Machine Learning & Data Mining
Fundamentals of AI (AE1FAI)

Slides created by Dr Rong Qu & Dr John Drake
Modified by Dr Huan Jin

OUTLINE

- **Machine learning overview**
  - Classification (supervised learning)
  - Clustering (unsupervised learning)

WHAT IS MACHINE LEARNING

Machine Learning relates with the study, design and development of the algorithms that give computers the capability to learn without being explicitly programmed

-- Arthur Samuel

An agent is learning if it improves its performance on future tasks after making observations about the world.

MACHINE LEARNING ≈ LOOKING FOR FUNCTION

Speech Recognition

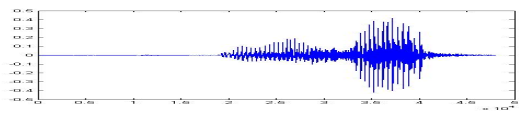
$f($  $) =$ "How are you"

Image Recognition

$f($  $) =$ "Cat"

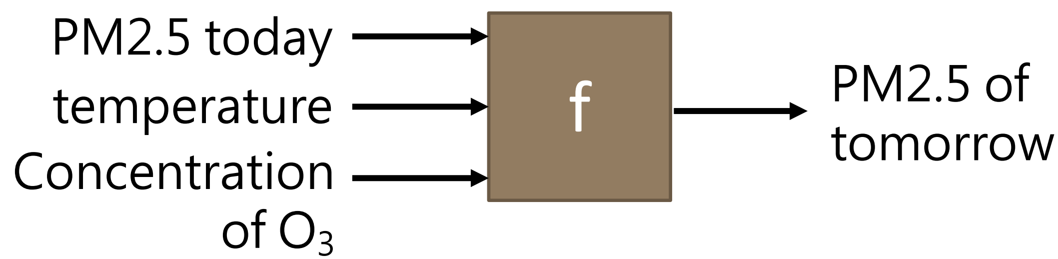
Playing Go

$f($  $) =$ "5-5" (next move)

DIFFERENT TYPES OF FUNCTIONS

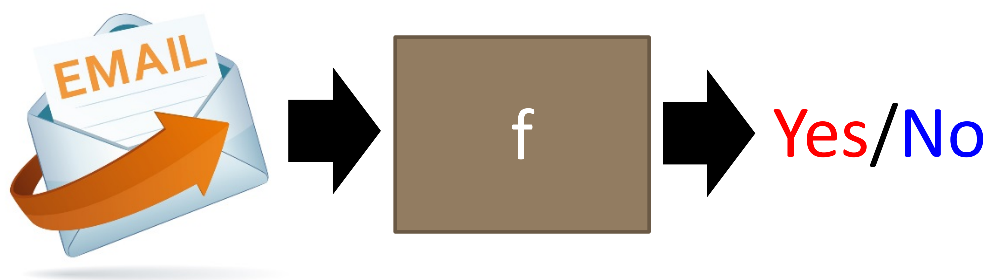
Regression: The function outputs a scalar.

Predict PM2.5



Classification: Given options (**classes**), the function outputs the correct one.

Spam filtering



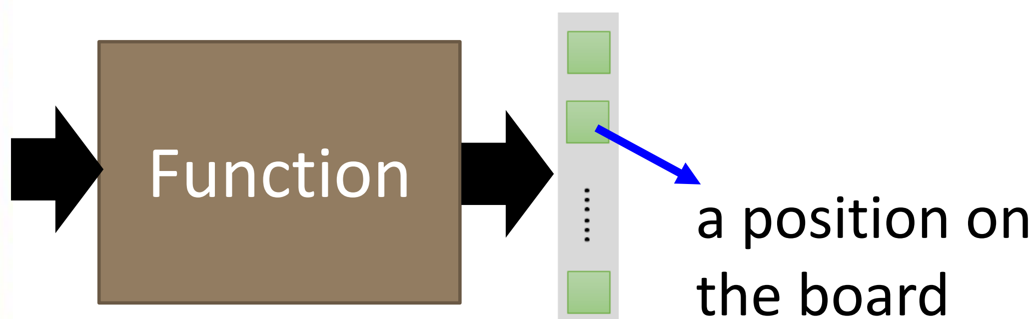
DIFFERENT TYPES OF FUNCTIONS

Classification: Given options (**classes**), the function outputs the correct one.

Each position is a class

(19 x 19 classes)

