DSC5103 Statistics

Session 10. Unsupervised Learning

Last time

- Boosting
  - Sequentially building models to fix the error of previous models

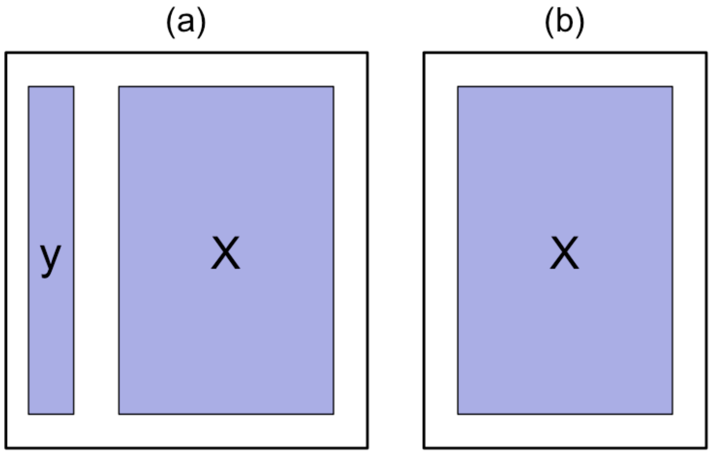
  - The tuning process

Plan for today

- Unsupervised Learning
  - Clustering Methods
    - K-mean clustering
    - Gaussian mixture model
    - Hierarchical clustering
  - Principle Components Analysis

Supervised vs. Unsupervised Learning

- Supervised Learning: finding the relationship between X and Y: Y = f(X) + ε



- Unsupervised Learning: where there is only X, no Y
  - Clustering: group data points with similar characteristics (X), define labels (Y)
  - PCA: combine predictors to reduce redundancy

- Challenges:
  - No obvious criterion to evaluate the output

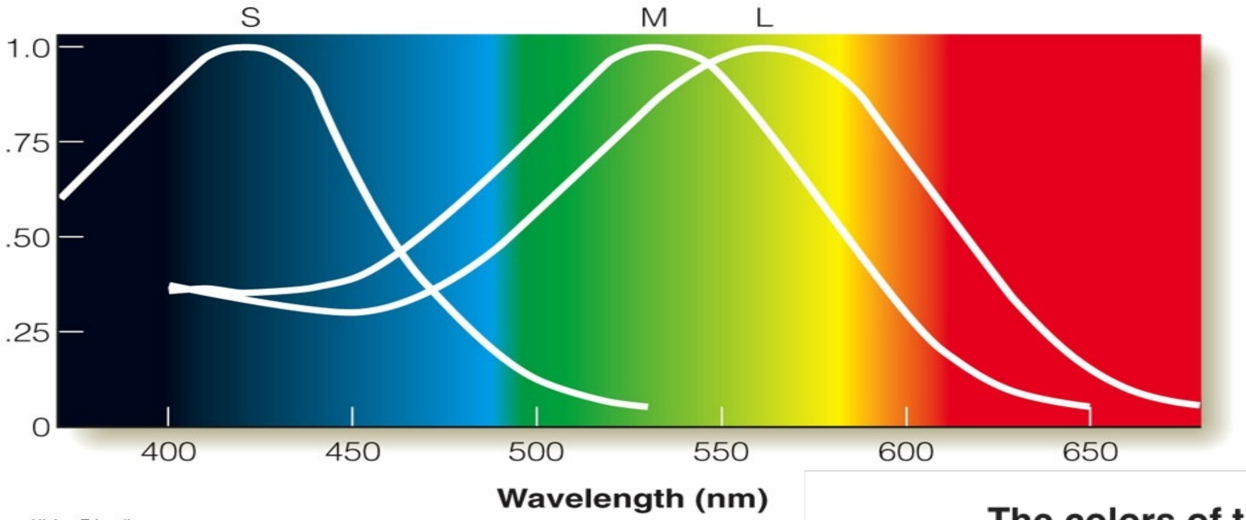
Clustering

- Clustering refers to a set of techniques for finding subgroups, or clusters, in a data set.

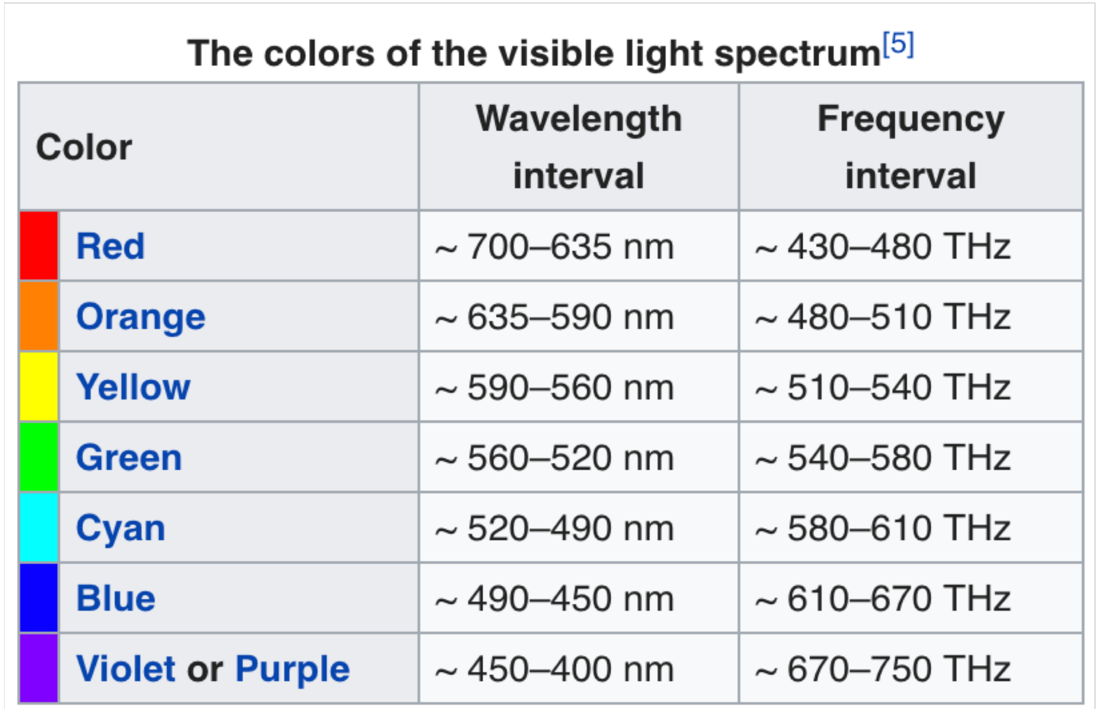
- A good clustering is one when the observations within a group are **similar** but between groups are very **different**.

