

# Data Mining

## Lecture 17

Ananya Jana  
CS360

Fall 2024



# Acknowledgement

---

Thanks to the authors Pang-Ning Tan, Michael Steinbach, Vipin Kumar for their slides

# Neural Network Applications

---

---

# Imbalanced Class problem

# Class Imbalance Problem

---

- Lots of classification problems where the classes are skewed (more records from one class than another)
  - Credit card fraud
  - Intrusion detection
  - Defective products in manufacturing assembly line

# Class Imbalance Problem

---

- Evaluation measures such as accuracy is not well-suited for imbalanced class
- Detecting the rare class is like finding needle in a haystack

# Confusion Matrix

---

- Confusion Matrix:

ACTUAL CLASS	PREDICTED CLASS	
	Class=Yes	Class=No
Class=Yes	a	b
	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

# Accuracy

ACTUAL CLASS	PREDICTED CLASS	
	Class=Yes	Class=No
Class=Yes	a (TP)	b (FN)
	c (FP)	d (TN)

- Most widely-used metric:

$$\text{Accuracy} = \frac{a + d}{a + b + c + d} = \frac{TP + TN}{TP + TN + FP + FN}$$



# Problem with Accuracy

---

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10

# Problem with Accuracy

---

- Consider a 2-class problem
  - Number of Class NO examples = 990
  - Number of Class YES examples = 10
- If a model predicts everything to be class NO, accuracy is  $990/1000 = 99\%$ 
  - This is misleading because the model does not detect any class YES example
  - Detecting the rare class is usually more interesting (e.g., frauds, intrusions, defects, etc)

# Alternative Measures

	PREDICTED CLASS		
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

$$\text{Precision (p)} = \frac{a}{a + c}$$

$$\text{Recall (r)} = \frac{a}{a + b}$$

$$\text{F - measure (F)} = \frac{2rp}{r + p} = \frac{2a}{2a + b + c}$$

# Alternative Measures

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	10	0
	Class=No	10	980

$$\text{Precision (p)} = \frac{10}{10+10} = 0.5$$

$$\text{Recall (r)} = \frac{10}{10+0} = 1$$

$$\text{F - measure (F)} = \frac{2 * 1 * 0.5}{1 + 0.5} = 0.62$$

$$\text{Accuracy} = \frac{990}{1000} = 0.99$$

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	1	9
	Class=No	0	990

$$\text{Precision (p)} = \frac{1}{1+0} = 1$$

$$\text{Recall (r)} = \frac{1}{1+9} = 0.1$$

$$\text{F - measure (F)} = \frac{2 * 0.1 * 1}{1 + 0.1} = 0.18$$

$$\text{Accuracy} = \frac{991}{1000} = 0.991$$

# Alternative Measures

---

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	40	10
	Class=No	10	40

Precision (p) = 0.8

Recall (r) = 0.8

F - measure (F) = 0.8

Accuracy = 0.8

# Alternative Measures

ACTUAL CLASS	PREDICTED CLASS		
		Class=Yes	Class=No
Class=Yes	Class=Yes	40	10
	Class=No	10	40

Precision (p) = 0.8

Recall (r) = 0.8

F - measure (F) = 0.8

Accuracy = 0.8

ACTUAL CLASS	PREDICTED CLASS		
		Class=Yes	Class=No
Class=Yes	Class=Yes	40	10
	Class=No	1000	4000

Precision (p) = ~ 0.04

Recall (r) = 0.8

F - measure (F) = ~ 0.08

Accuracy = ~ 0.8

# Measures of classification performance

ACTUAL CLASS	PREDICTED CLASS		
		Yes	No
		Yes	No
ACTUAL CLASS	Yes	TP	FN
	No	FP	TN

$\alpha$  is the probability that we reject the null hypothesis when it is true. This is a Type I error or a false positive (FP).

$\beta$  is the probability that we accept the null hypothesis when it is false. This is a Type II error or a false negative (FN).