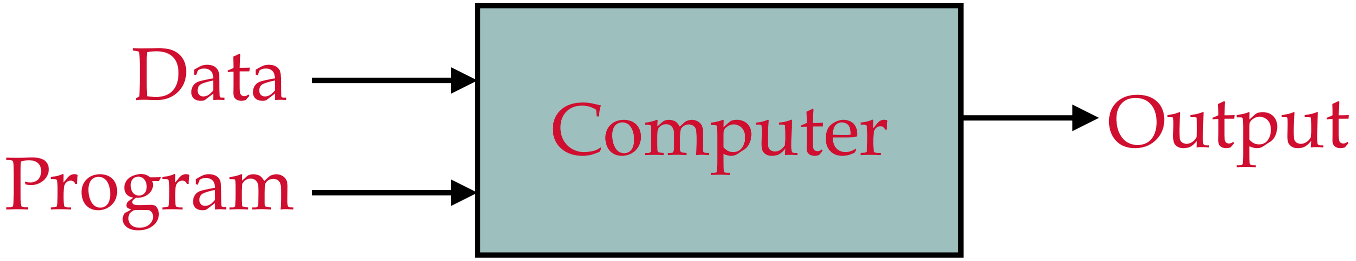
Playing GO



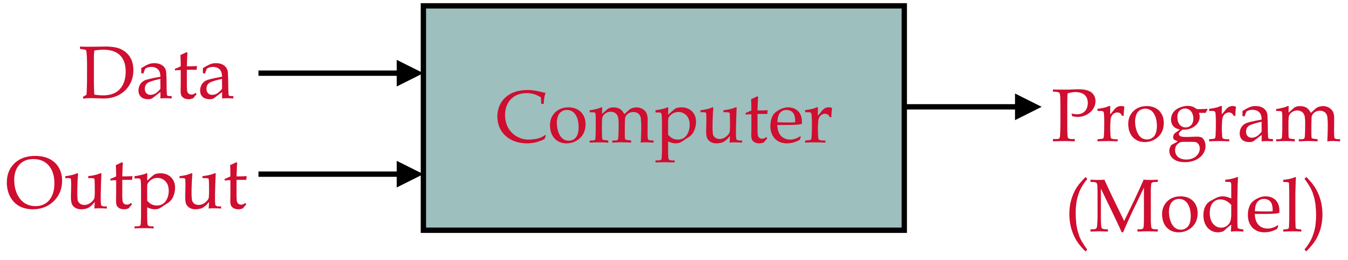
Next move

MACHINE LEARNING VS. TRADITIONAL PROGRAMMING

Traditional Programming



Machine Learning

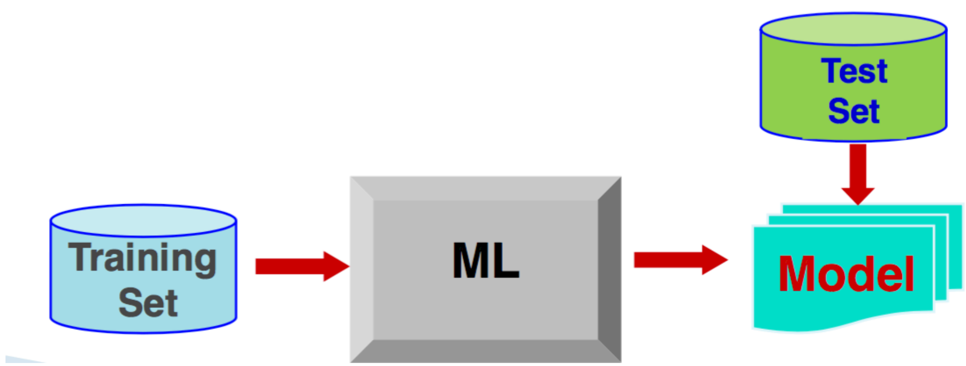


MACHINE LEARNING: PROCESS

Partition the total dataset into subsets:

Training set: Learning the parameters of the model

Test set: How the results will generalize to an independent (novel) data set



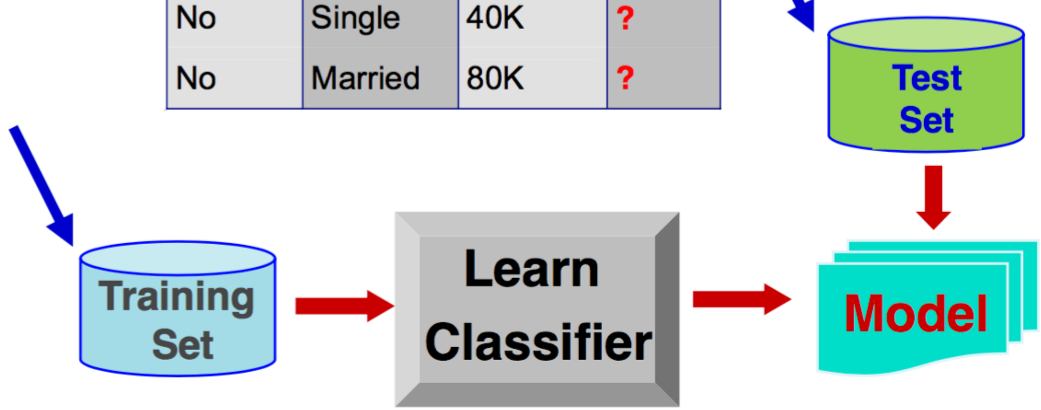
HOW MACHINE LEARNING WORKS

categorical categorical continuous class

Tid	Saving	Marital Status	Taxable Income	Good
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

record

Saving	Marital Status	Taxable Income	Good
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



HOW MACHINE LEARNING WORKS

Training

Training set may be noisy, e.g., $(x, (f(x) + \varepsilon))$

Training Images

Image Features → **x** → Training → Learned Model

Training Labels → **y** → Training

f()

Training: given a *training set* of labeled examples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, the function f is estimated by minimising the prediction error on the training set

Testing: apply f to a never seen before *test example* **x** and output the predicted value y = f(**x**)

Testing

Test Image

Image Features → **x** → Learned Model → Prediction

f(**x**)

y

Slide credit: D. Hoiem and L. Lazebnik

MACHINE LEARNING TASKS:
SUPERVISED LEARNING

- **Supervised learning:** the agent observes some example input-output pairs and learns a function that maps from input to output

- given input samples (**x**) and labeled outputs (**y**) of a function y = f(**x**), "learn" f, and evaluate it on new data

  – Classification: y is discrete (class labels). Learn a decision boundary that separates one class from another

  – Regression: y is continuous, e.g. linear regression. Learn a continuous input-output mapping, also known as "curve fitting" and "function approximation"

MACHINE LEARNING TASKS:
SUPERVISED LEARNING

- Examples:
  – is this image a cat, dog, car, house?
  – how would this user score that restaurant?
  – is this email spam?
  – what will be the sales, stock price next year?