- How to define “similar” and “different” is often context dependent…

Color Categorization



mson Higher Education



The colors of the visible light spectrum[5]

Color	Wavelength interval	Frequency interval
🟥 Red	~ 700–635 nm	~ 430–480 THz
🟧 Orange	~ 635–590 nm	~ 480–510 THz
🟨 Yellow	~ 590–560 nm	~ 510–540 THz
🟩 Green	~ 560–520 nm	~ 540–580 THz
🟦 Cyan	~ 520–490 nm	~ 580–610 THz
🟦 Blue	~ 490–450 nm	~ 610–670 THz
🟪 Violet or Purple	~ 450–400 nm	~ 670–750 THz

http://h3stogram.herokuapp.com/

Examples of Clustering

- Cluster customers into segments with similar profile
- Cluster products with similar characteristics
- Cluster shops with similar demand pattern
- Cluster patients with similar gene expressions

Different Clustering Methods

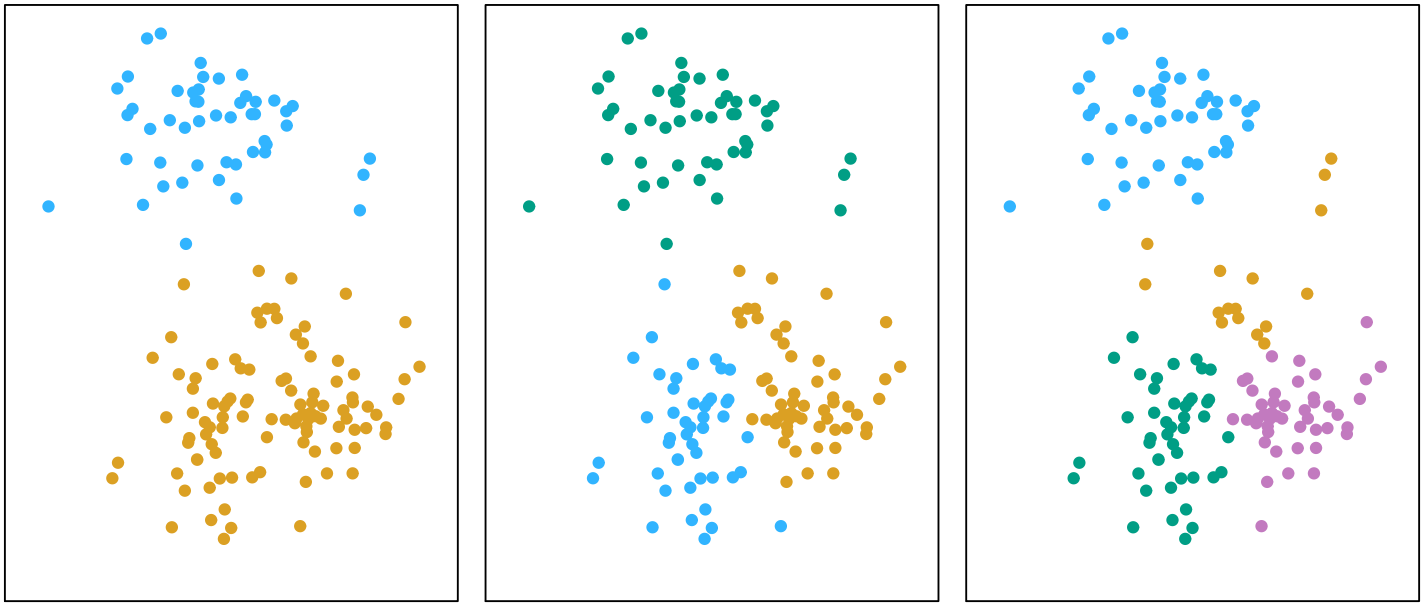
- K-means Clustering
  - Partition the data into a **pre-specified** number (K) of clusters

- Gaussian Mixture Models
  - Generalization of K-means

- Hierarchical Clustering
  - Clustering representation for each possible number of clusters

K-Means Clustering

- One must first specify the desired number of clusters K
- The K-means algorithm will assign each observation to exactly one of the K clusters



K=2

K=3

K=4

  - Note that there is no ordering of the clusters, so the cluster coloring is arbitrary

How does K-Means work?

- We would like to partition that data set into K clusters $C_1, \ldots, C_K$
  - Each observation belongs to one and only one of the K clusters

- The objective is to have a minimal “within-cluster-variation”, i.e. the elements within a cluster should be as similar as possible

- One way of achieving this is to minimize the total sum of the Euclidean distances to its cluster mean:

$$\min_{C_1,\ldots,C_K} \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - \bar{x}_k||^2$$

  - Note: it is decreasing in K.

