CSCI4622 Machine Learning

2020 Spring

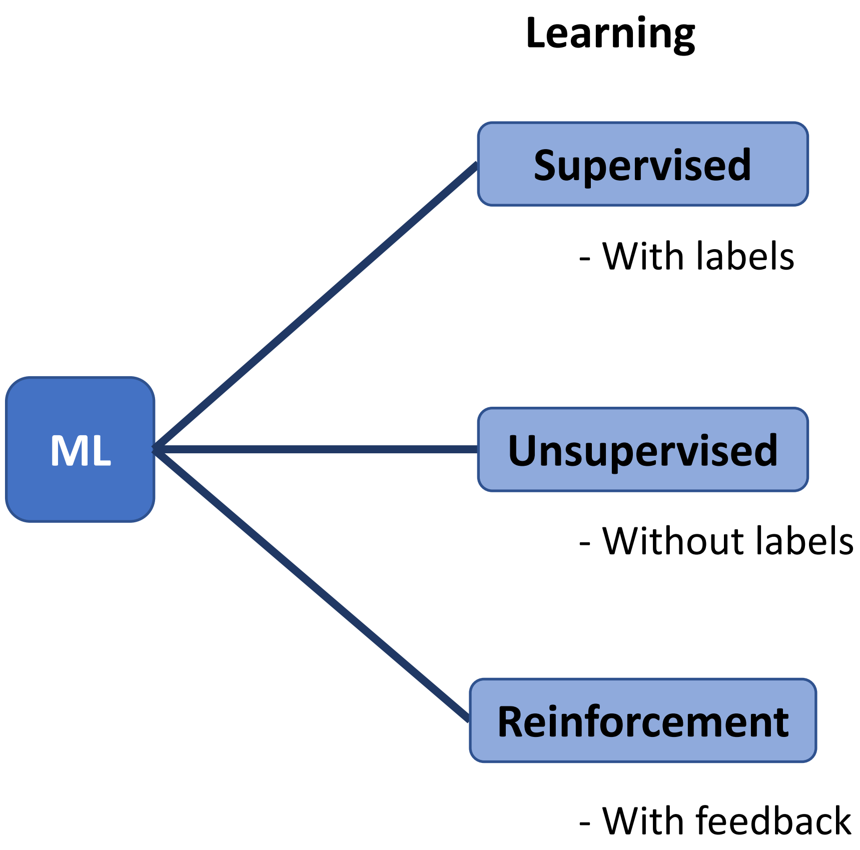
Unsupervised Learning

Geena Kim

Unsupervised Learning

*Slide contents adopted from ISLR material

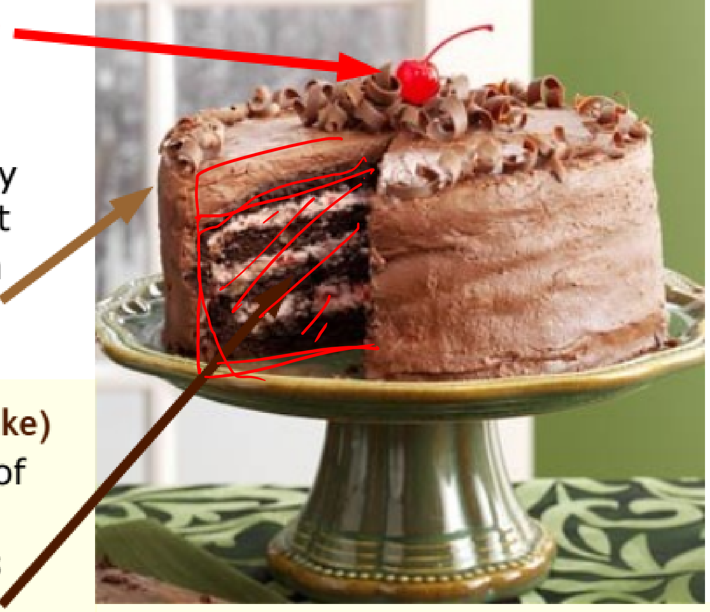
Types of machine learning problems



Yann LeCun says about Unsupervised Learning…

in terms of data availability

- "Pure" Reinforcement Learning (cherry)
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- Supervised Learning (icing)
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- Unsupervised/Predictive Learning (cake)
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**
- (Yes, I know, this picture is slightly offensive to RL folks. But I'll make it up)