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

$$ErrorRate = 1 - accuracy$$

$$Precision = Positive Predictive Value = \frac{TP}{TP + FP}$$

$$Recall = Sensitivity = TP Rate = \frac{TP}{TP + FN}$$

$$Specificity = TN Rate = \frac{TN}{TN + FP}$$

$$FP Rate = \alpha = \frac{FP}{TN + FP} = 1 - specificity$$

$$FN Rate = \beta = \frac{FN}{FN + TP} = 1 - sensitivity$$

$$Power = sensitivity = 1 - \beta$$

# Alternative Measures

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	40	10
	Class=No	10	40

Precision (p) = 0.8

TPR = Recall (r) = 0.8

FPR = 0.2

F - measure (F) = 0.8

Accuracy = 0.8

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	40	10
	Class=No	1000	4000

Precision (p) = ~ 0.04

TPR = Recall (r) = 0.8

FPR = 0.2

F - measure (F) = ~ 0.08

Accuracy = ~ 0.8



# Alternative Measures

	PREDICTED CLASS		
ACTUAL CLASS		Class= Yes	Class= No
	Class= Yes	10	40
	Class= No	10	40

Precision (p) = 0.5

TPR = Recall (r) = 0.2

FPR = 0.2

	PREDICTED CLASS		
ACTUAL CLASS		Class= Yes	Class= No
	Class= Yes	25	25
	Class= No	25	25

Precision (p) = 0.5

TPR = Recall (r) = 0.5

FPR = 0.5

	PREDICTED CLASS		
ACTUAL CLASS		Class= Yes	Class= No
	Class= Yes	40	10
	Class= No	40	10

Precision (p) = 0.5

TPR = Recall (r) = 0.8

FPR = 0.8

# ROC (Receiver Operating Characteristic)

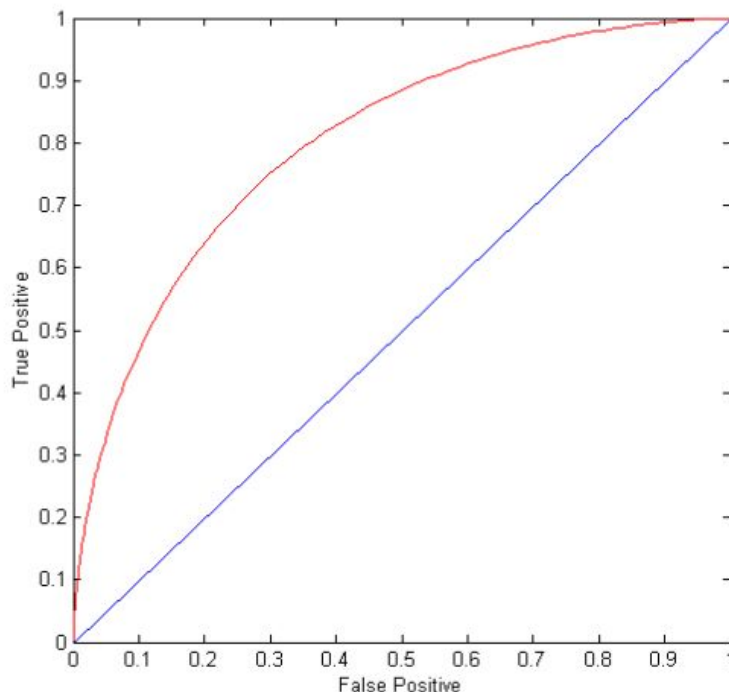
---

- A graphical approach for displaying trade-off between detection rate and false alarm rate
- Developed in 1950s for signal detection theory to analyze noisy signals
- ROC curve plots TPR against FPR
  - Performance of a model represented as a point in an ROC curve
  - Changing the threshold parameter of classifier changes the location of the point

# ROC Curve

(TPR, FPR):

- (0,0): declare everything to be negative class
- (1,1): declare everything to be positive class
- (1,0): ideal
- Diagonal line:
  - Random guessing
  - Below diagonal line:
    - ◆ prediction is opposite of the true class