K-Means Algorithm

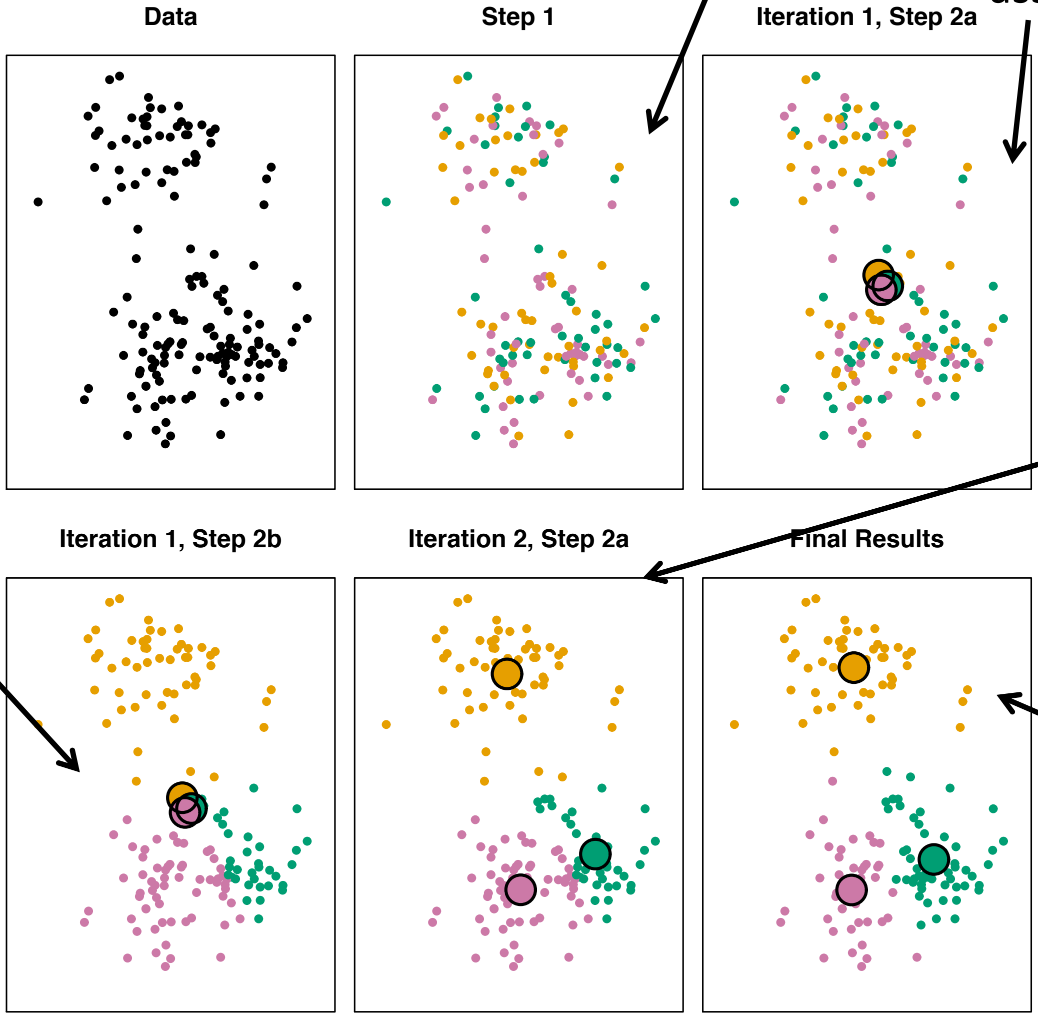
1. Randomly assign each observation to one of K clusters

2. Iterate until the cluster assignments stop changing:
   - Center: for each of the K clusters, compute the cluster centroid (mean).

   - Classify: assign each observation to the cluster whose centroid is closest (in terms of Euclidean distance)

