Tutorial 4

Clustering

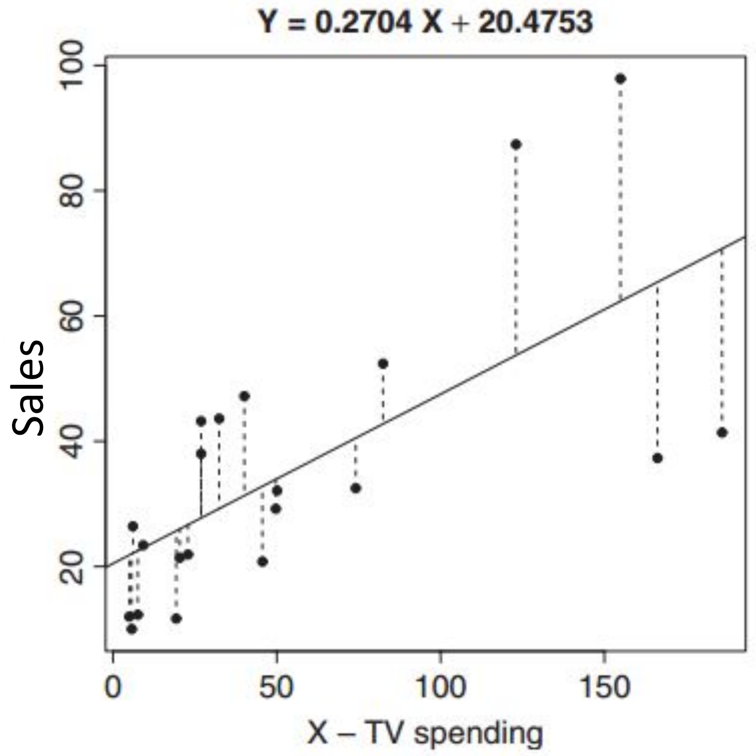
Supervised Learning vs. Unsupervised Learning

- Supervised Learning:
  - Linear Regression
  - Classification
    - Naïve Byesian Classifier
    - Decision trees (TDIDT)
    - Logistic Regression
    - Neural Networks
    - Support Vector Machine

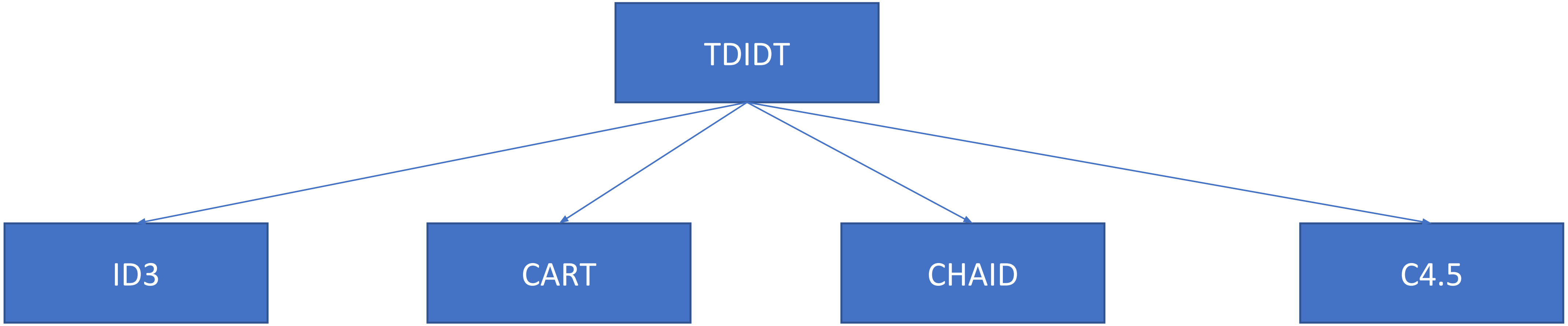
- Unsupervised Learning:
  - Clustering
  - K-Means
  - Hierarchical Clustering
  - Density-based Clustering

Regression (Recap)