Goals of Unsupervised Learning

Not interested in prediction but to discover interesting things about the data

Informative visualization

Finding subgroups ← Clustering

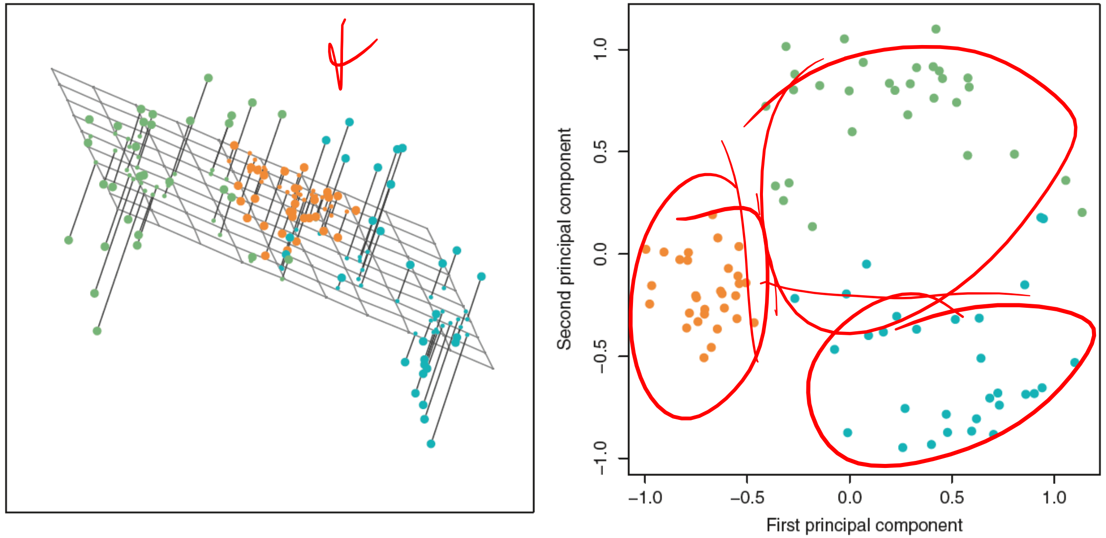
Dimensionality Reduction ✓

Preprocessing

Data synthesis

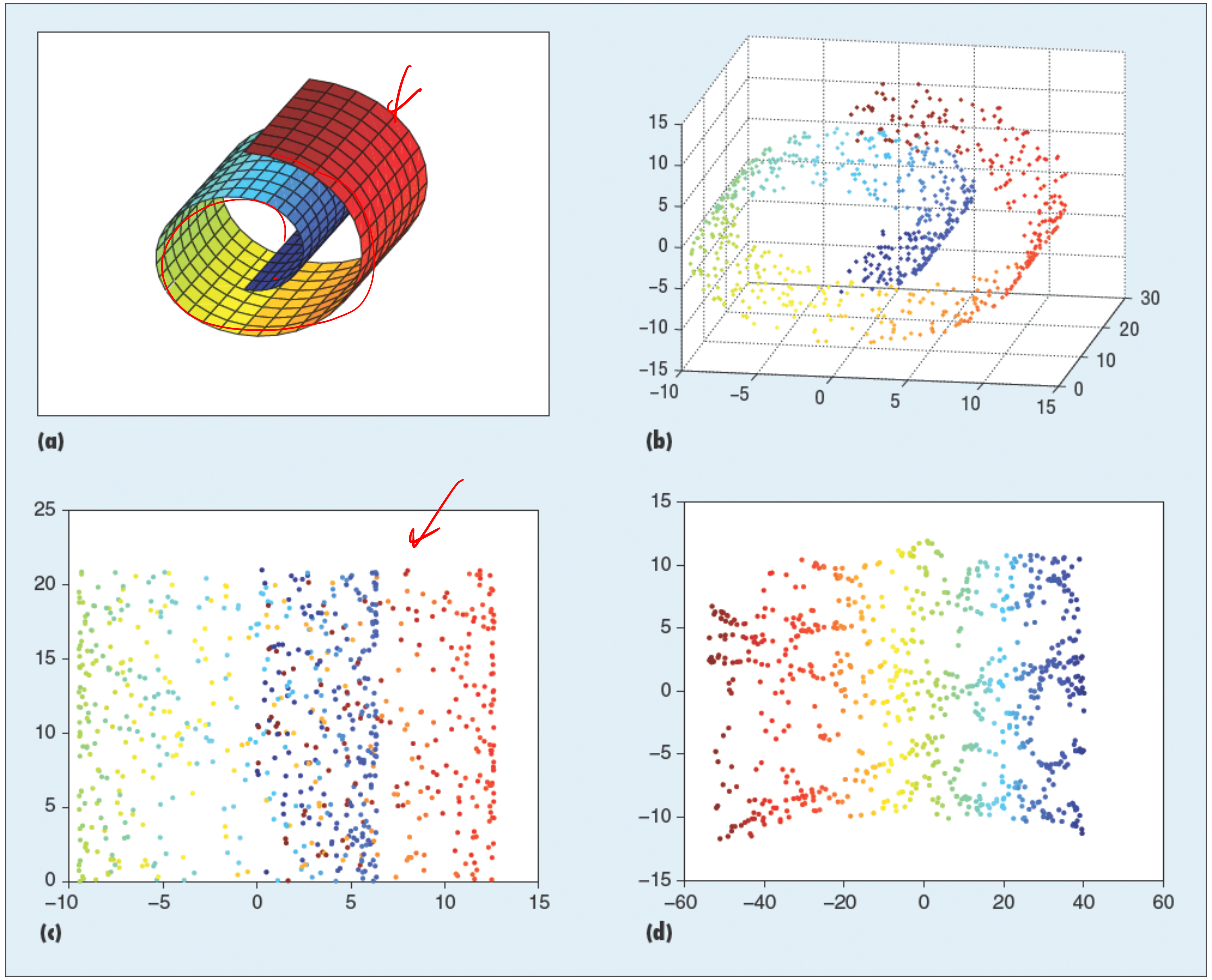
Dimensionality Reduction

Projection to low-dimension



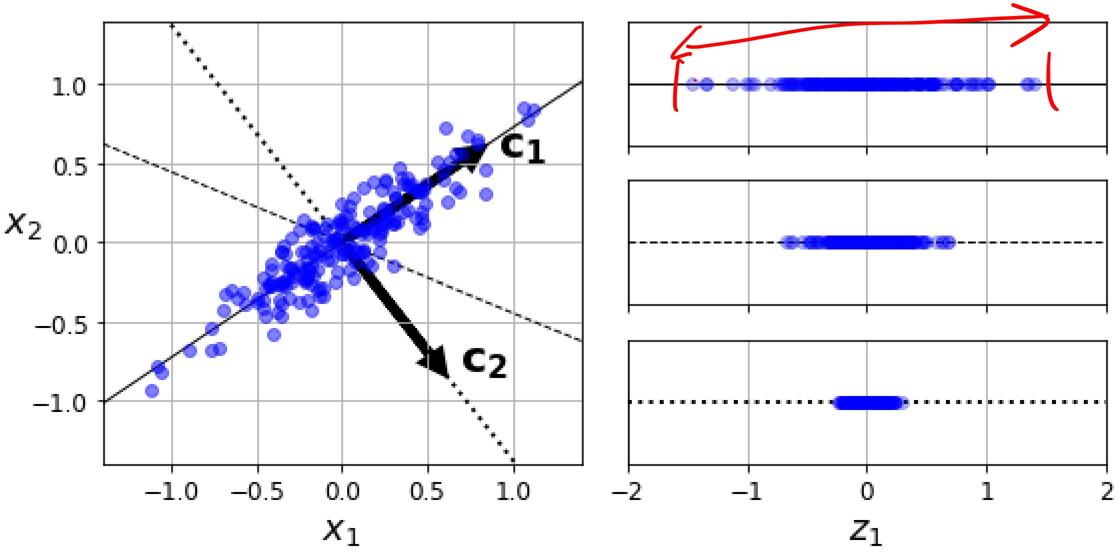
Manifold learning

t SNE



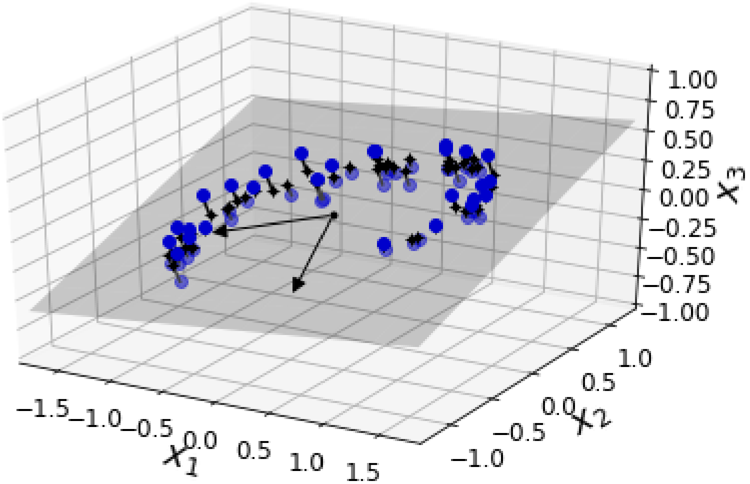
Principal Component Analysis (PCA)

How to choose the principal components?



Method 1. Preserve the maximum variance

Method 2. Choose axis that minimize the mean squared distance between the original dataset and its projection onto the axis



Principal Component Analysis (PCA)

$$\begin{pmatrix} \lambda_1 & & & & \\ & \lambda_2 & & & \\ & & \lambda_3 & & \\ & & & \ddots & \\ & & & & \lambda_n \end{pmatrix}$$

Facts about $C^X = \frac{1}{m-1} X^T X$

- Symmetric
- All eigenvalues are real (b/c symmetric)
- All eigenvalues are nonnegative (because it is positive semidefinite)
- C^X has ~~D~~ n mutually orthogonal eigenvectors (which can be scaled to unit length)

$$C^X = \sum_{i=1}^{D} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^T ,$$

where λ_i are the eigenvalues $(\lambda_1 \geq \lambda_2 \geq \ldots, \geq \lambda_D)$, $\boldsymbol{v}_i$ is the eigenvector associated with λ_i.