3. Stop when no further changes

An Illustration of the K-Means Algorithm



Random Assignment of points

Compute cluster centers from initial assignments

Computer new cluster centers

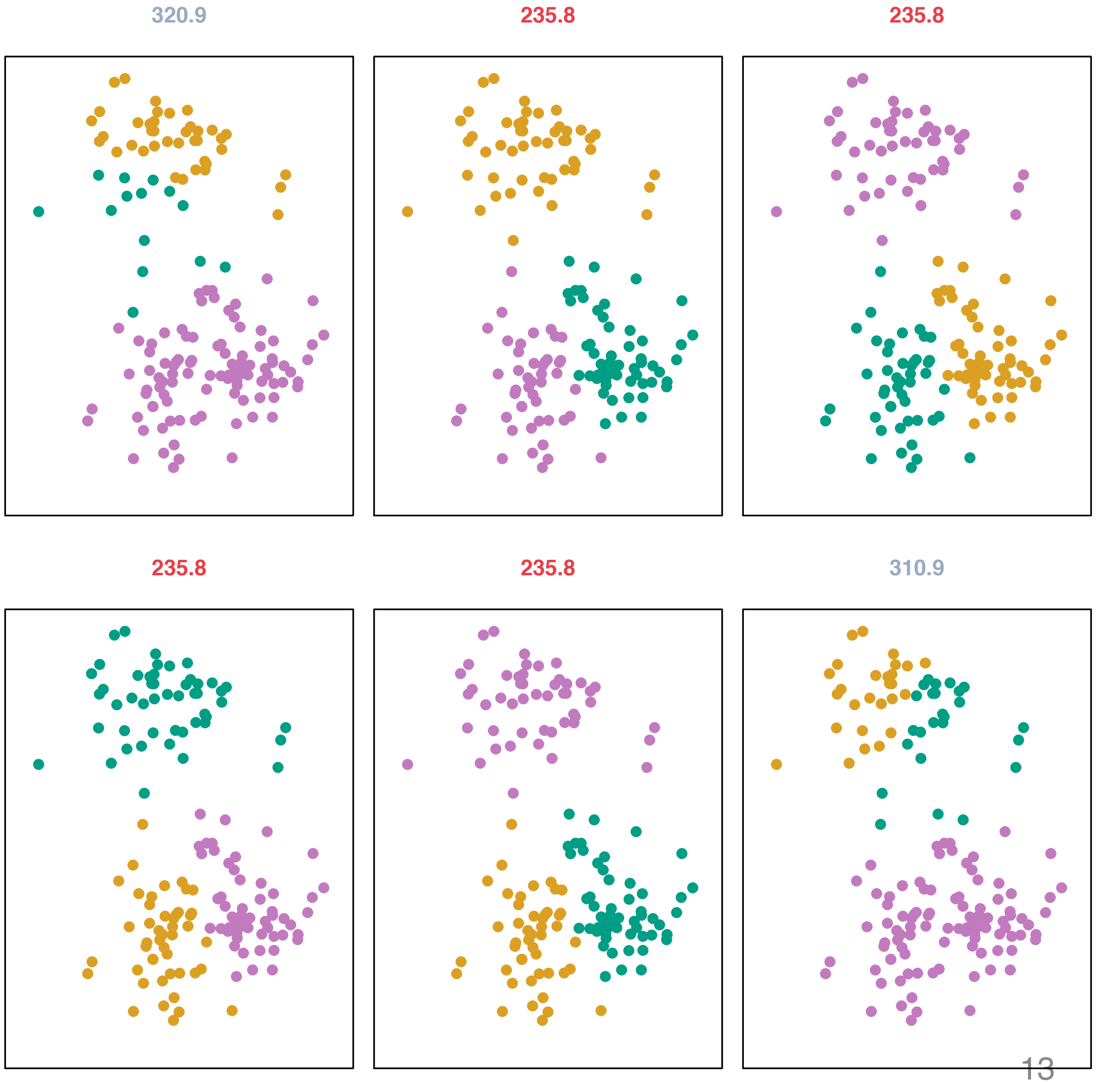
Assign points to closest cluster center

Now there is no further change so stop

Local Optimums

- The K-means algorithm can get stuck in “local optimum” and not find the best solution

- Hence, it is important to run the algorithm with multiple starting points to find a good solution

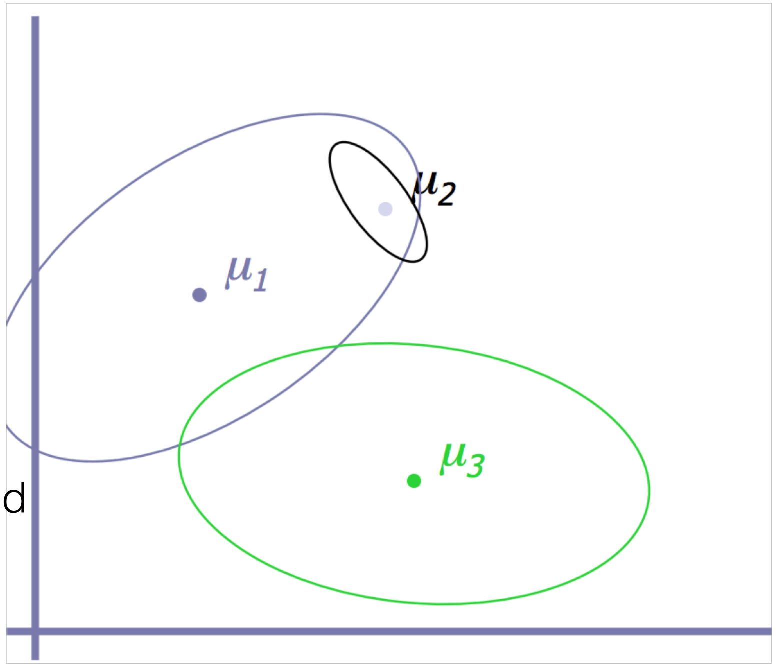


Gaussian Mixture Model (GMM)

- Data generating model:
  - K clusters
  - K Gaussian distribution on the predictor space $N(\mu_k, \Sigma_k)$
    - μ_k determines the center
    - Σ_k determines the shape of the ellipsoid
  - Mixture probabilities p_k: probability of a data is born in cluster k

- Estimation by the E-M algorithm
  - E-step: Assign data to clusters (classify)
  - M-step: Parameters for the clusters μ_k, Σ_k

- Generalization of K-means
  - Ellipsoid instead of spheres
  - Soft (Probabilistic) clustering instead of hard



Hierarchical Clustering

- K-Means/GMM clustering requires choosing the number of clusters K.

- Hierarchical Clustering
  - It does not require to commit to a particular K

  - Bottom-up (Agglomerative) clustering: start from the leaves and combine clusters up to the trunk