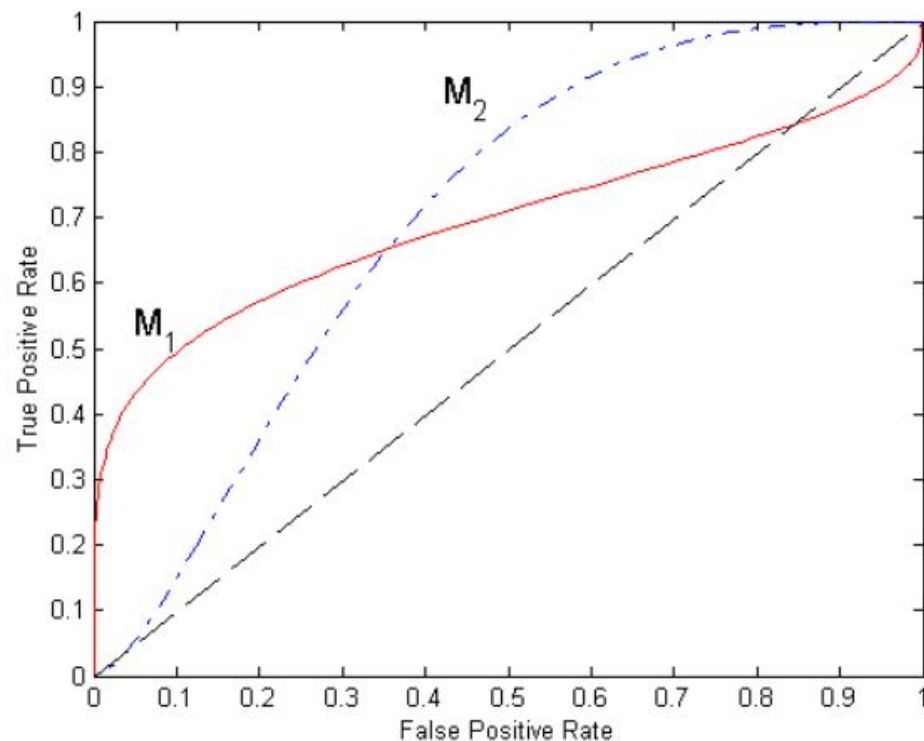


# ROC Curve

---

- To draw ROC curve, classifier must produce continuous-valued output
  - Outputs are used to rank test records, from the most likely positive class record to the least likely positive class record
- Many classifiers produce only discrete outputs (i.e., predicted class)
  - How to get continuous-valued outputs?
    - ◆ Decision trees, rule-based classifiers, neural networks, Bayesian classifiers, k-nearest neighbors, SVM

# Using ROC for Model Comparison



- No model consistently outperform the other
  - $M_1$  is better for small FPR
  - $M_2$  is better for large FPR
- Area Under the ROC curve
  - Ideal:
    - Area = 1
  - Random guess:
    - Area = 0.5

# How to construct ROC curve

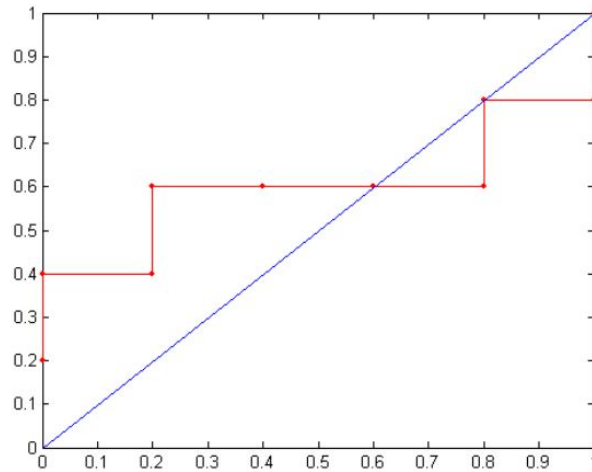
Instance	Score	True Class
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

- Use a classifier that produces a continuous-valued score for each instance
  - The more likely it is for the instance to be in the + class, the higher the score
- Sort the instances in decreasing order according to the score
- Apply a threshold at each unique value of the score
- Count the number of TP, FP, TN, FN at each threshold
  - $TPR = TP / (TP + FN)$
  - $FPR = FP / (FP + TN)$