Company	TV spending (M$)	Sales (M$)
MILLER.LITE	50.1	32.1
PEPSI	74.1	32.5
STROH'S	19.3	11.7
FEDERAL.EXPRESS	22.9	21.9
BURGER.KING	82.4	52.4
COCA-COLA	40.1	47.2
MC.DONALD'S	185.9	41.4
MCI	26.9	43.2
DIET.COLA	20.4	21.4
FORD	166.2	37.3
LEVI'S	123	87.4
BUD.LITE	45.6	20.8
ATT.BELL	154.9	97.9
CALVIN.KLEIN	5	12
WENDY'S	49.7	29.2
POLAROID	26.9	38
SHASTA	5.7	10
MEOW.MIX	7.6	12.3
OSCAR.MEYER	9.2	23.4
CREST	32.4	43.6
KIBBLES.N.BITS	6.1	26.4



Decision Tree



ID3

Iterative Dichotomiser 3

Use “Information Gain” to measure purity

CART

Classification And Regression Tree

Use “Gini Index” to measure purity

CHAID

CHi-squared Automatic Interaction Detector

Use “CHi-square” to measure purity

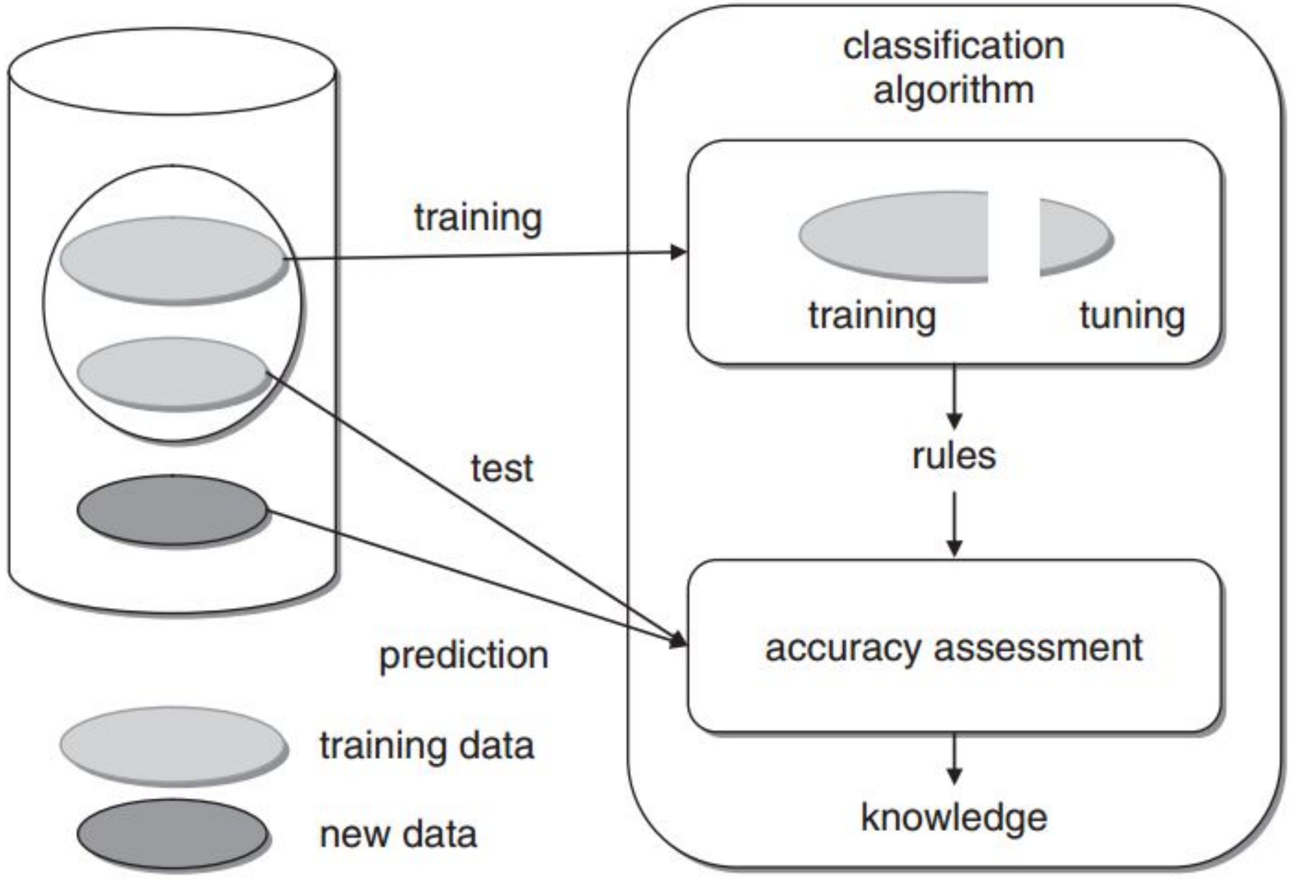
C4.5

successor of ID3

Use “Gain Ratio” to measure purity

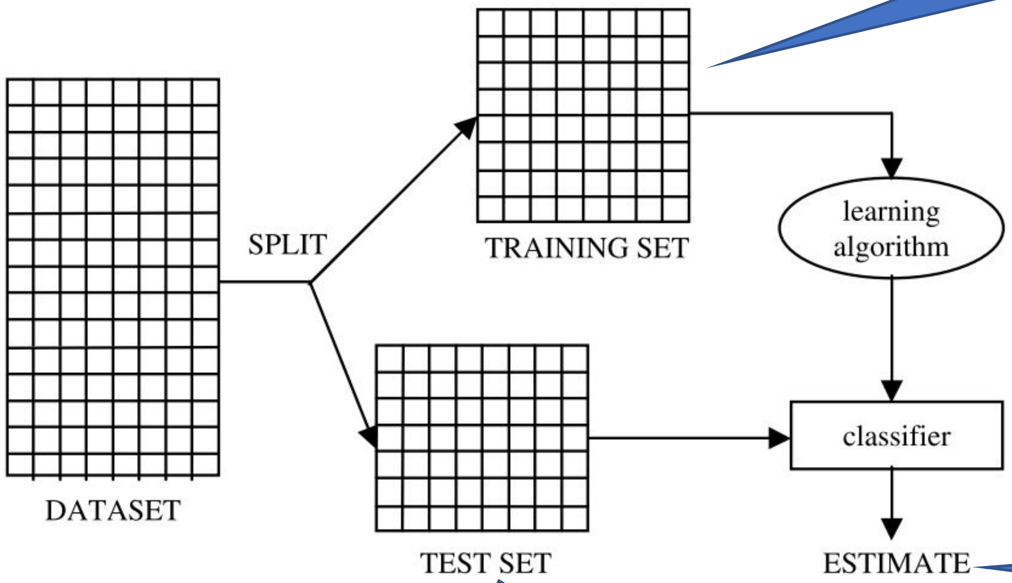
Classification

- Training phase
- Test phase
- Prediction phase



Predictive Accuracy