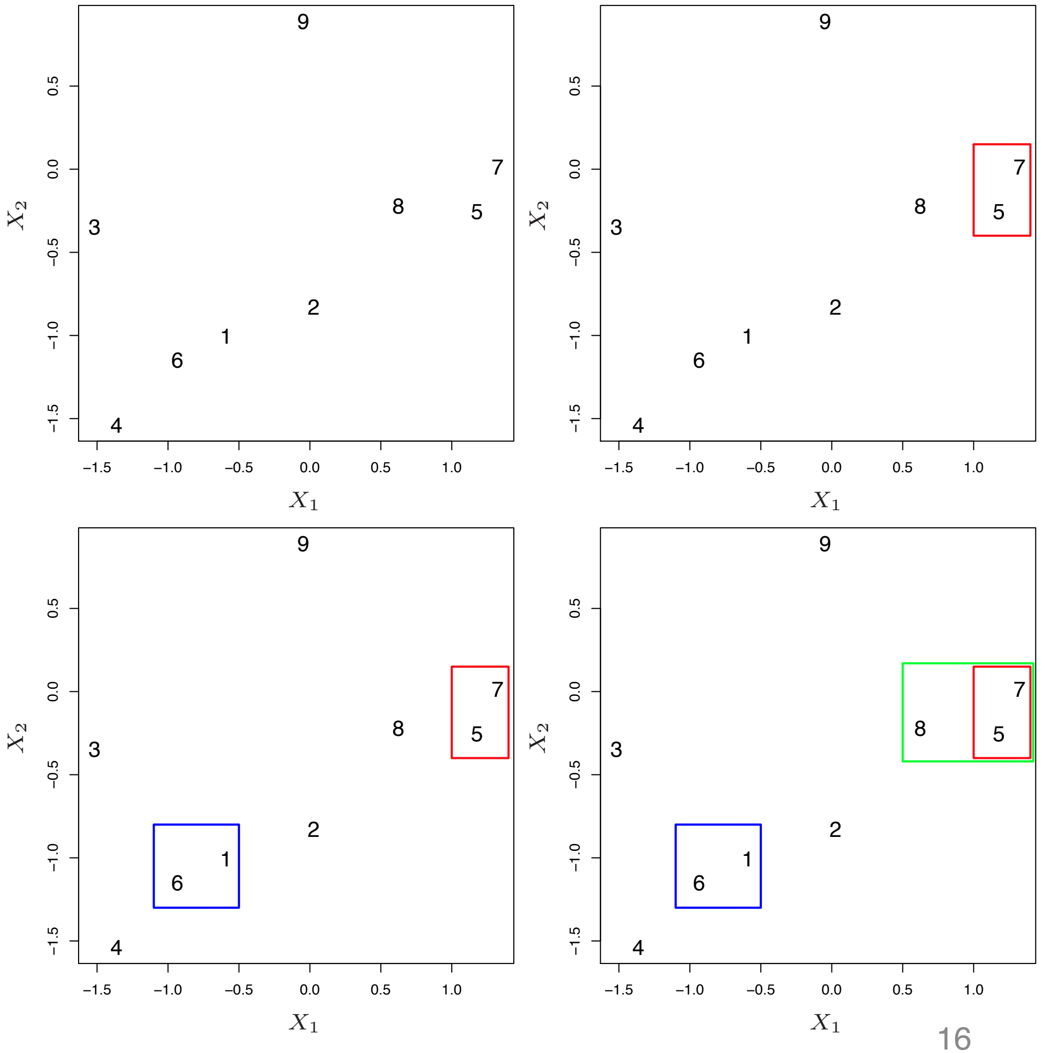
  - It produces a tree based representation of the observations, called a Dendogram

An Example

- Start with 9 points as 9 clusters

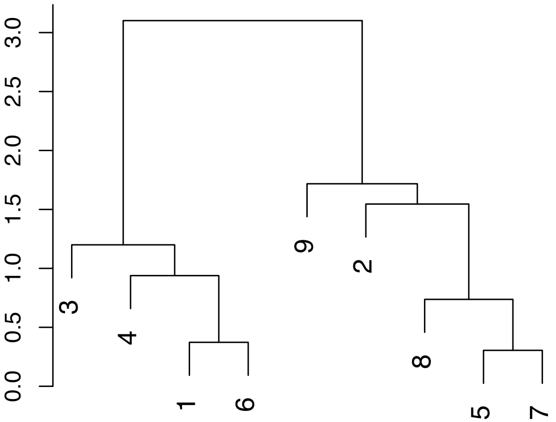
- Sequentially identify the closest two clusters and merge them
  - Merge 5 and 7
  - Merge 6 and 1
  - Merge the (5,7) cluster with 8
  - …

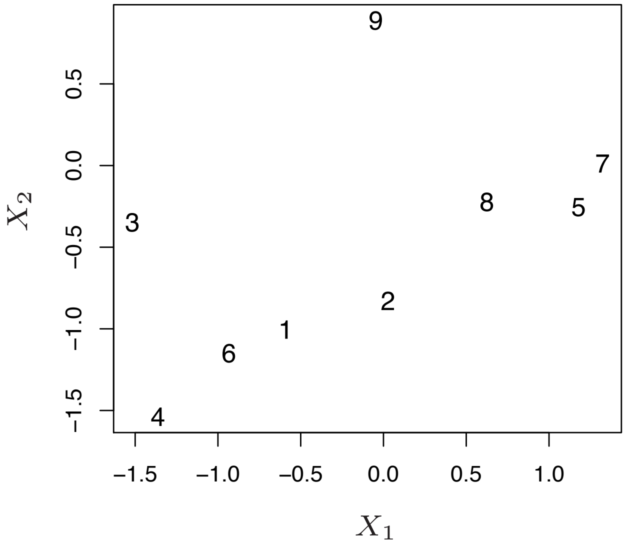
- Continue until all observations are merged in a single cluster



Dendograms

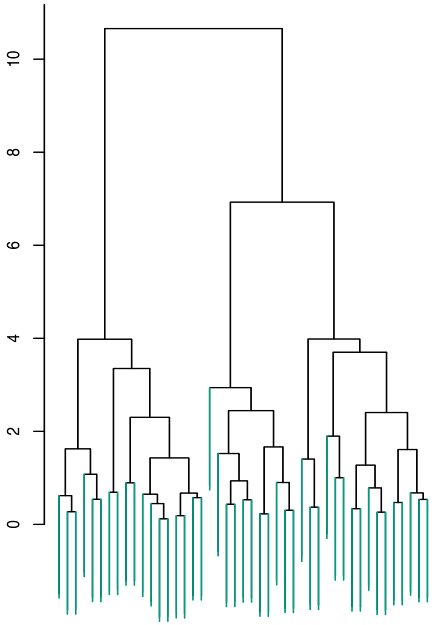
- A tree representation of the sequential merging
- Height on vertical axis indicates how similar the points are