# How to construct ROC curve

Class	+	-	+	-	-	-	+	-	+	+	
Threshold >=	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0

ROC Curve:



# Tasks

---

1. Calculate the metrics Accuracy, Precision, Recall, F1 score, Specificity, False Positive Rate for the following tables

	Predicted Class Yes	Predicted Class No
Actual Class Yes	200	200
Actual Class No	200	200

	Predicted Class Yes	Predicted Class No
Actual Class Yes	0	10
Actual Class No	0	790



# Tasks

---

2. Draw the ROC curve for the following table

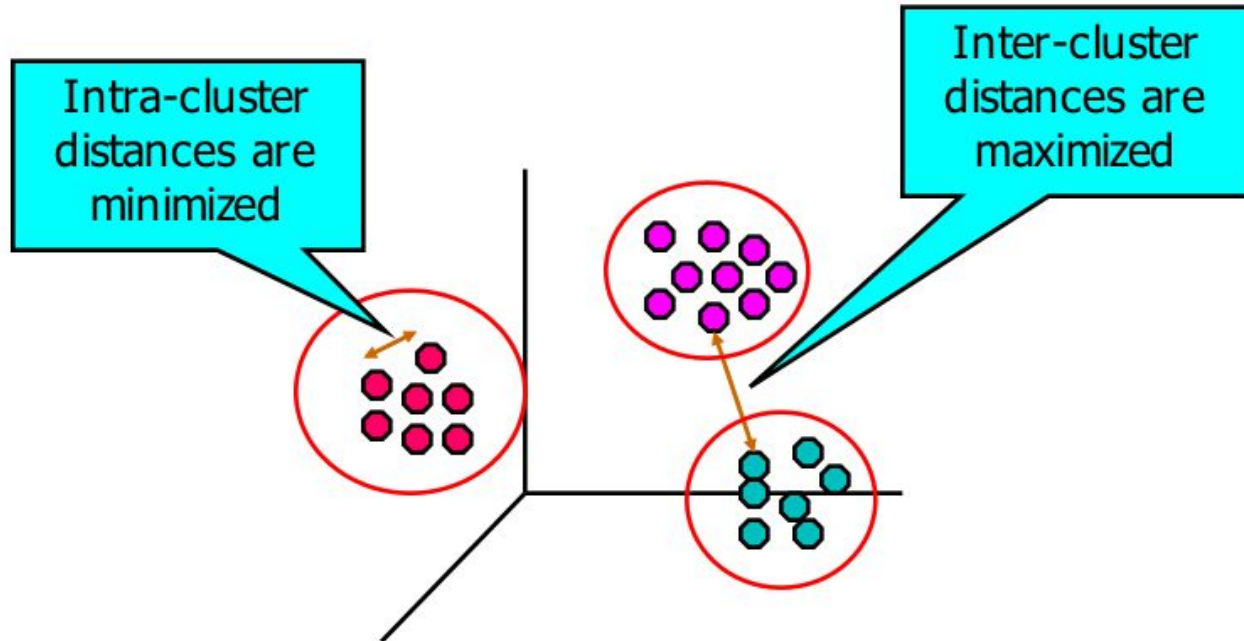
Instance	Score	True Class
1	0.95	+
2	0.35	-
3	0.85	+
4	0.25	+