The training set is used to construct a classifier (decision tree, neural net etc.)



If the test set contains N instances of which C are correctly classified the predictive accuracy of the classifier for the test set is p = C/N

Unseen instances

Confusion Matrix

When Type I error is more important to be avoided? (search engine info retrieval)

		Condition (as determined by "Gold standard")		
		Condition positive	Condition negative	
Test outcome	Test outcome positive	True positive TP	False positive (Type I error) FP	Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$
	Test outcome negative	False negative (Type II error) FN	True negative TN	Negative predictive value = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$
		Sensitivity = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	Specificity = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Accuracy

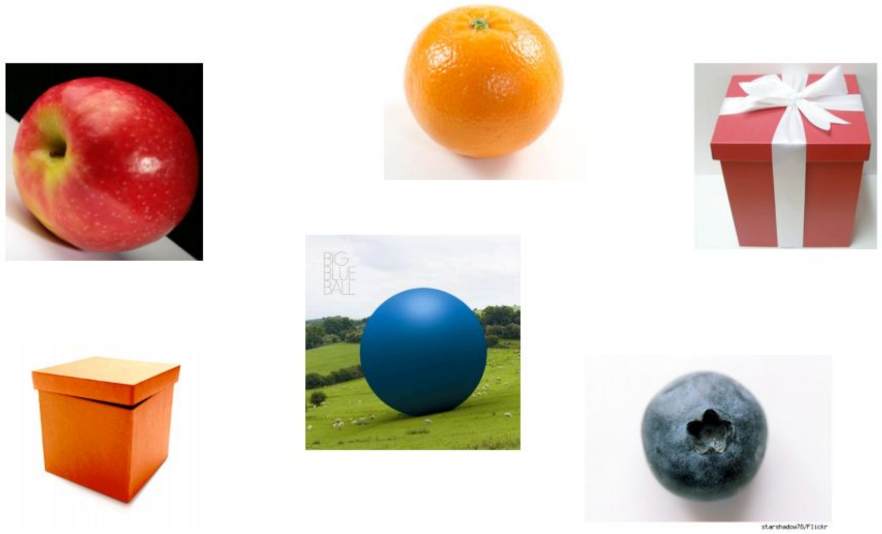
When Type II error is more important to be avoided? (medical cases)