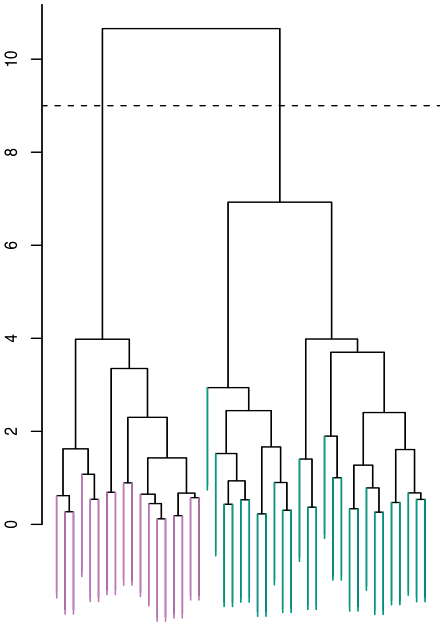


Choosing Clusters

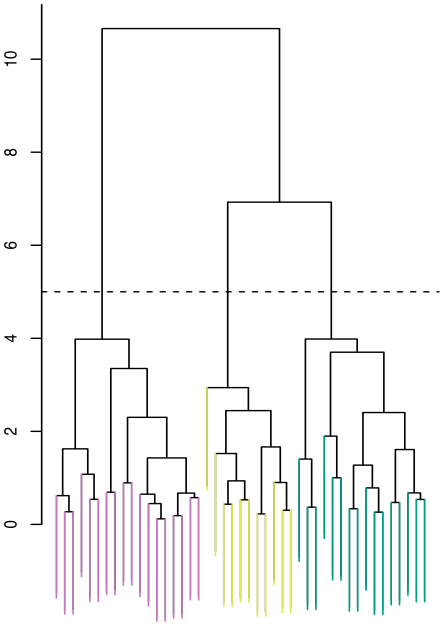
- To choose clusters we draw lines across the dendogram
- We can form any number of clusters depending on where we draw the break point



One Cluster



Two Clusters



Three Clusters

Algorithm (Agglomerative Approach)

1. Start with each point as a separate cluster (n clusters)

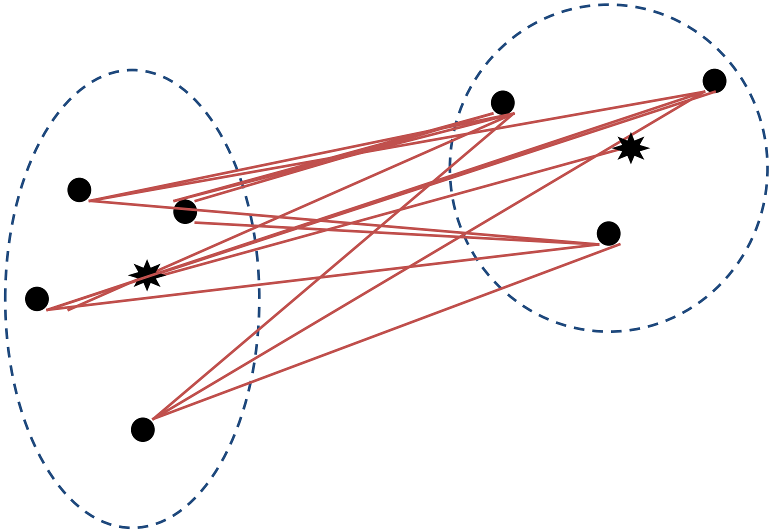
2. Repeat the following until there is only one cluster left

   – Calculate a measure of dissimilarity between all clusters

   – Merge two clusters that are most similar

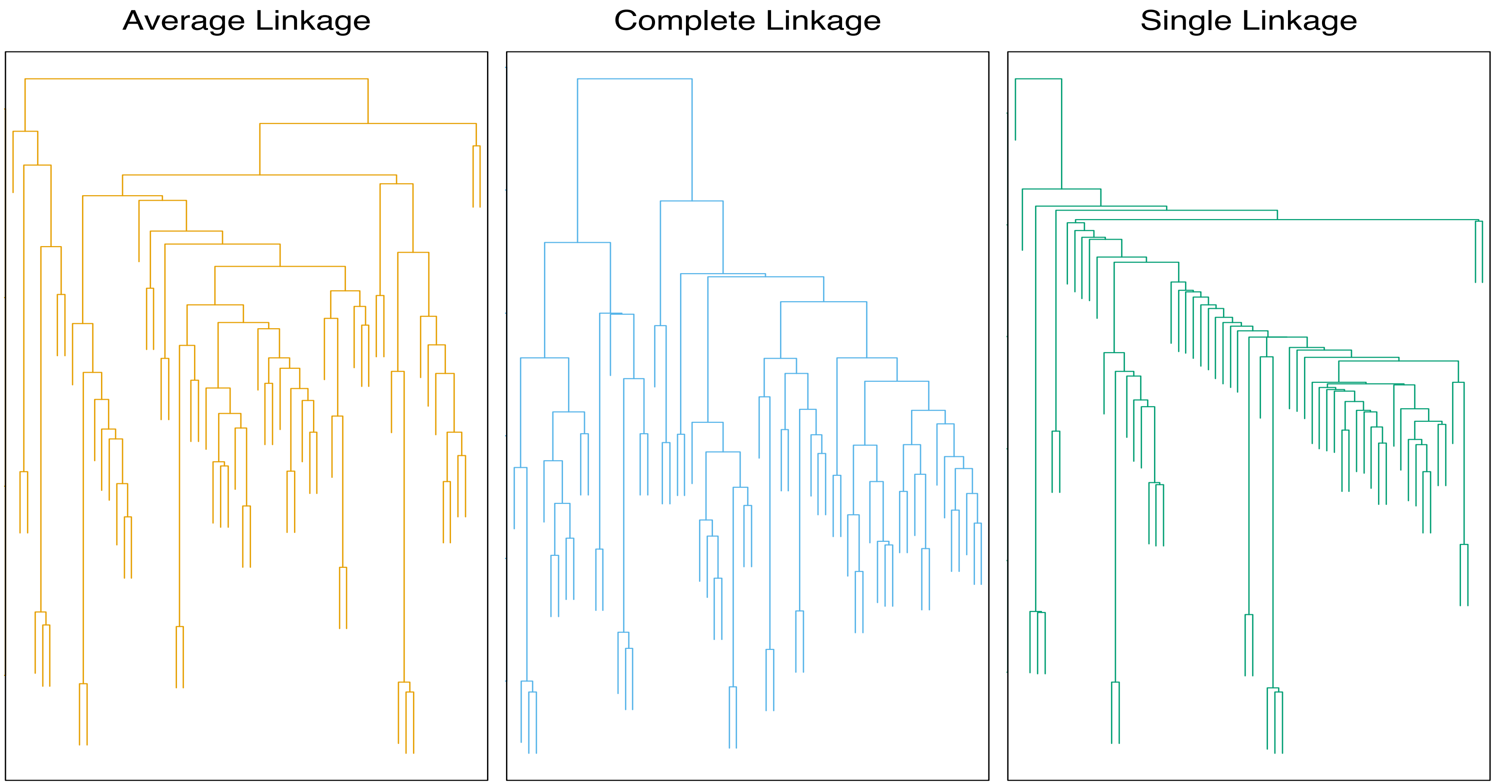
Linkage Methods: Distance Between Clusters

