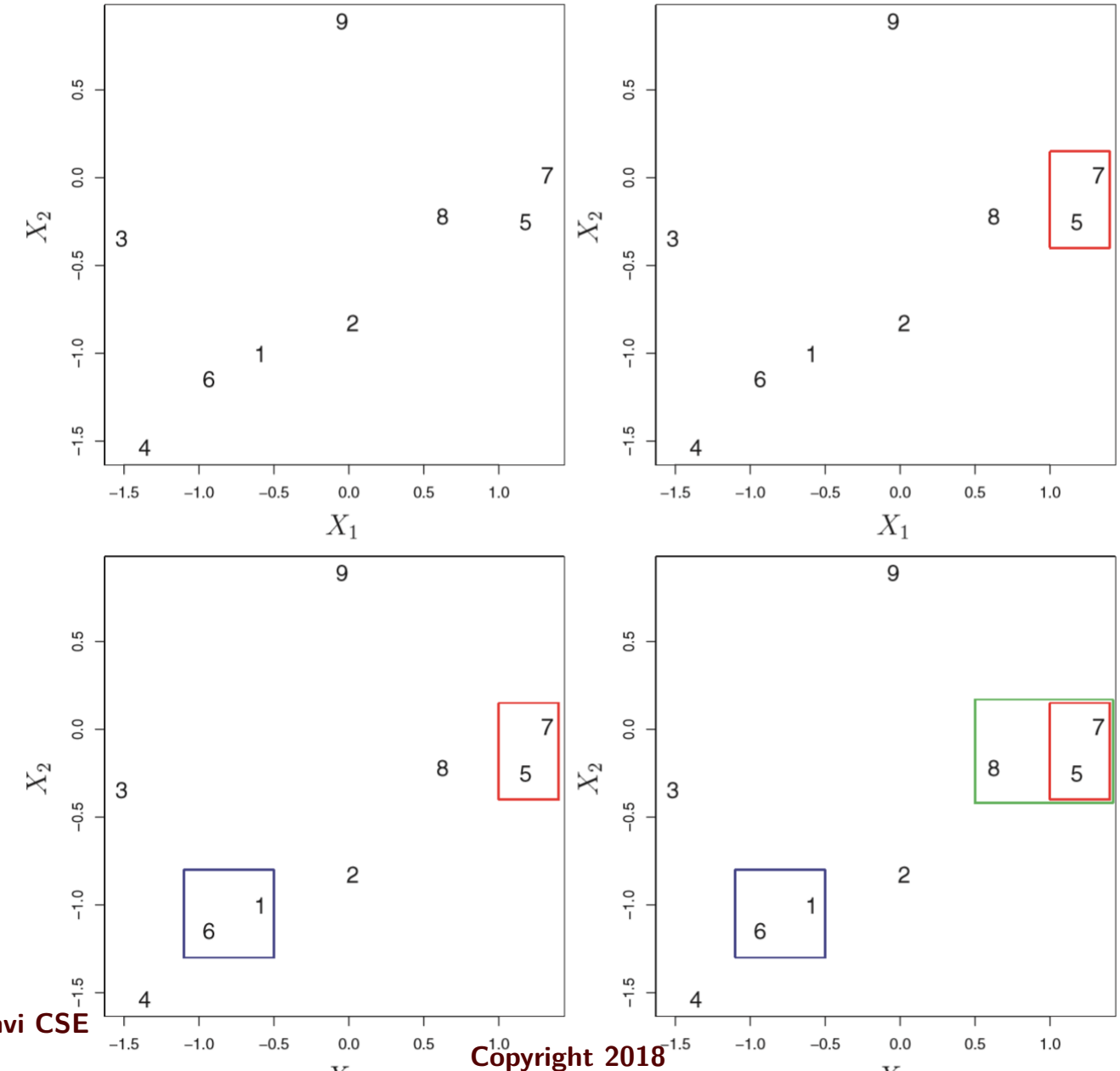
Hierarchical Clustering

- If grouping of clusters is hierarchical, can decide the right level of split
- If 3 clusters comes as a result of taking 2 clusters and splitting one then great
- If not, this might do poorly compared to k-means

Hierarchical Algorithm Principals

- Define a dissimilarity measure between pairs of observations (often Euclidean distance)
- Iteratively treat each n observations as their own clusters
- fuse the most similar items together to a cluster - so now you have $n - 1$ clusters
-

Clustering



Algorithm

Algorithm 10.2 *Hierarchical Clustering*

1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.