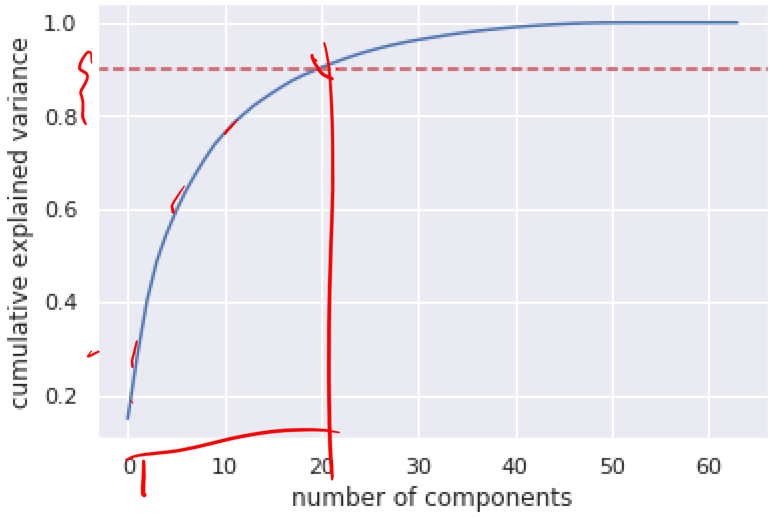
Principal Component Analysis (PCA)

How many dimensions should we choose to use?

$$= \frac{V_j}{V(j=1 \cdots N)}$$

1 , 2 , . . .

"elbow" plot



1

What is explained variance?

What is explained variance ratio?

x%

0,0000001

$$f_1 = \phi_1 x_1 + \phi_2 x_2 \cdots \phi_n \cancel{x_n}$$

$$1 = \sum \phi^2$$

Clustering

- PCA: looks for a low-dimensional representation of observations that are useful

- Clustering: looks for homogeneous subgroups among observations

What Clustering if for

- Get a meaningful intuition of the structure of the

- Cluster-then-predict where different models will be built for different subgroups if we believe there is a wide variation in the behaviors of different subgroups.
- (ex) clustering patients into different subgroups and build a model for each subgroup to predict the probability of the risk of having certain disease; e.g. genetic data

Clustering applications

- Customer segmentation: identifying subgroups of people who might like to purchase particular types of products
- Advertising: identifying subgroups of people who might respond to particular types of advertising
- Document clustering: identifying documents that are similar (same applies to movies and music)
- Genetics study: identifying subgroups of disease type using gene expression data (e.g. type 1 and type 2 diabetes)
- Image segmentation or image compression, preprocessing