- How do we define the dissimilarity (linkage) between two clusters?
  - Complete Linkage: Largest inter-cluster pairwise distance
  - Single Linkage: Smallest inter-cluster pairwise distance
  - Average Linkage: Average inter-cluster pairwise distance
  - Centroid: distance between centroids of the two clusters



Linkage Can be Important

- Complete and average linkage tend to yield evenly sized clusters whereas single linkage tends to yield extended clusters to which single leaves are merged one by one.

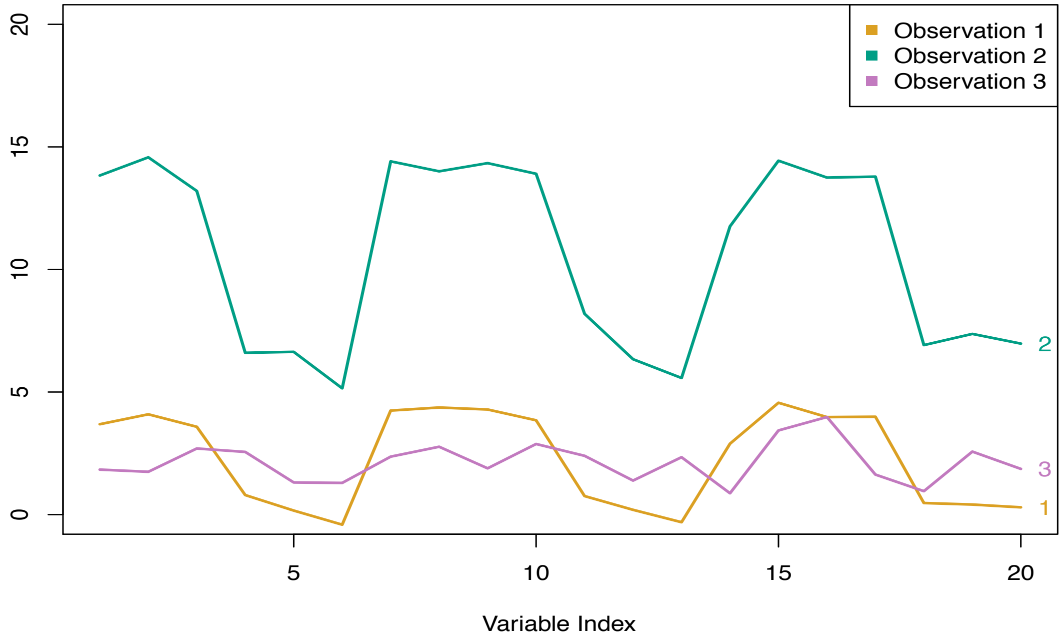


Choice of Dissimilarity Measure

- Distance between two points
  - Euclidean distance is the common dissimilarity measure
  - An alternative measure that could make sense in some cases is the correlation based distance
    - Considers two observations to be similar if their features are highly correlated
    - Ignore the difference in magnitude

Comparing Dissimilarity Measures

- n=3 observations and p=20 variables
- In terms of Euclidean distance, observation 1 and 3 are similar
- Observation 1 and 2 are highly correlated so would be considered similar in terms of correlation measure



Online Shopping Example

- Suppose we record the number of purchases of each item (columns) for each customer (rows)

- Using Euclidean distance, customers who have purchases very little will be clustered together

- Using correlation measure, customers who tend to purchase the same types of products will be clustered together even if the magnitude of their purchase may be quite different

Practical Issues in Clustering

- Should the features first be standardized? i.e. Have the variables centered to have a mean of zero and standard deviation of one.

- In case of K-means/GMM clustering:
  - How many clusters should we look for the data?

- In case of hierarchical clustering:
  - What dissimilarity measure should be used?
  - What type of linkage should be used?
  - Where should we cut the dendogram in order to obtain clusters?

- How to incorporate categorical variables?!

- There is no single right answer!