MACHINE LEARNING TASKS: UNSUPERVISED LEARNING

- **Unsupervised learning**: given only samples x of the data, infers a function f such that y = f(x) describes the hidden structure of the unlabeled data - more of an exploratory/descriptive data analysis
  - – Clustering: y is discrete. Learn any intrinsic structure that is present in the data
  - – Dimensional Reduction: y is continuous. Discover a lower- dimensional surface on which the data lives

SUPERVISED LEARNING

- F(x): function
- D: training sample (x, F(x))

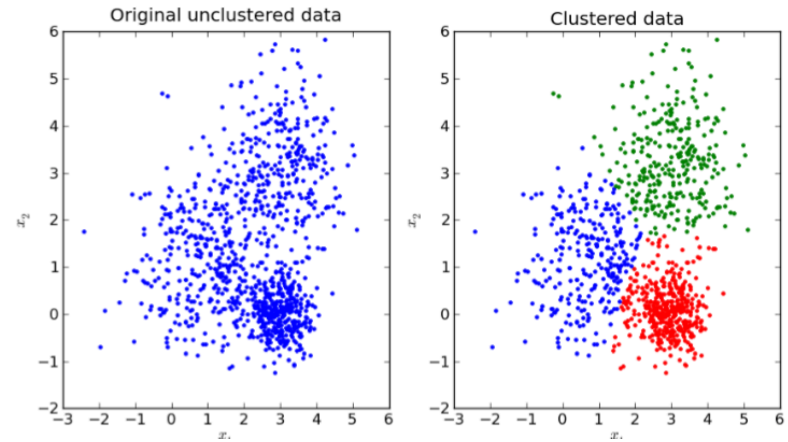
57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0	1

- G(x): model learned from D

71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0 ?

- Goal: minimise E[(T - O)2] for future use

UNSUPERVISED LEARNING



Training dataset:

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,1,1,0,0,0,0,0,0,0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0
40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1

71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0 ?

SUPERVIDED VS. UNSUPERVISED

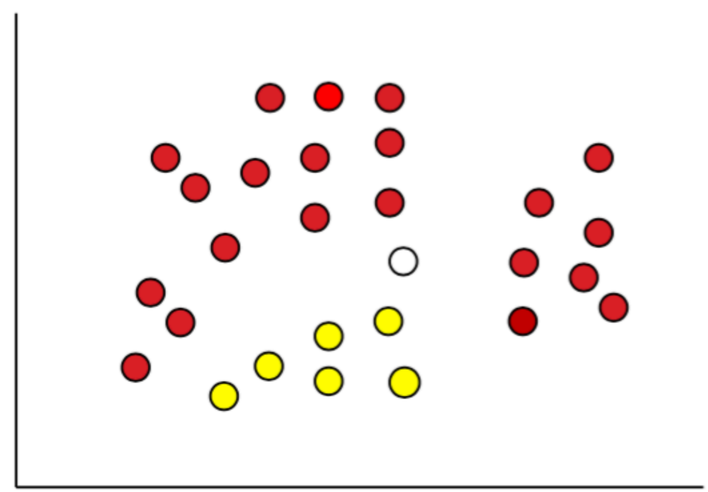
Supervised	Un-supervised
y = F(x): function	**y = ?**: no function
D: labeled training set	**D**: unlabeled data set
Learn: G(x): model trained to predict labels of new cases	**Learn:** ?
Goal: $E[(F(x)-G(x))^2] \approx 0$	**Goal:** ?

OUTLINE

- Machine learning overview
  - **Classification (supervised learning)**
  - Clustering (unsupervised learning)

CLASSIFICATION
(SUPERVISED LEARNING)

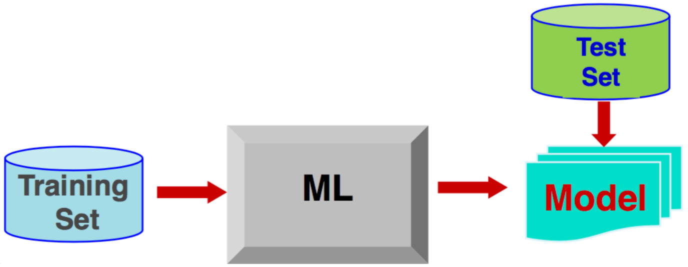
- Learn a method to predict the instance class from pre-labeled (classified) instances



Given a set of points from classes

what is the class of new point ?