Supervised Learning vs. Unsupervised Learning

- Unsupervised Learning:
  - Clustering
    - K-Means
    - Hierarchical Clustering
    - Density-based Clustering

Clustering

Cluster Analysis



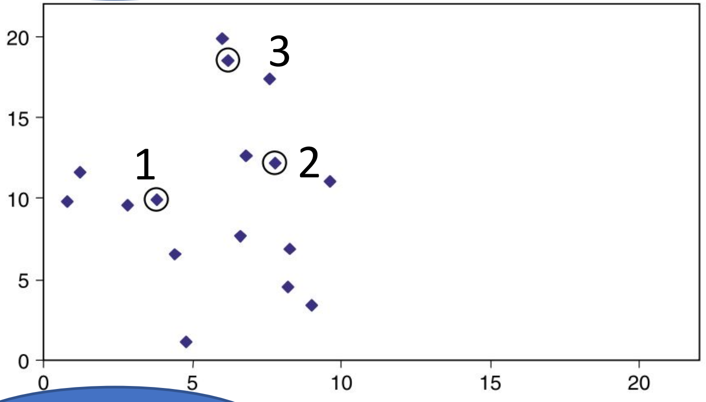
How many clusters do you expect?

Clustering

- K-means

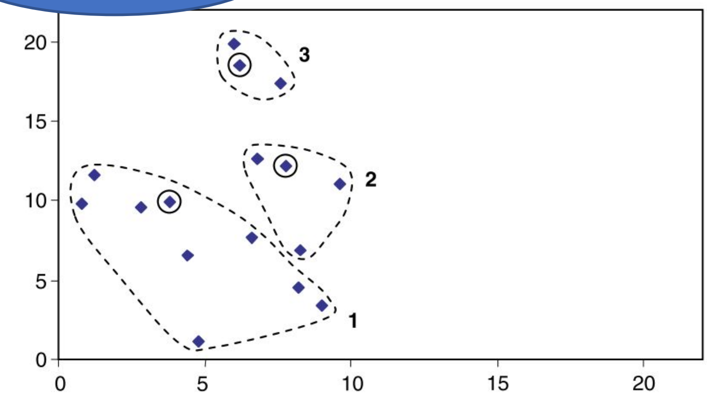
x	y
6.8	12.6
0.8	9.8
1.2	11.6
2.8	9.6
3.8	9.9
4.4	6.5
4.8	1.1
6.0	19.9
6.2	18.5
7.6	17.4
7.8	12.2
6.6	7.7
8.2	4.5
8.4	6.9
9.0	3.4
9.6	11.1

Initial Set Up



	Initial	
	x	y
Centroid 1	3.8	9.9
Centroid 2	7.8	12.2
Centroid 3	6.2	18.5

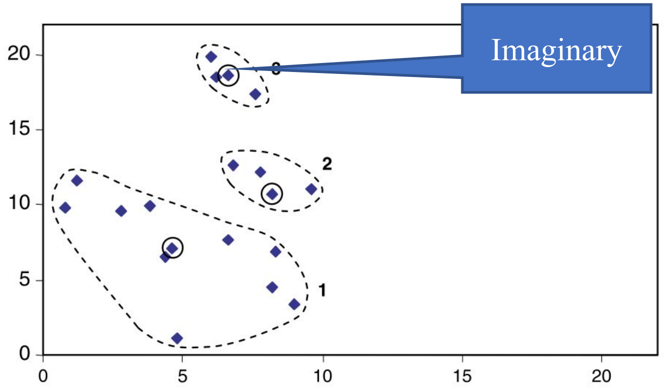
Iteration 1



	Initial		After first iteration	
	x	y	x	y
Centroid 1	3.8	9.9	4.6	7.1
Centroid 2	7.8	12.2	8.2	10.7
Centroid 3	6.2	18.5	6.6	18.6

Iteration 2…n

Repeat…until the centroids no longer move



	Initial		After first iteration		After second iteration	
	x	y	x	y	x	y
Centroid 1	3.8	9.9	4.6	7.1	5.0	7.1
Centroid 2	7.8	12.2	8.2	10.7	8.1	12.0
Centroid 3	6.2	18.5	6.6	18.6	6.6	18.6