---

# Cluster analysis

# What is Cluster Analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



# Applications of Cluster Analysis

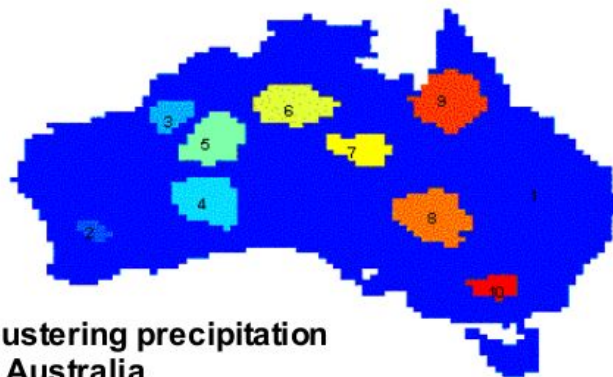
## ● Understanding

- Group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations

	<i>Discovered Clusters</i>	<i>Industry Group</i>
<b>1</b>	Applied-Matl-DOWN, Bay-Network-DOWN, 3-COM-DOWN, Cabletron-Sys-DOWN, CISCO-DOWN, HP-DOWN, DSC-Comm-DOWN, INTEL-DOWN, LSI-Logic-DOWN, Micron-Tech-DOWN, Texas-Inst-Down, Tellabs-Inc-Down, Natl-Semiconduct-DOWN, Oracl-DOWN, SGI-DOWN, Sun-DOWN	Technology1-DOWN
<b>2</b>	Apple-Comp-DOWN, Autodesk-DOWN, DEC-DOWN, ADV-Micro-Device-DOWN, Andrew-Corp-DOWN, Computer-Assoc-DOWN, Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN, Microsoft-DOWN, Scientific-Atl-DOWN	Technology2-DOWN
<b>3</b>	Fannie-Mae-DOWN, Fed-Home-Loan-DOWN, MBNA-Corp-DOWN, Morgan-Stanley-DOWN	Financial-DOWN
<b>4</b>	Baker-Hughes-UP, Dresser-Inds-UP, Halliburton-HLD-UP, Louisiana-Land-UP, Phillips-Petro-UP, Unocal-UP, Schlumberger-UP	Oil-UP

## ● Summarization

- Reduce the size of large data sets



Clustering precipitation  
in Australia

# What is not Cluster Analysis?

---

- Simple segmentation

- Dividing students into different registration groups alphabetically, by last name

- Results of a query

- Groupings are a result of an external specification
- Clustering is a grouping of objects based on the data

- Supervised classification

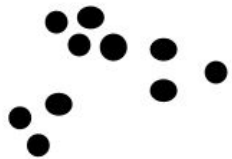
- Have class label information

- Association Analysis

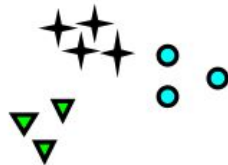
- Local vs. global connections

# Notion of a Cluster can be Ambiguous

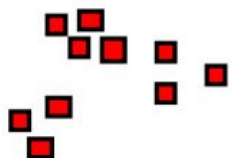
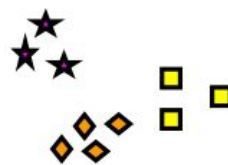
---



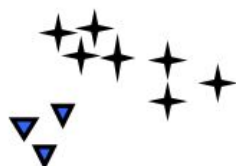
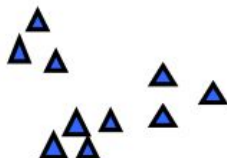
How many clusters?



Six Clusters



Two Clusters



Four Clusters



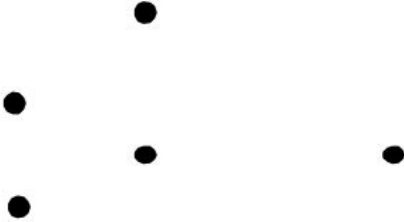
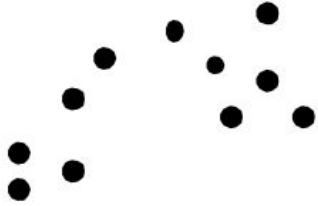
# Types of Clustering