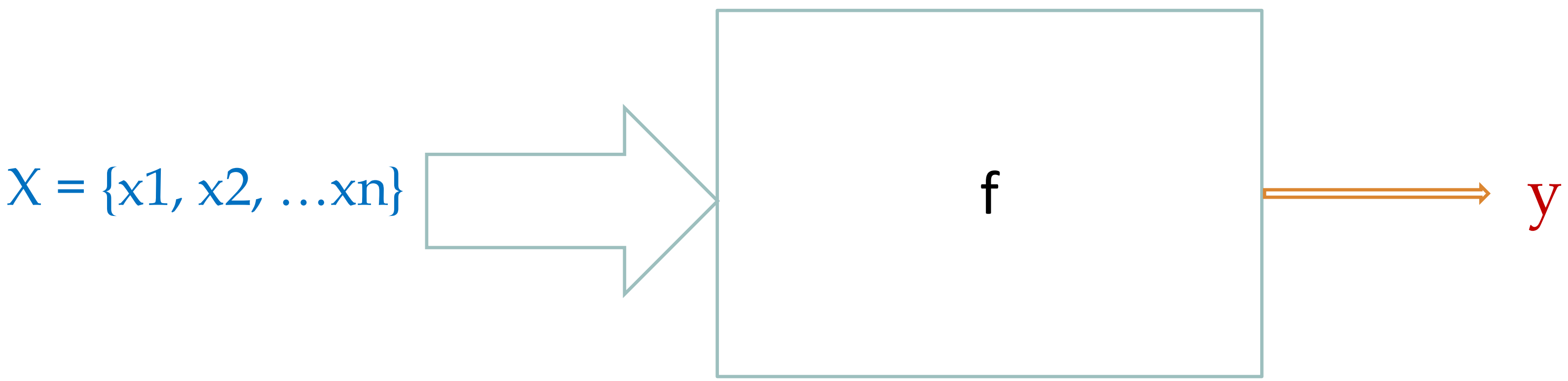
CLASSIFICATION
(SUPERVISED LEARNING)



- Data: a collection of records
  - Each record contains a set of attributes
  - One of the attributes is the class attribute
- Find a model for class attribute as a function of the other attributes
- Goal: assign a class to unseen records correctly
- Process
  - Divide the given data set into training & test sets
  - Use training set to build the model
  - y test set to validate the model

CLASSIFICATION
(SUPERVISED LEARNING)

- **Goal**: Predict class y = f(x1, x2, .. Xn)



CLASSIFICATION: APPLICATION

Target marketing

- **Goal:** Reduce cost of mailing by targeting consumers who are likely to buy a new cell-phone product.

- **Approach:**
  - Find the old data for a similar product.
  - Collect information of all customers.
    - Business type, where they stay, how much they earn, …
  - We know previous customers decision. This *{buy, don’t buy}* decision forms the *class attribute*.
  - Use this information to learn a classifier model.

CLASSIFICATION: APPLICATION

- **Goal**: Predict class y = f(x1, x2, .. Xn)

X = {Business type, where they stay, how much they earn, …}

y = {buy, don’t buy}

OTHER SUPERVISED LEARNING

- **Regression**: (linear or any other polynomial)

  $a*x_1 + b*x_2 + c = y$

- **Decision trees**: divide decision space into piecewise constant regions.

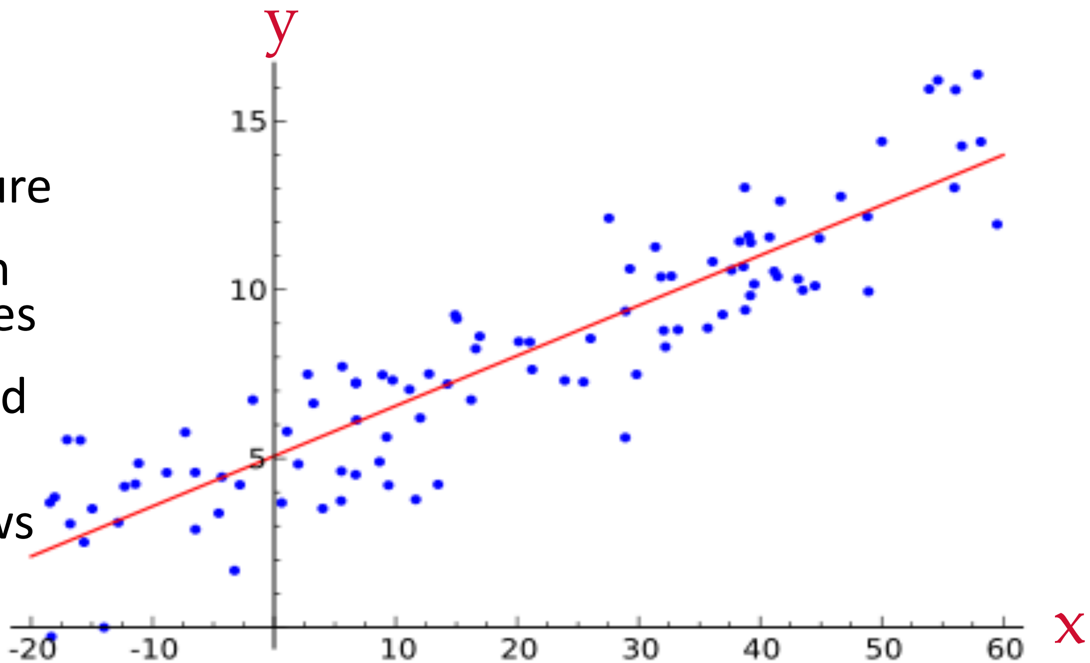
- **Neural networks**: partition by non-linear boundaries