Clustering

- Hierarchical Clustering

	a	b	c	d	e	f
a	0	12	6	3	25	4
b	12	0	19	8	14	15
c	6	19	0	12	5	18
d	3	8	12	0	11	9
e	25	14	5	11	0	7
f	4	15	18	9	7	0

Distance matrix

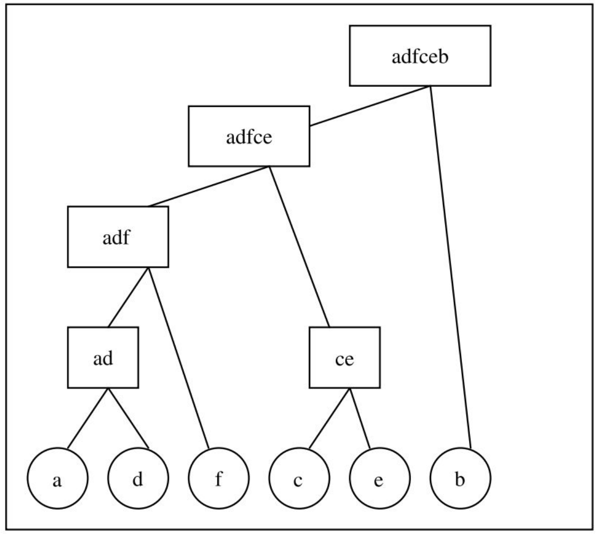
	ad	b	c	e	f
ad	0	8	6	11	4
b	8	0	19	14	15
c	6	19	0	5	18
e	11	14	5	0	7
	4	15	18	7	0

Distance Matrix
1. Single-link
2. Complete-link
3. Average-link

	adf	b	c	e
adf	0	8	6	7
b	8	0	19	14
c	6	19	0	5
e	7	14	5	0

	adf	b	ce
adf	0	8	6
b	8	0	14
ce	6	14	0

	adfce	b
adfce	0	8
b	8	0

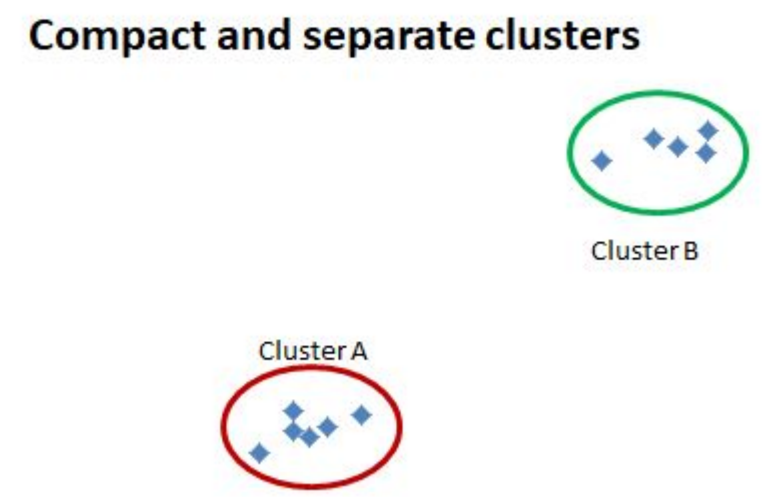


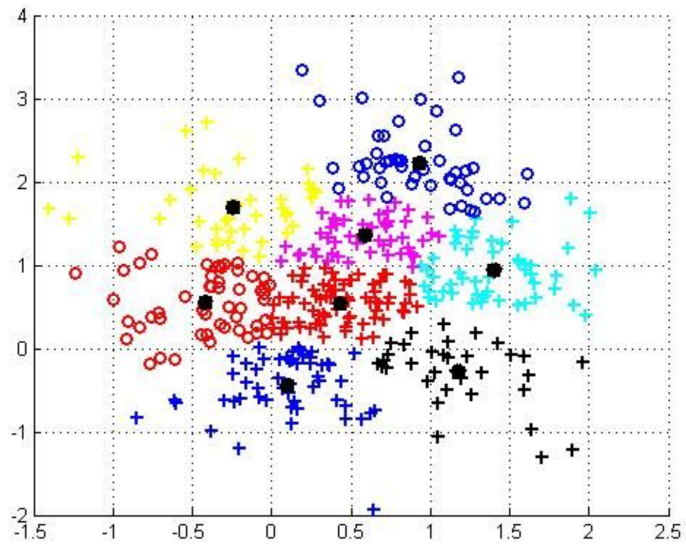
Dendrogram

Clustering Evaluation

For example:

- Compactness (e.g., within-groups/clusters sum of squares)
- Seperation (e.g., average euclidean distance between cluster centroids)





Clustering Evaluation

- Silhouette index
- Davies-Bouldin
- Calinski-Harabasz
- Dunn index
- R-squared index
- Hubert-Levin (C-index)
- Krzanowski-Lai index
- Hartigan index
- Root-mean-square standard deviation (RMSSTD) index
- Semi-partial R-squared (SPR) index
- Distance between two clusters (CD) index
- weighted inter-intra index
- Homogeneity index
- Separation index

Application of Clustering

- Marketing research
  - Identify different groups of customers
  - Customization

- Social Network
  - Identify different SN users
  - Personalized recommendation

Data Preparation

- When do we need data standardization?

  Example: Person=(age, marathon distance)
  A. (22, 10000m)
  B. (22, 20000m)
  C. (80, 5000m)

  Question: Who is more similar to A?
  B or C?

Data Preparation

- **Decimal scaling**

$$x'_{ij} = \frac{x_{ij}}{10^h},$$

- **Min-max**

$$x'_{ij} = \frac{x_{ij} - x_{\min,j}}{x_{\max,j} - x_{\min,j}}(x'_{\max,j} - x'_{\min,j}) + x'_{\min,j},$$

$$x_{\min,j} = \min_i \; x_{ij}, \quad x_{\max,j} = \max_i \; x_{ij},$$

- ***z*-index**

$$x'_{ij} = \frac{x_{ij} - \bar{\mu}_j}{\bar{\sigma}_j},$$

K-means Clustering

How to choose k?

Choose k based on the how results will be used
e.g., “How many market segments do we want?”

Also experiment with slightly different k’s
Initial partition into clusters can be random, or based on domain knowledge

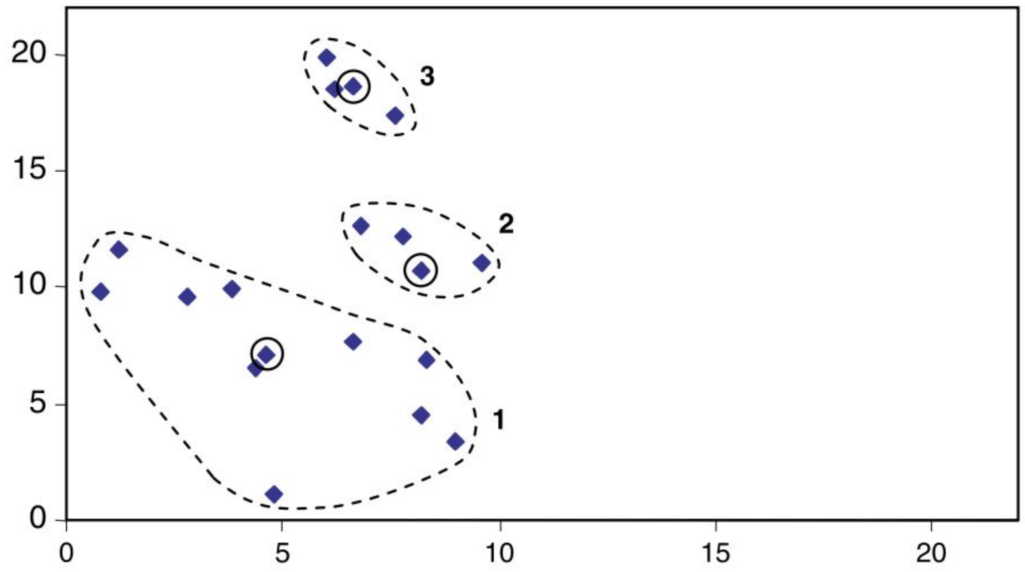
If random partition, repeat the process with different random
partitions

K-means Clustering

The within-groups/clusters sum of squares (WSS):

$$WSS(k) = \sum_{i=1}^{n} \sum_{j=0}^{p} (x_{ij} - mean(x_{kj}))^2$$

where, k is the cluster, x_{ij} is the value of the j^{th} variable for the i^{th} observation, and $mean(x_{kj})$ is the mean of the j^{th} variable for the k^{th} cluster.