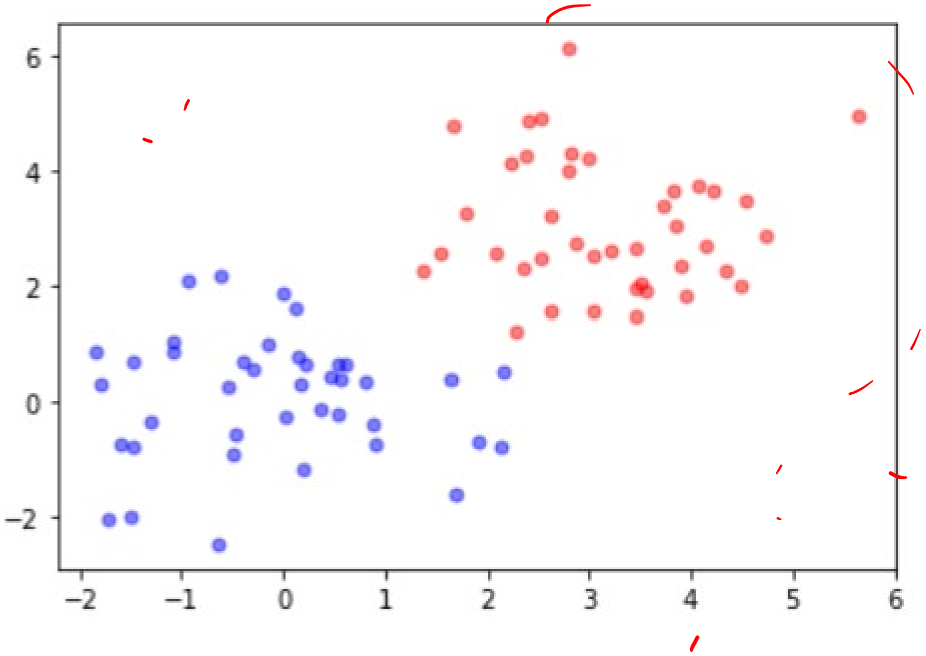
Popular Clustering Methods

- K-means clustering

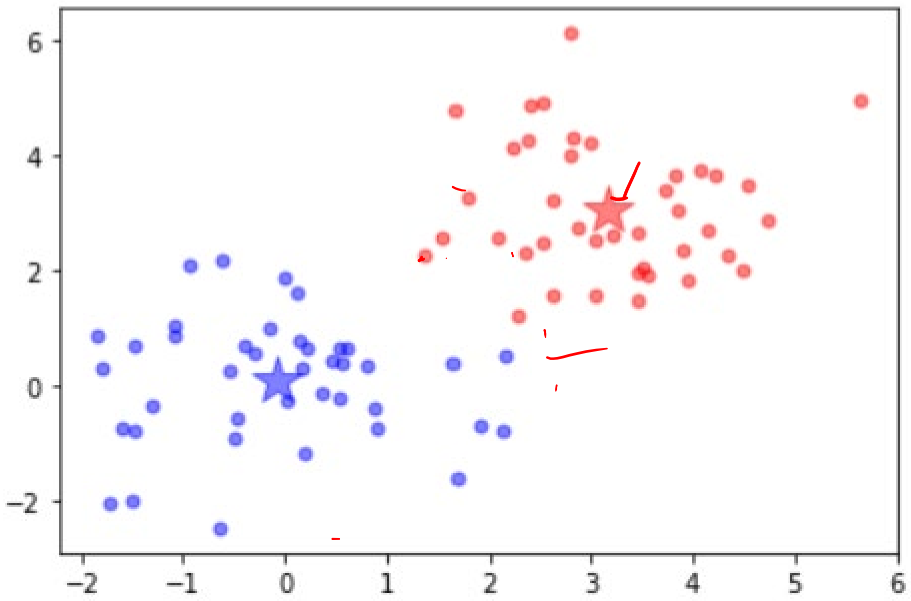
- hierarchical clustering

K-means Clustering

What is a cluster?



What is a centroid?



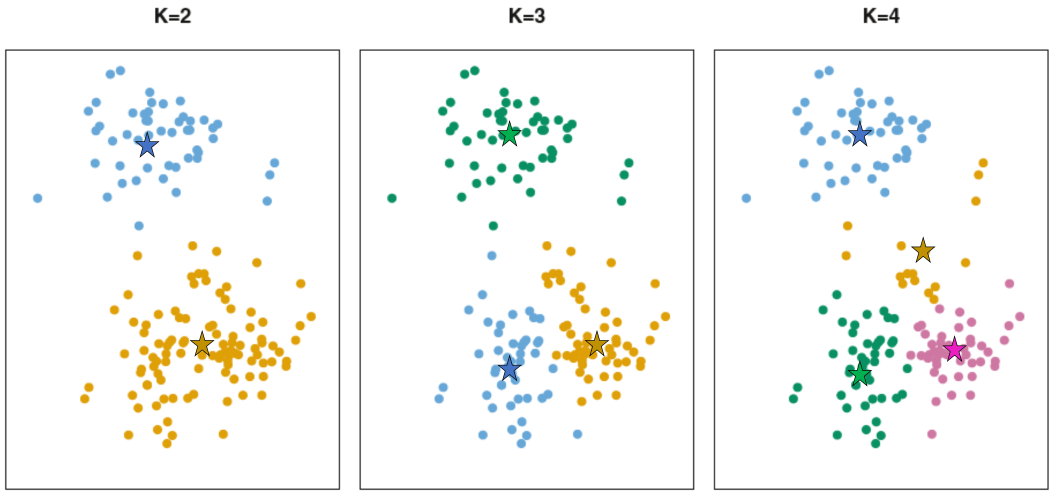
K-means Clustering

What is K-means clustering?

Cluster

Centroid

Euclidean distance



K-means Clustering

K-means Clustering Algorithm

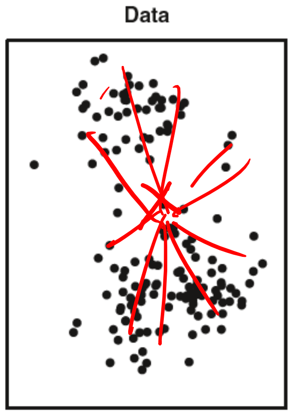
2, 3

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

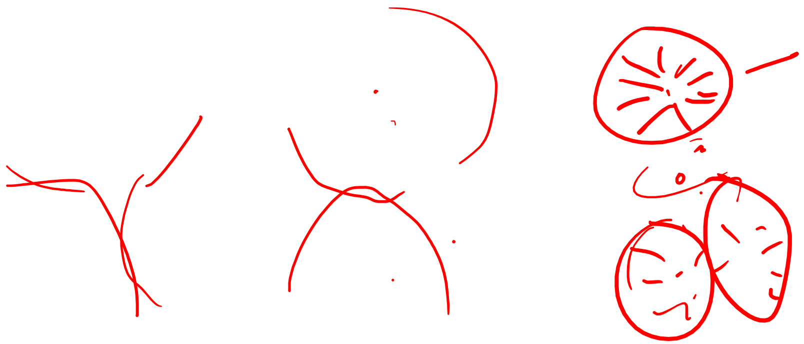
2. Iterate until the cluster assignments stop changing:

   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).