- **Support vector machines, ...**

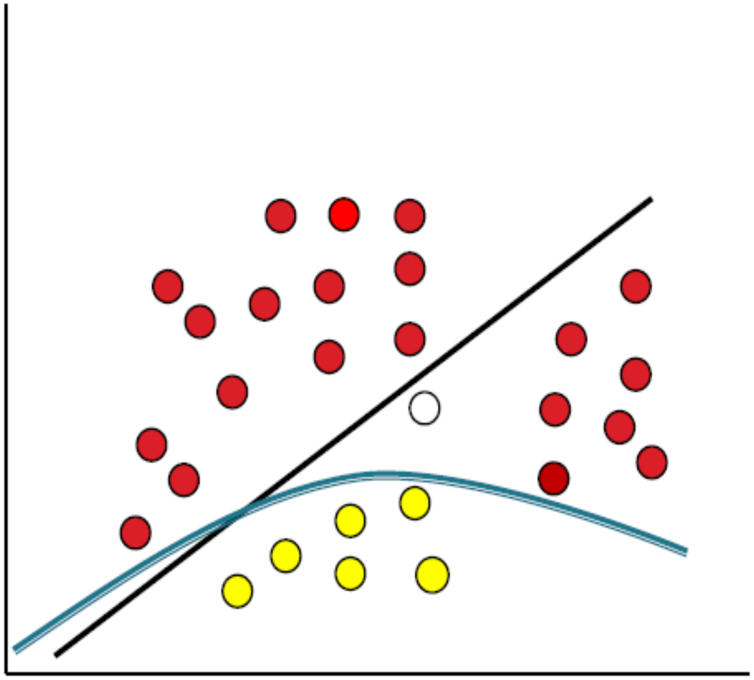
REGRESSION

- To find the best line (linear function y=f(x)) to explain the data

  - predict sales of new products based on advertising expenditure
  - time series prediction of stock market indices
  - Estimate weight based on BMI
  - predict the no of views of a youtuber.

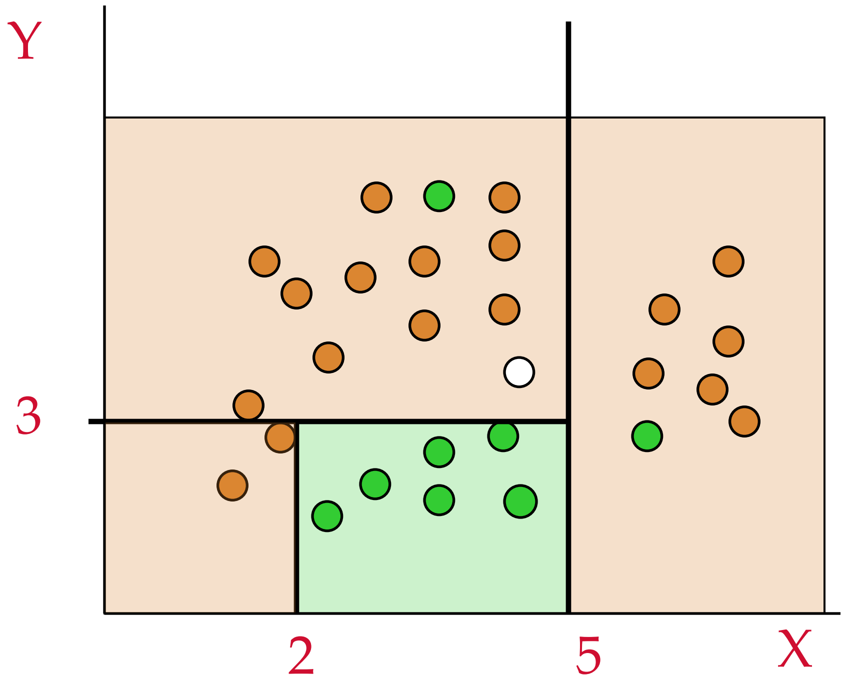


REGRESSION



- Linear Regression
- $w_0 + w_1 x = y$
- Regression computes w_i from data to minimise squared error to 'fit' the data
- Not flexible enough