K-means Clustering

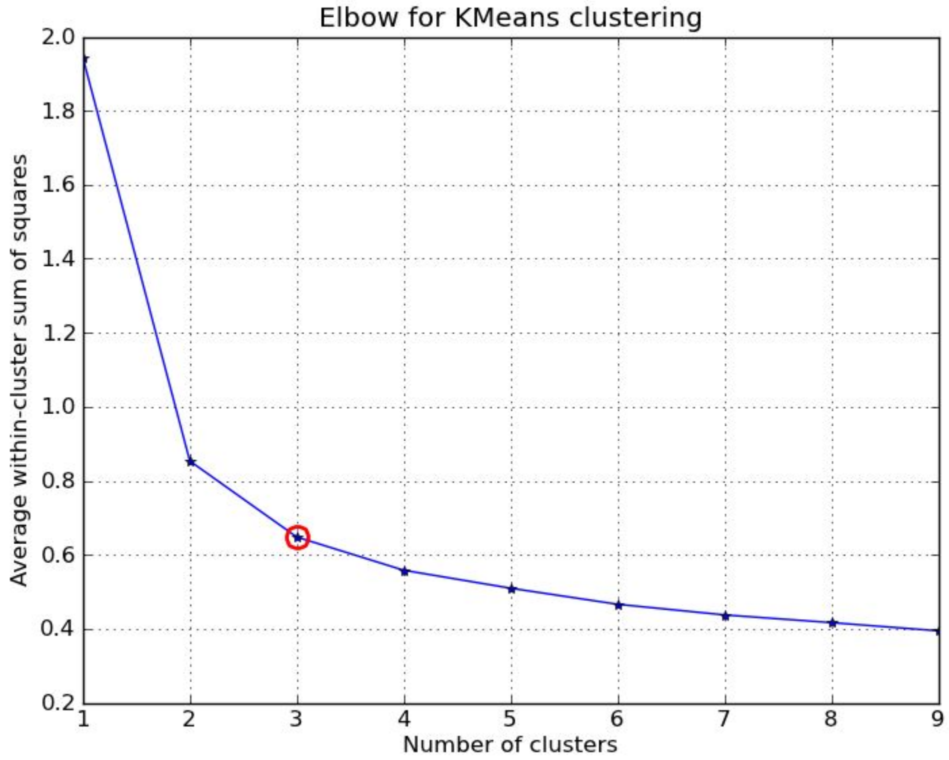
How to choose k?

Elbow method

– Gauge how the heterogeneity within clusters changes for various of k.

• The heterogeneity within clusters is expected to decreases with more clusters.

• The heterogeneity is measured by within-clusters/groups sum of squares (WSS)



Programming Assignment 4

Using the BT2101 Tutorial 4 Programming code (Clustering.ipynb), please answer the questions in the jupyter notebook

Answer all in the jupyter notebook.

Instructions

Submit Python Notebook to the submission folder and Named:
AXXXX_T4_program.ipynb

Include your answers in the jupyter notebook

- You need to show outputs, instead of just showing functions.

Submit a FINAL program by Sep-25 (by 12:00pm noon)

- Based on Clustering.ipynb

Thank you!

Reminder - Matrix Math

- Scalar

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

$$2A = 2 \cdot \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 2\cdot 1 & 2\cdot 2 \\ 2\cdot 3 & 2\cdot 4 \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 6 & 8 \end{bmatrix}$$

- Matrix Multiplication

AB – n*p

If $\mathbf{A}$ is an $n \times m$ matrix and $\mathbf{B}$ is an $m \times p$ matrix,

$$\mathbf{A} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nm} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1p} \\ B_{21} & B_{22} & \cdots & B_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ B_{m1} & B_{m2} & \cdots & B_{mp} \end{pmatrix}$$

$$\mathbf{AB} = \begin{pmatrix} (\mathbf{AB})_{11} & (\mathbf{AB})_{12} & \cdots & (\mathbf{AB})_{1p} \\ (\mathbf{AB})_{21} & (\mathbf{AB})_{22} & \cdots & (\mathbf{AB})_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ (\mathbf{AB})_{n1} & (\mathbf{AB})_{n2} & \cdots & (\mathbf{AB})_{np} \end{pmatrix}$$