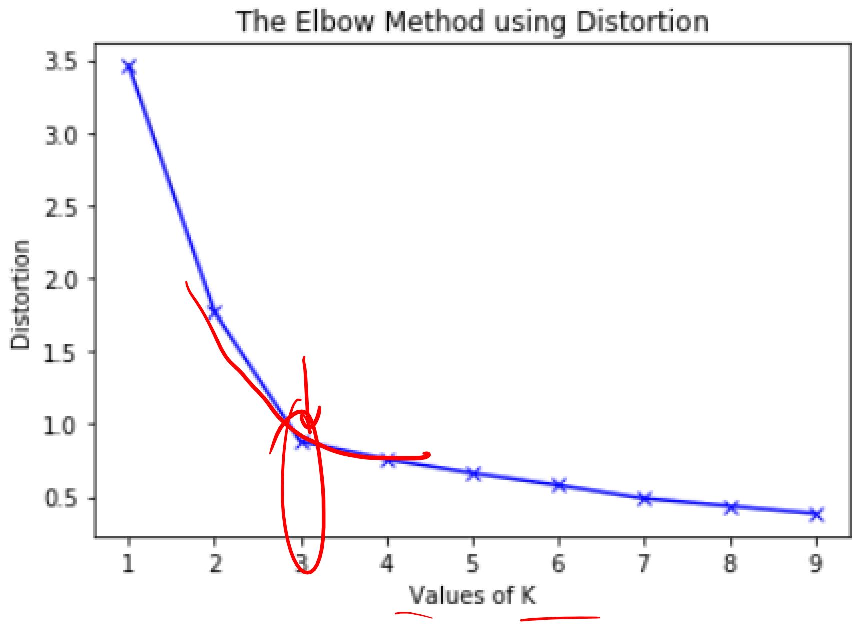
d

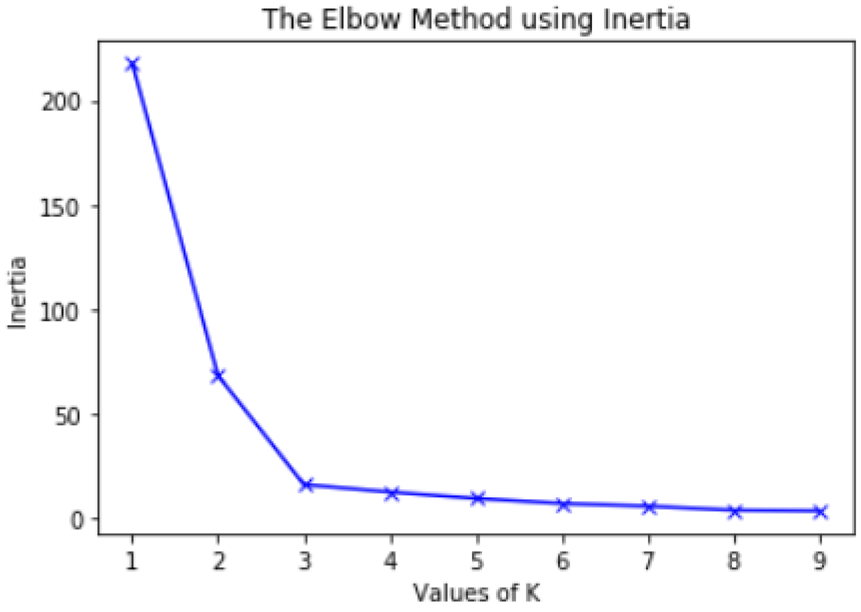


K-means Clustering

How to choose K?

Metric:
Distortion (the mean of square distance within a cluster)
Inertia (the sum of square distance within a cluster)





K-means Clustering

1 2 3 4 100

1 5 10 15 20 50 500

K-means Clustering

Need to decide how many clusters (K) before trying

Vulnerable to curse of dimensionality PCA preprocessing helps

"Given enough time," K-means will always converge

Finds local minimum, not global minimum

The local minimum is highly dependent on the initialization of the centroids

sklearn’s KMeans can initialize better if `init=‘k-means++’` is used

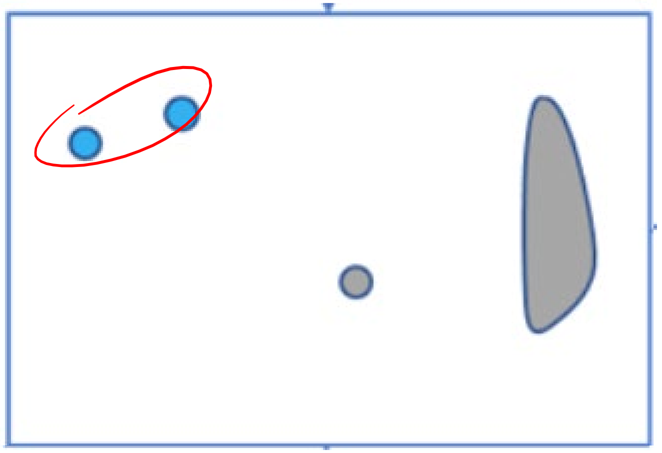
`MiniBatchKmeans` uses mini-batches to reduce the computation time