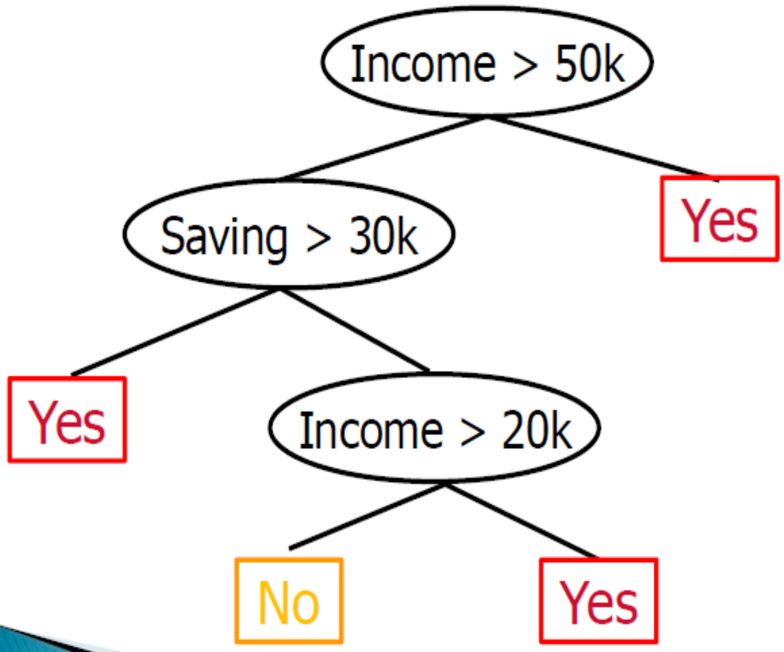
DECISION TREES

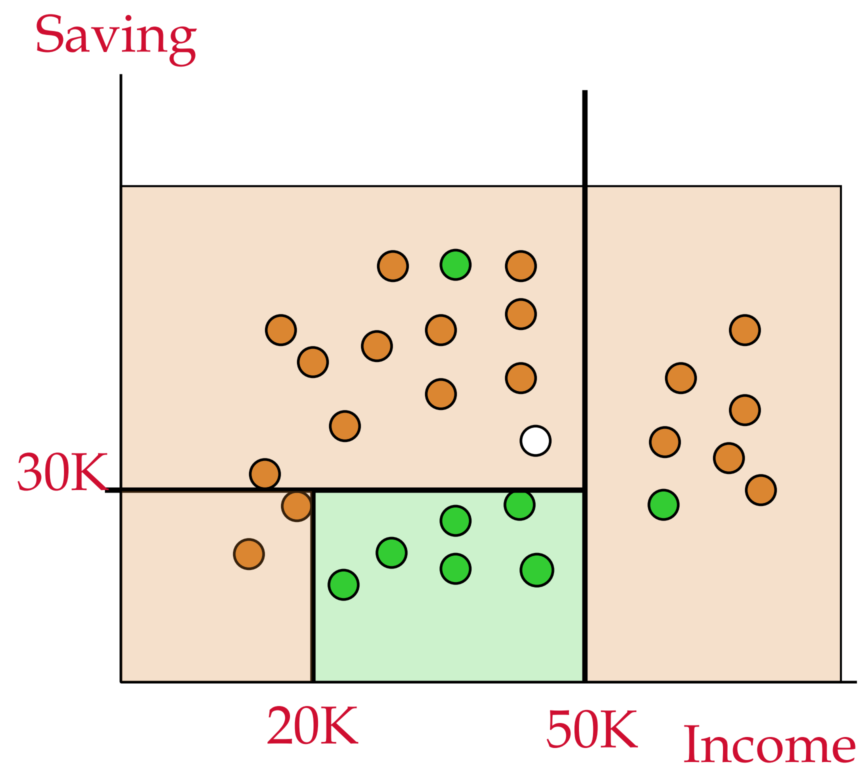


- if X > 5 then orange
- else if Y > 3 then orange
- else if X > 2 then green
- else orange

DECISION TREES

- **Internal node**: decision rule on one or more attributes
- **Leaf node:** a predicted class label

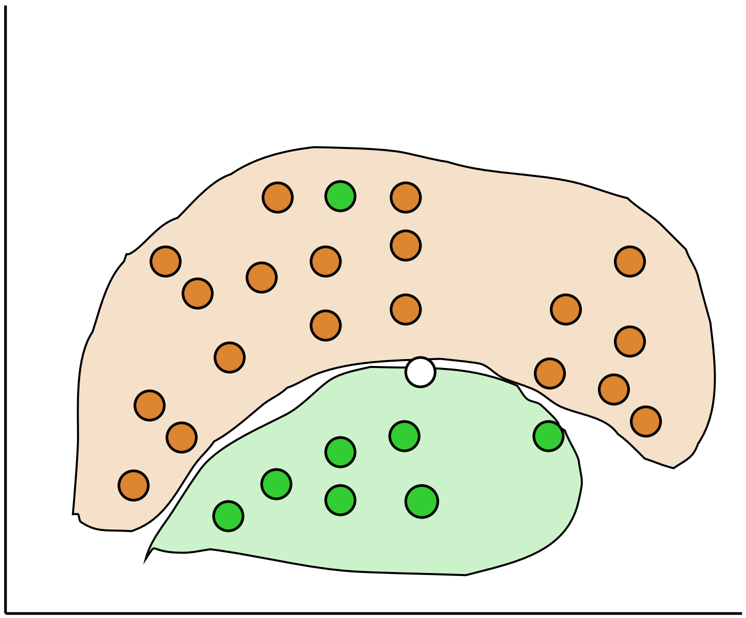


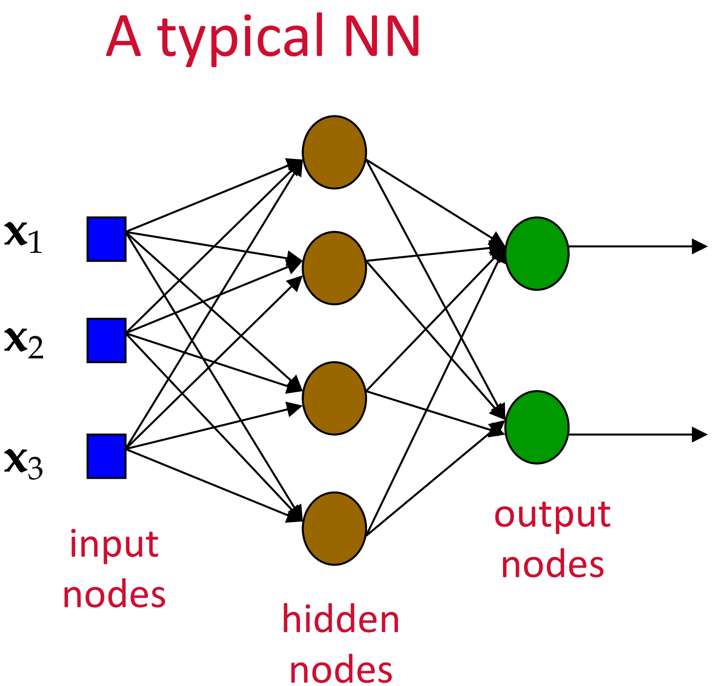


DECISION TREES

Pros	Cons
Reasonable training time	Simple decision boundaries
Can handle large number of attributes	Problems with lots of missing data
Easy to implement	Cannot handle complicated relationship between
Easy to interpret	