Principle Components Analysis

- PCA produces a low-dimensional representation of the data
  - Pre-processing for supervised learning
  - Low-dimensional visualization

- Idea: decompose the data X into a sequence of principle components that are
  - linear combinations of the variables
  - of maximal variance
  - mutually uncorrelated

Principal Components Analysis: details

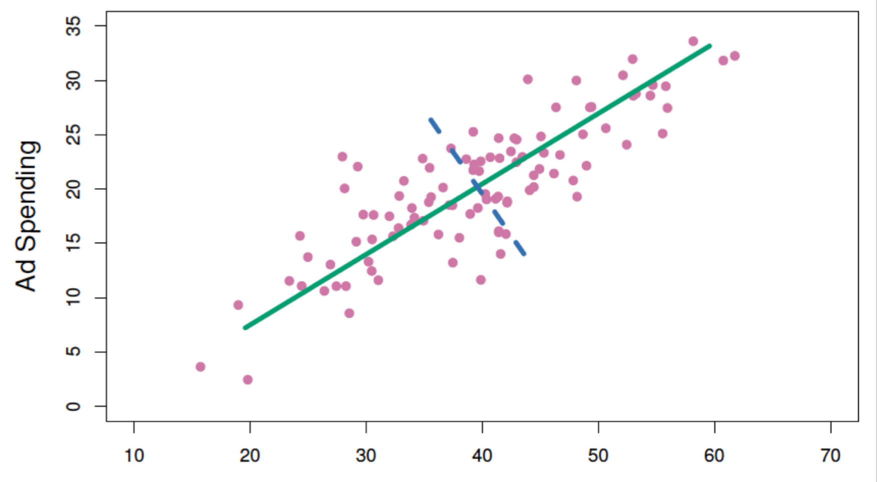
- The first principle component Z_1 is a normalized linear combination of X

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p$$

- where $\phi_1 = (\phi_{11}, \ldots, \phi_{p1})$ is the loading of Z_1 and satisfies

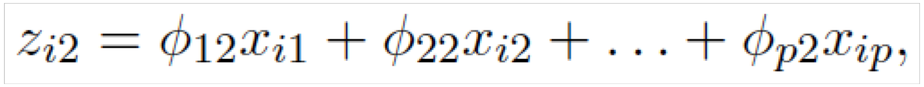
$$\sum_{j=1}^{p} \phi_{j1}^2 = 1$$

- Loading ϕ_1 is found by maximize Var[Z_1] via Singular Value Decomposition (SVD)
- ϕ_1 defines a direction in feature space along which the data vary the most



Principal Components Analysis: details

- The second principal component is the linear combination of X that has maximal variance among all linear combinations that are **uncorrelated** with Z_1



$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \ldots + \phi_{p2}x_{ip},$$

- ϕ_2 is the loading vector of Z_2 and is **orthogonal** to ϕ_1.

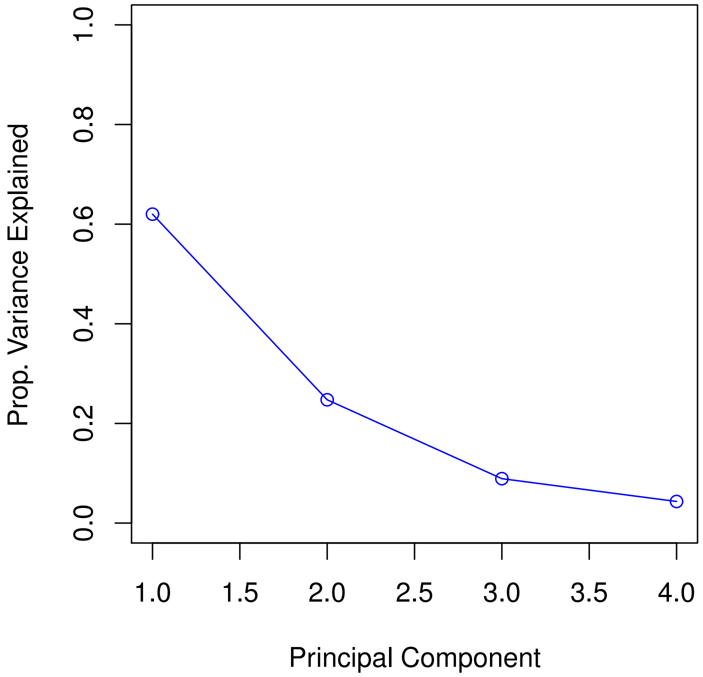
- The principal component directions ϕ_1, ϕ_2, ... are the ordered sequence of right singular vectors of the matrix X, which can be obtained by SVD.

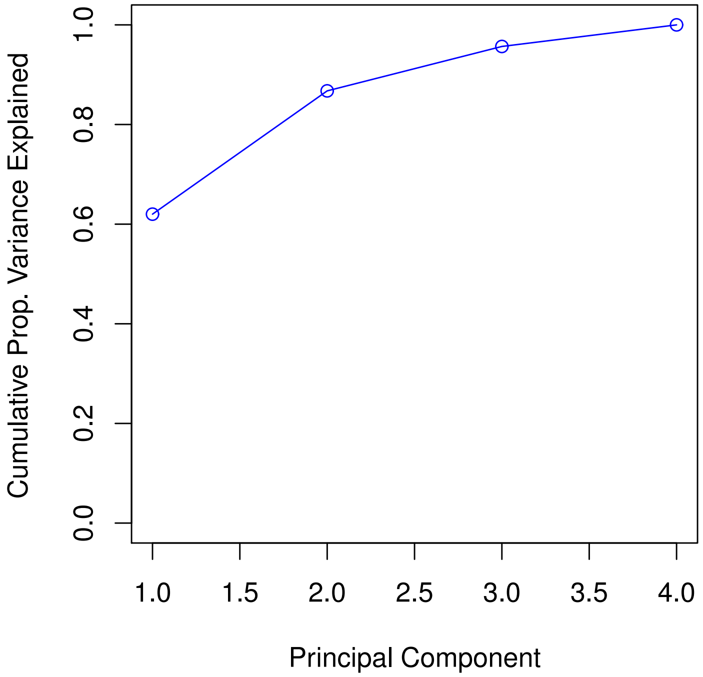
- There are at most min(n – 1, p) principal components.

Proportion of Variance Explained

- The PVE of the m-th principal component is given by the positive quantity between 0 and 1

$$\frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}.$$





Final Remarks

- Unsupervised learning is important for understanding unlabeled data
- Unsupervised learning can be a useful pre-processor for supervised learning
- It is intrinsically more difficult than supervised learning because there is no gold standard or single objective