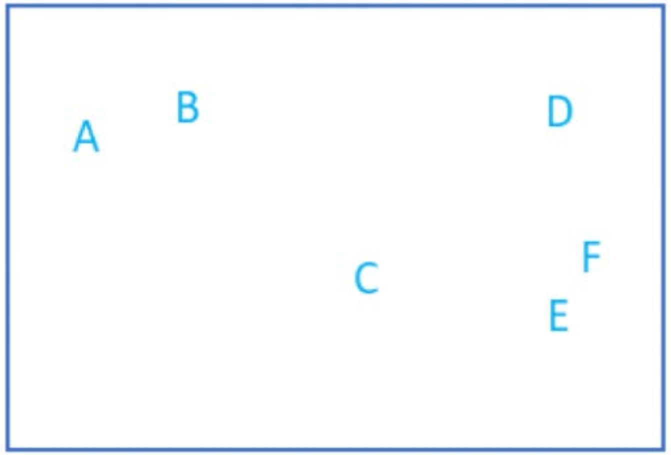
Hierarchical Clustering

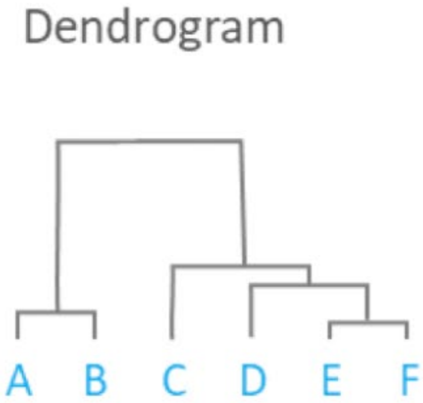
It does not need to know K in advance!

Dendrogram (upside down tree)