NEURAL NETWORKS

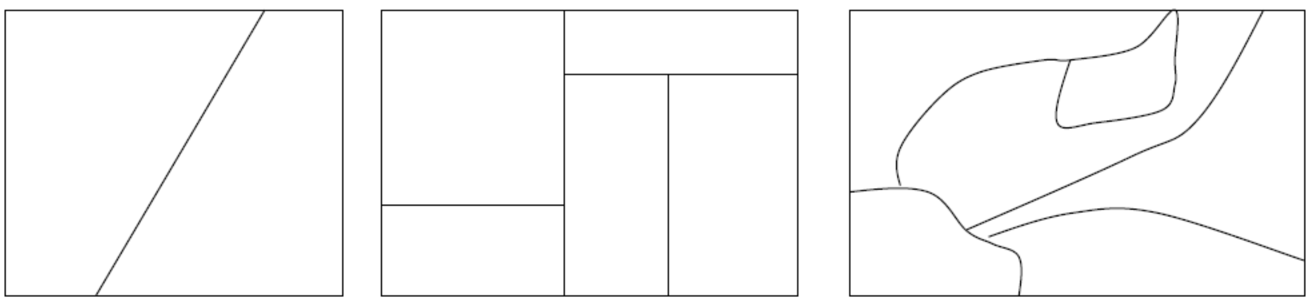




NEURAL NETWORKS

- Useful for learning complex data like speech, image and handwriting recognition

Decision boundaries:



Linear regression **Decision tree** **Neural network**

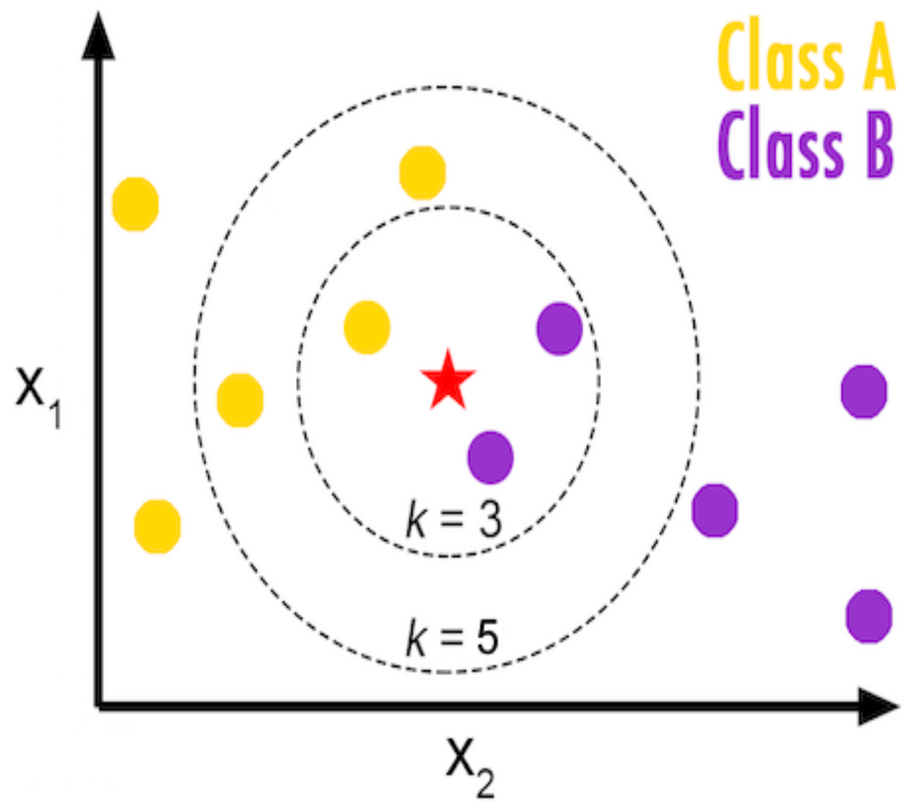
- Regression: use of linear or any other polynomial
- Decision Trees: divide decision space into piecewise regions
- Neural Networks: partition by nonlinear boundaries