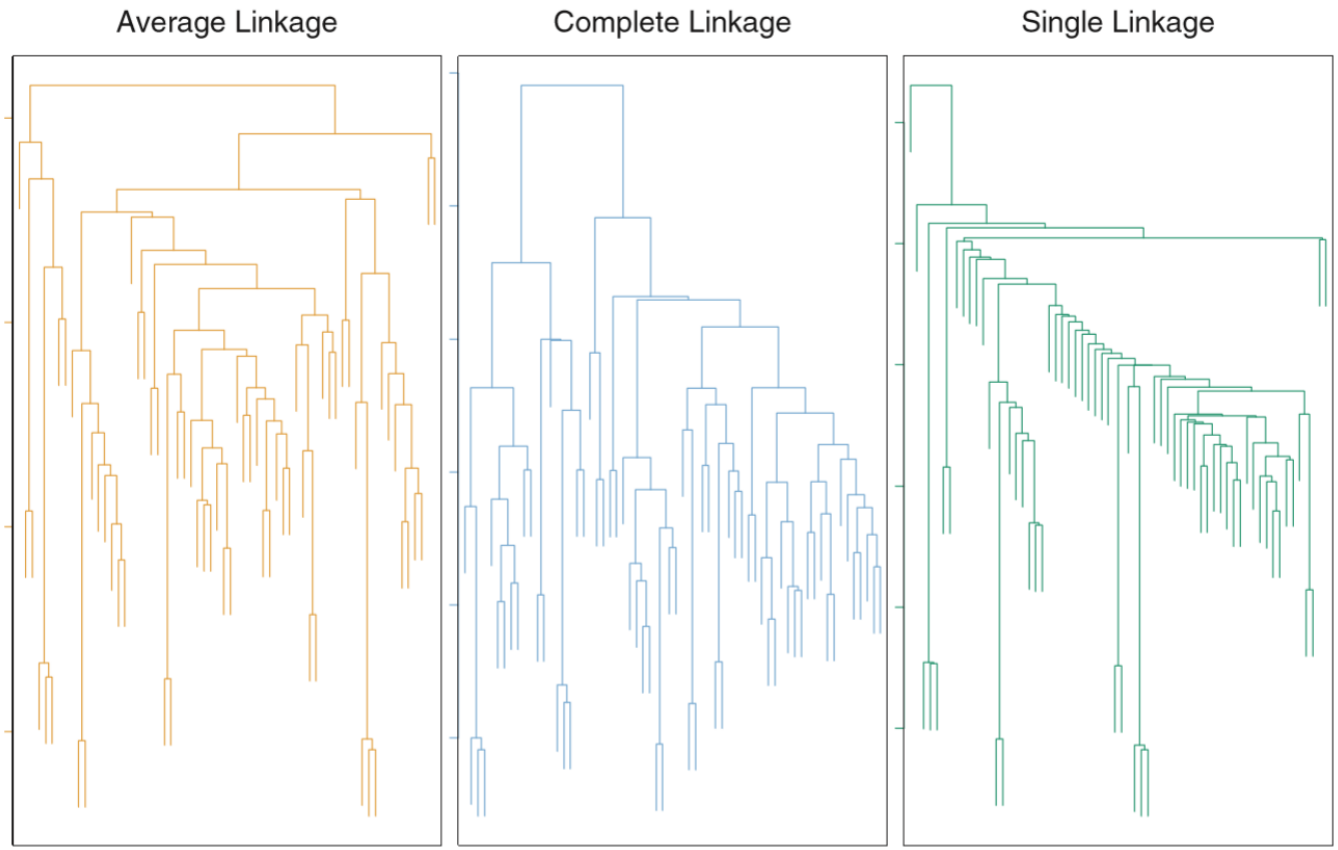
Measures

- How did we decide 8 clusters with 5 and 7?
- How do i define dissimilarity between an observation and a cluster?
- Can extend idea of similarity measures to groups of observations