agglomerative hierarchical clustering



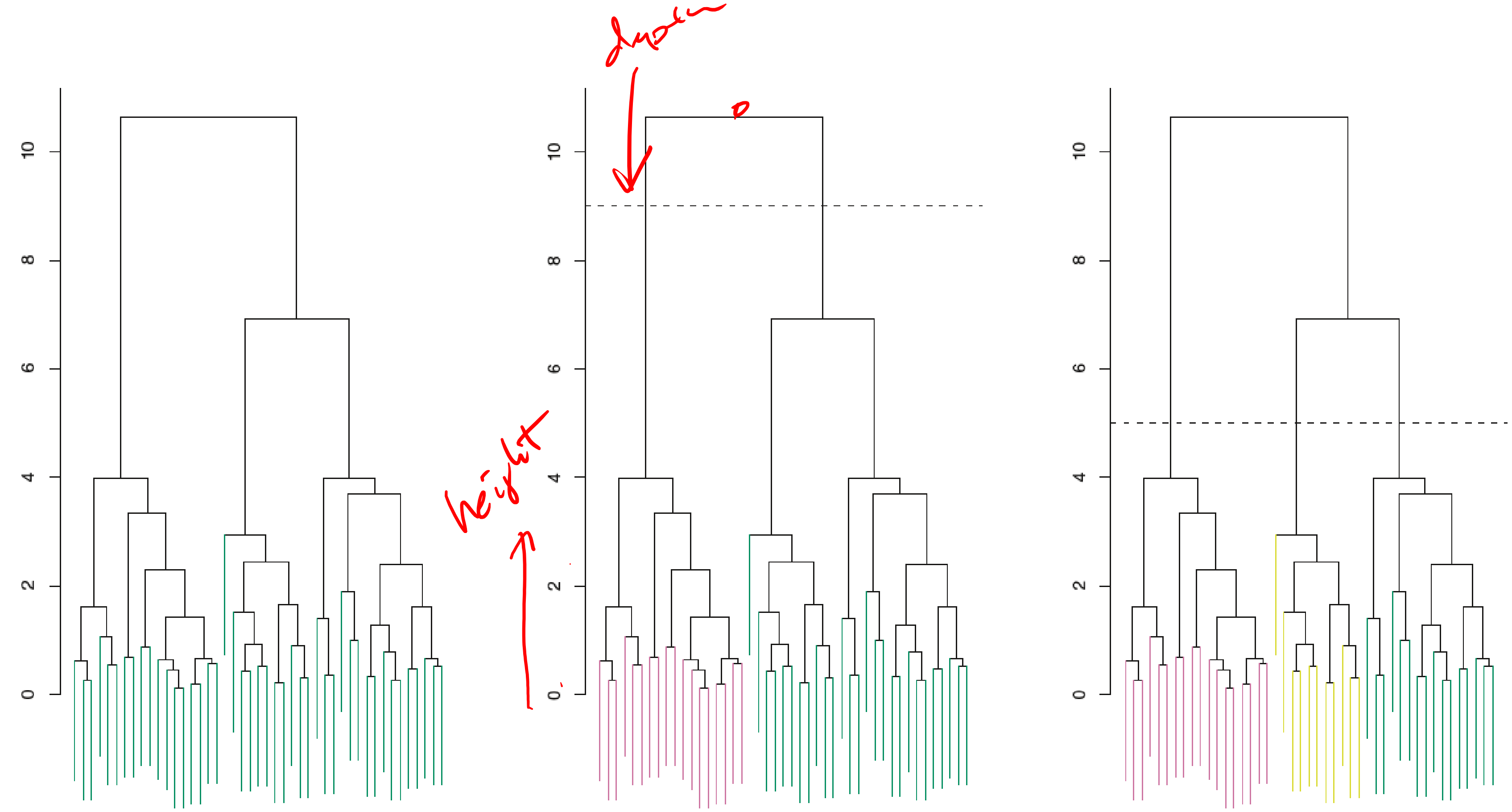




Distance: Euclidean, Correlation-based

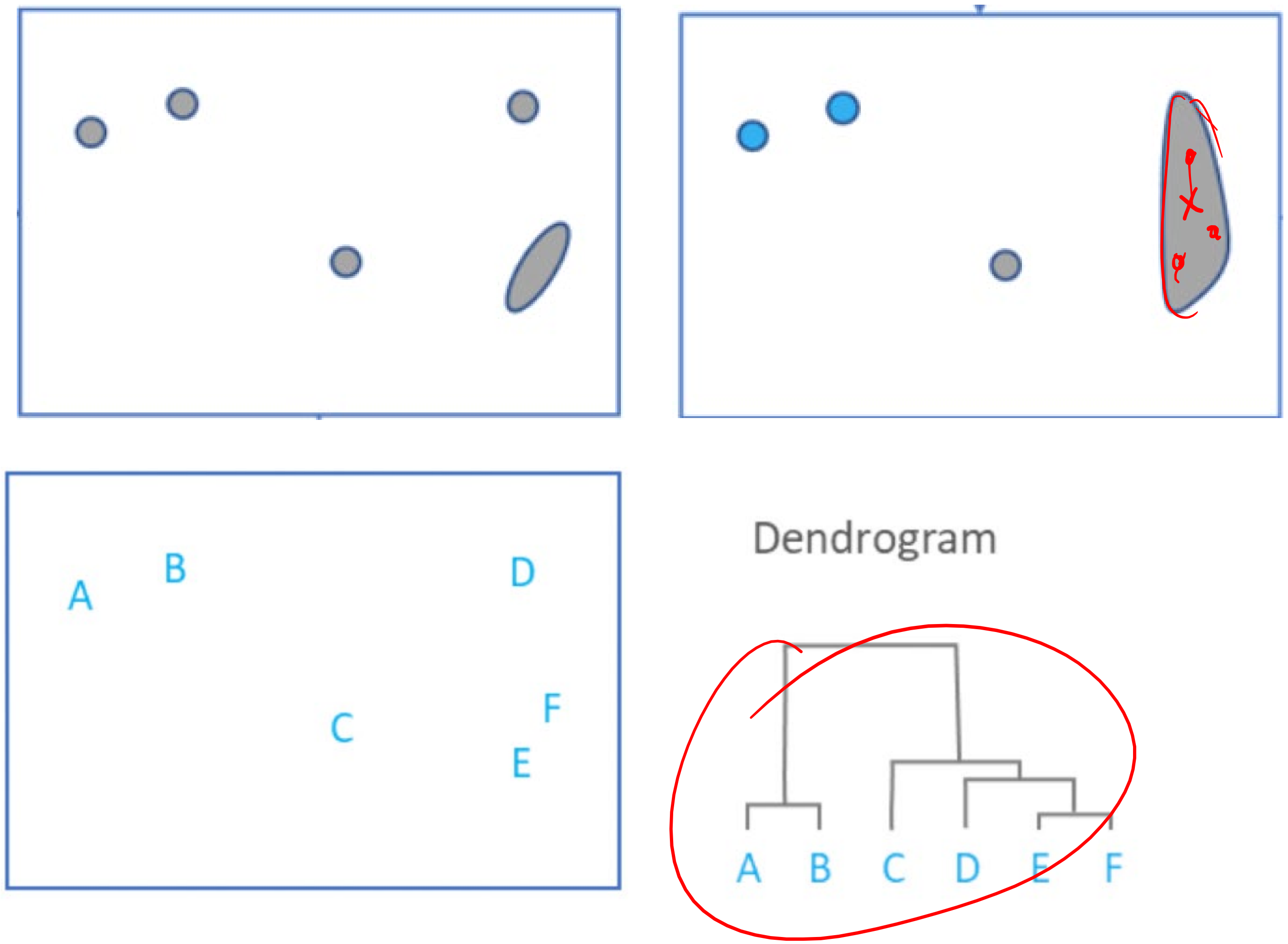
Hierarchical Clustering

Finding clusters from the dendrogram



Hierarchical Clustering

Choice of linkage type matters!



- Complete max
- Single → min
- Average → Avg
- Centroid

Effect of (dis)similarity metric choice

Choice of similarity metric is very important

Example: identifying subgroups of shoppers

Data-> 100 millions of shoppers (rows) and 500 millions of items

What happens if we use Euclidean distance?

What if we use correlation?

Effect of feature scaling

Features may have very different range of values

Consider shopping frequency of certain items
(e.g.) phone charging cable vs. laptop

The solution: standardize