---

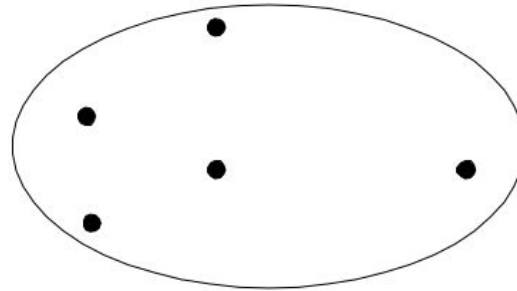
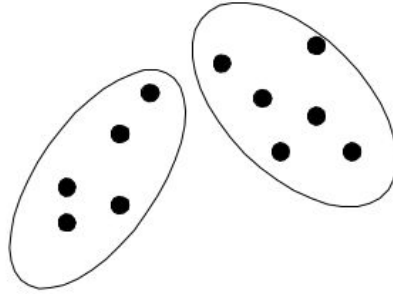
- A **clustering** is a set of clusters
- Important distinction between **hierarchical** and **partitional** sets of clusters
- **Partitional Clustering**
  - A division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- **Hierarchical clustering**
  - A set of nested clusters organized as a hierarchical tree

# Partitional Clustering

---



**Original Points**

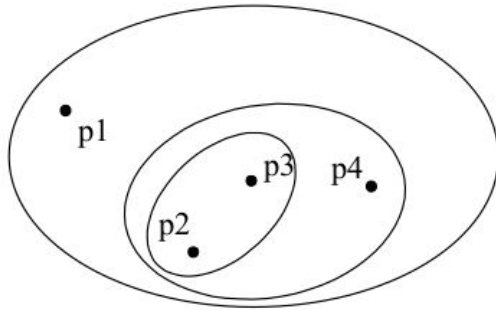


**A Partitional Clustering**

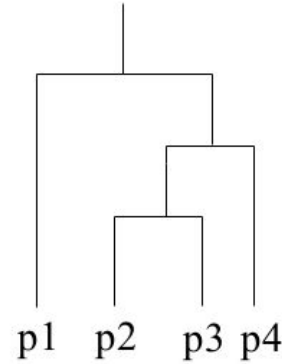


# Hierarchical Clustering

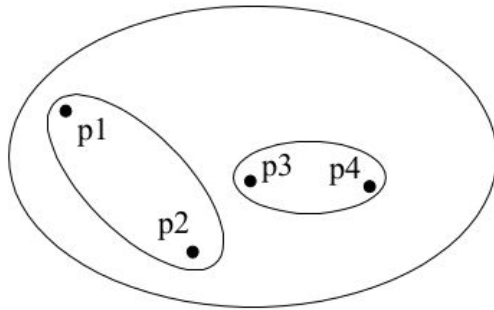
---



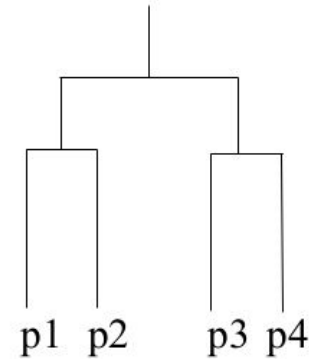
**Traditional Hierarchical Clustering**



**Traditional Dendrogram**



**Non-traditional Hierarchical Clustering**



**Non-traditional Dendrogram**

# Clustering Algorithms

---

- K Means Algorithm
- Hierarchical Algorithm

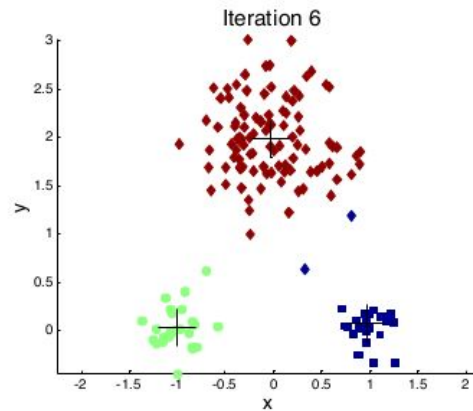
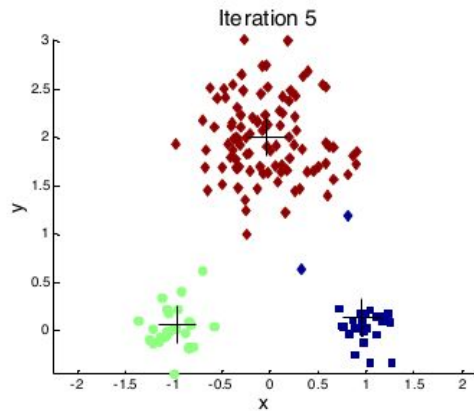
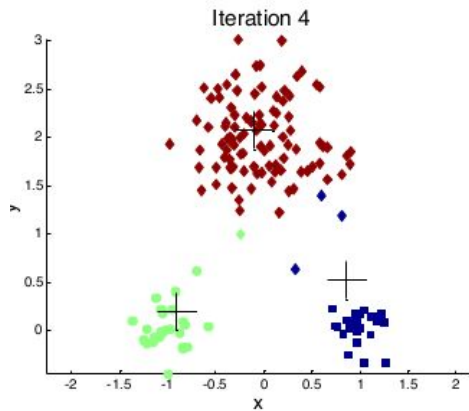
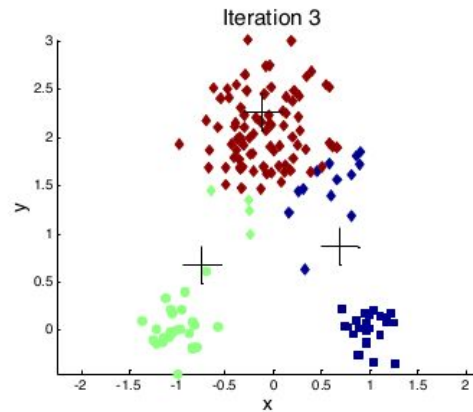
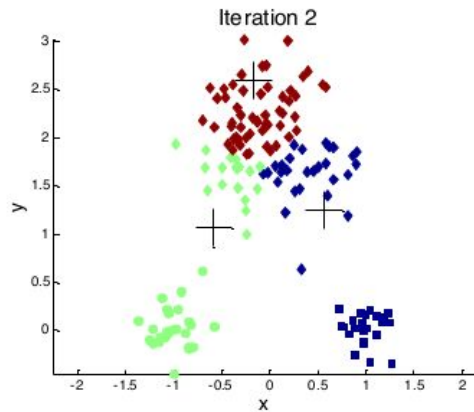
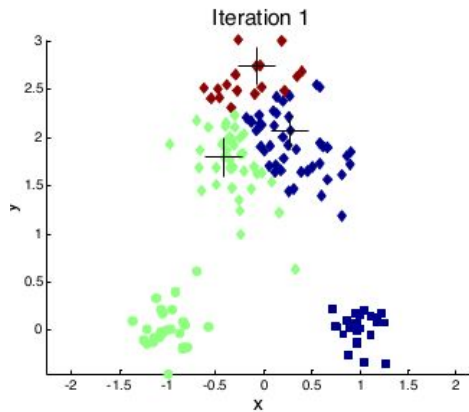
# K-means Clustering

---

- Partitional clustering approach
- Number of clusters,  $K$ , must be specified
- Each cluster is associated with a **centroid** (center point)
- Each point is assigned to the cluster with the closest centroid
- The basic algorithm is very simple

- 
- 1: Select  $K$  points as the initial centroids.
  - 2: **repeat**
  - 3:   Form  $K$  clusters by assigning all points to the closest centroid.
  - 4:   Recompute the centroid of each cluster.
  - 5: **until** The centroids don't change
-

# Iterations of K-means Clustering



# K-means Clustering – Details

---

- Initial centroids are often chosen randomly.
  - Clusters produced vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster.
- ‘Closeness’ is measured by Euclidean distance, cosine similarity, correlation, etc.
- K-means will converge for common similarity measures mentioned above.
- Most of the convergence happens in the first few iterations.
  - Often the stopping condition is changed to ‘Until relatively few points change clusters’
- Complexity is  $O(n * K * I * d)$ 
  - $n$  = number of points,  $K$  = number of clusters,  
 $I$  = number of iterations,  $d$  = number of attributes

# Task

---

Find the K-means clusters for the following dataset. Assume there are two clusters.  
Apply the K-means algorithms for three iterations.

(Please note that you can select two of the following points A, B, C, D as the initial centroids too)

Record	X1	X2
A	1	4
B	5	1
C	6	2
D	3	0