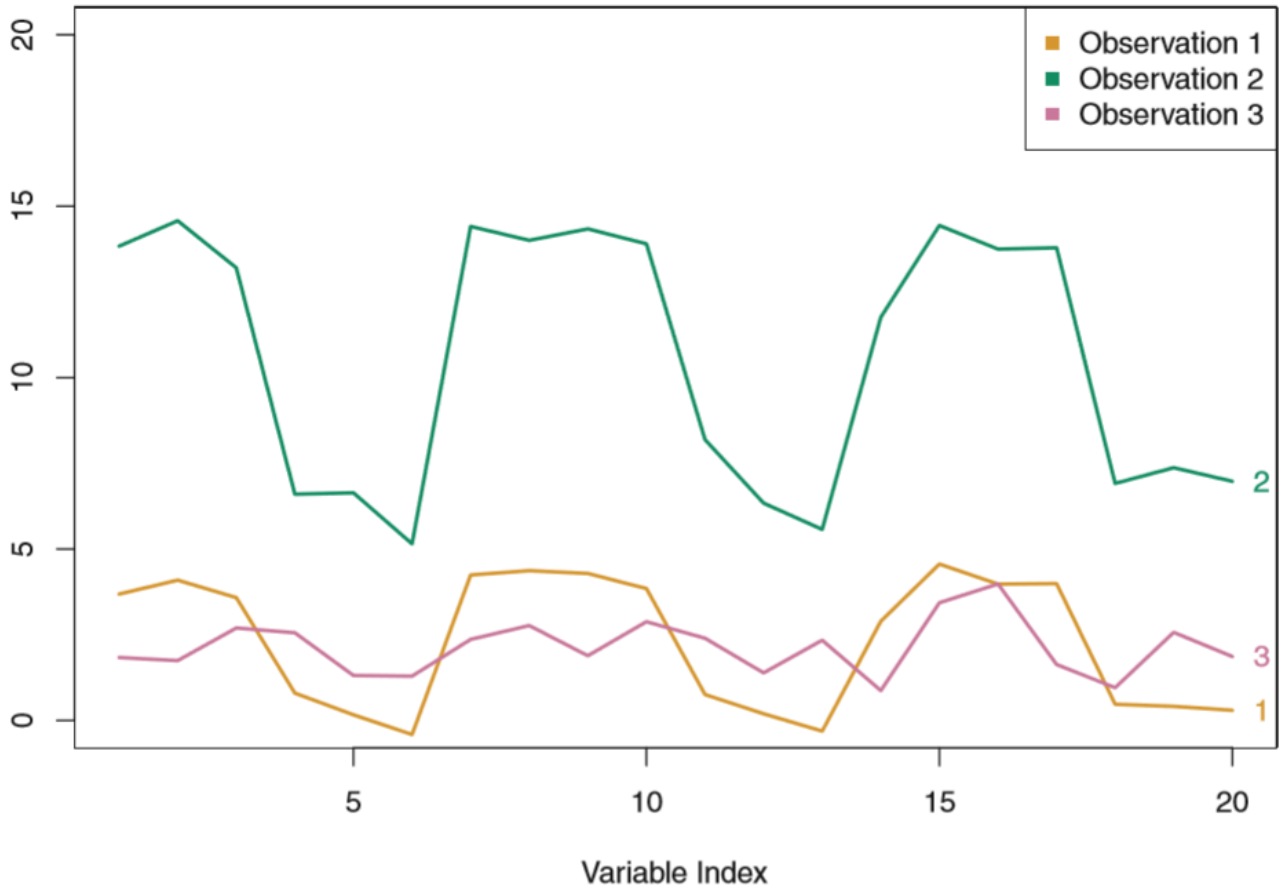
Common measures

Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

Examples



Correlation or Euclidean Distance?



Measures

For groups G and H
Single Link (nearest neighbor):

$$d_{SL}(G, H) = min_{i \in G, i' \in H} d_{i,i'}$$

Complete Link (furthest neighbor):

$$d_{CL}(G, H) = max_{i \in G, i' \in H} d_{i,i'}$$

Average Link:

$$d_{avg}(G, H) = \frac{1}{n_G n_H} \sum_{i \in G} \sum_{i' \in H} d_{i,i'}$$

, where n_G and n_H are number of elements in each group. This creates relatively compact clusters that are relatively far apart from each other.

Biclustering

- Up until now - all these techniques have involved using features to cluster observations
- What if we try to cluster observations as well as cluster features?
- For a 2D matrix - we cluster the rows and the columns - common in bioinformatics
- If rows $r_i \in \{1, \cdots, K^r\}$ and columns $c_j \in \{1, \cdots, K^c\}$ have these latent indicators, then

$$p(x|r, c, \theta) = \prod_i \prod_j p(x_{ij}|r_i, c_j, \theta) = p(x_{ij}|\theta_{r_i, c_j})$$

Biclustering

O1	killer whale, blue whale, humpback, seal, walrus, dolphin
O2	antelope, horse, giraffe, zebra, deer
O3	monkey, gorilla, chimp
O4	hippo, elephant, rhino
O5	grizzly bear, polar bear
F1	flippers, strain teeth, swims, arctic, coastal, ocean, water
F2	hooves, long neck, horns
F3	hands, bipedal, jungle, tree
F4	bulbous body shape, slow, inactive
F5	meat teeth, eats meat, hunter, fierce
F6	walks, quadrapedal, ground

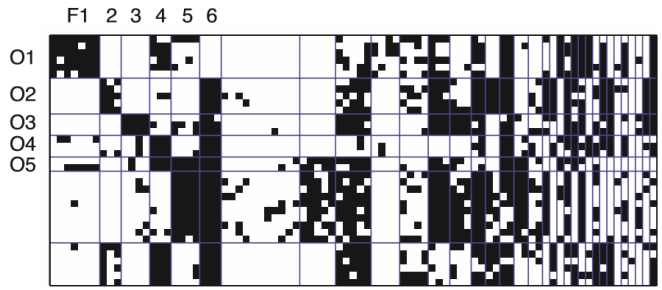


Figure 25.18 Illustration of biclustering . We show 5 of the 12 animal clusters, and 6 of the 33 feature clusters. The original data matrix is shown, partitioned according to the discovered clusters. From Figure 3 of (Kemp et al. 2006). Used with kind permission of Charles Kemp.

Practical Challenges with Clustering

- Should features/observations be standardized first?
- What measure of distance/similarity should be used?
- If we are setting number of components, how do we determine what number to use?
- How do you validate clustering in validation approaches? BIC, Intercluster variability, Intracluster variability
- Strict clustering vs. soft clustering?
- Use different techniques/parameters - see how consistent your results are!

Takeaways and Next Time

- Hierarchical Clustering
- Next Time: Neural Networks