NEURAL NETWORKS

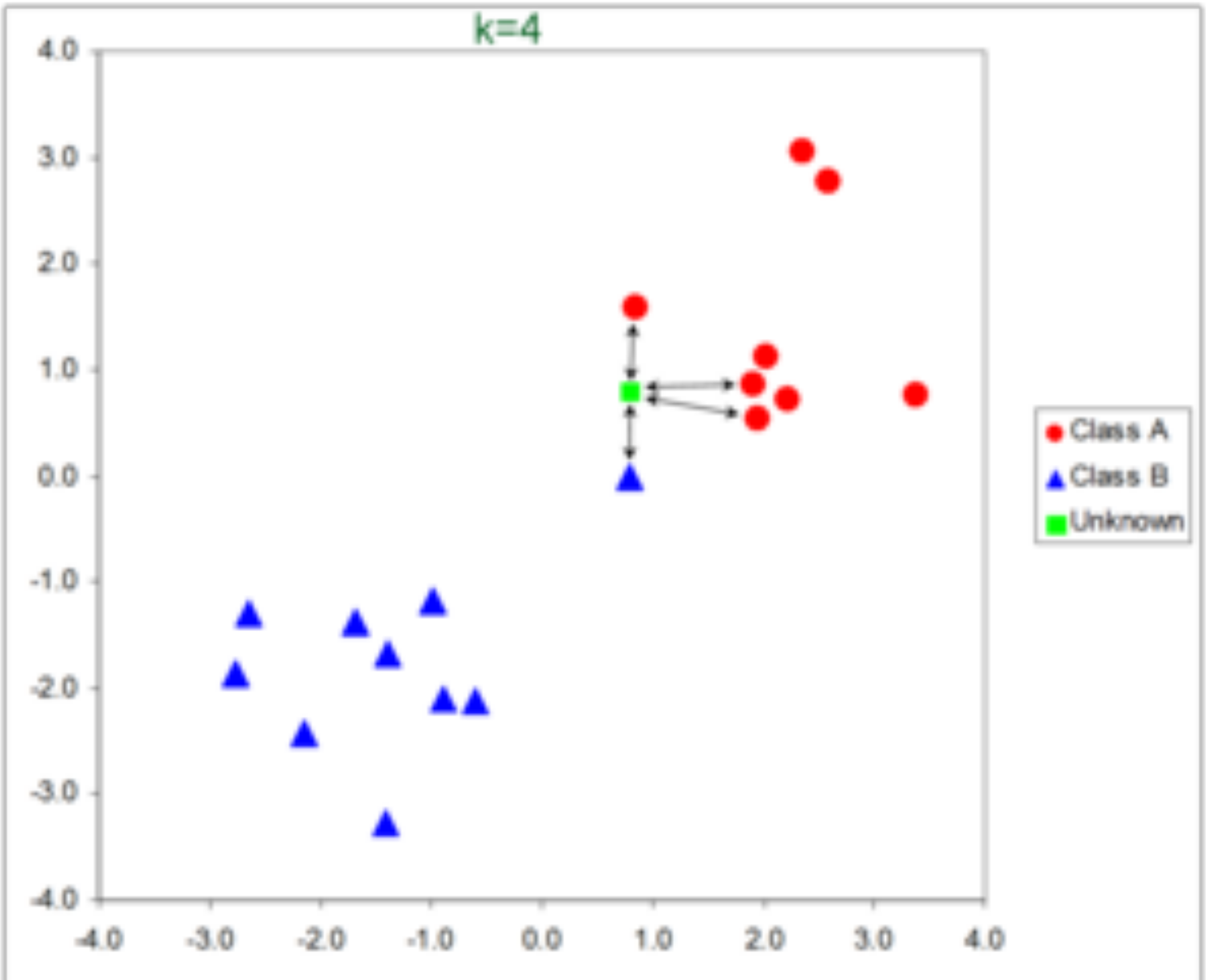
Pros	Cons
Can learn more complicated class boundaries	Hard to implement: trial and error for choosing parameters and network structure
Can be more accurate	Slow training time
Can handle large number of features	Can over-fit the data: find patterns in random noise
	Hard to interpret

K-NEAREST NEIGHBOR (KNN)

- One of the first choices for a classification study when there is little or no prior knowledge about the distribution of the data.
- K: hyper parameter
- A new data point is assigned the class of the plurality of its nearest neighbors in the training set, considering the nearest k neighbors
- K=3: Purple
- K=5: Yellow



K-NEAREST NEIGHBOR

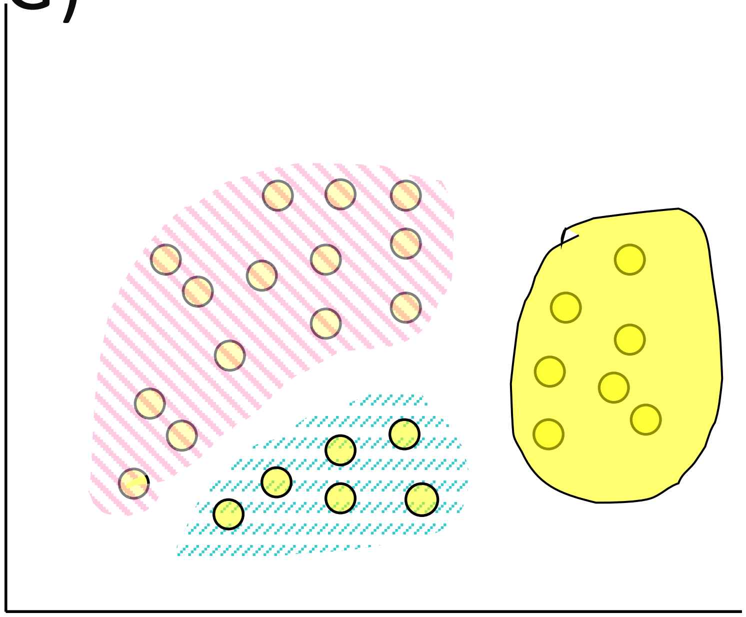


OUTLINE

- Machine learning overview
- Data mining overview
- Data mining tasks
  - Classification (supervised learning)
  - **Clustering (unsupervised learning)**

CLUSTERING
(UNSUPERVISED LEARNING)

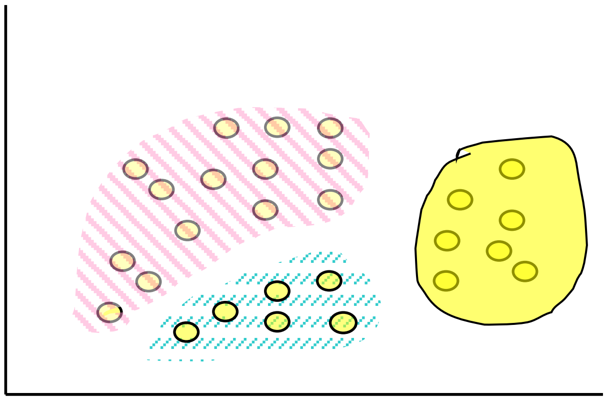
- What we have
  - a set of un-labeled data points, each with a set of attributes
  - a similarity measure
- What we need
  - find "natural" partitioning of data, or groups of similar/close items



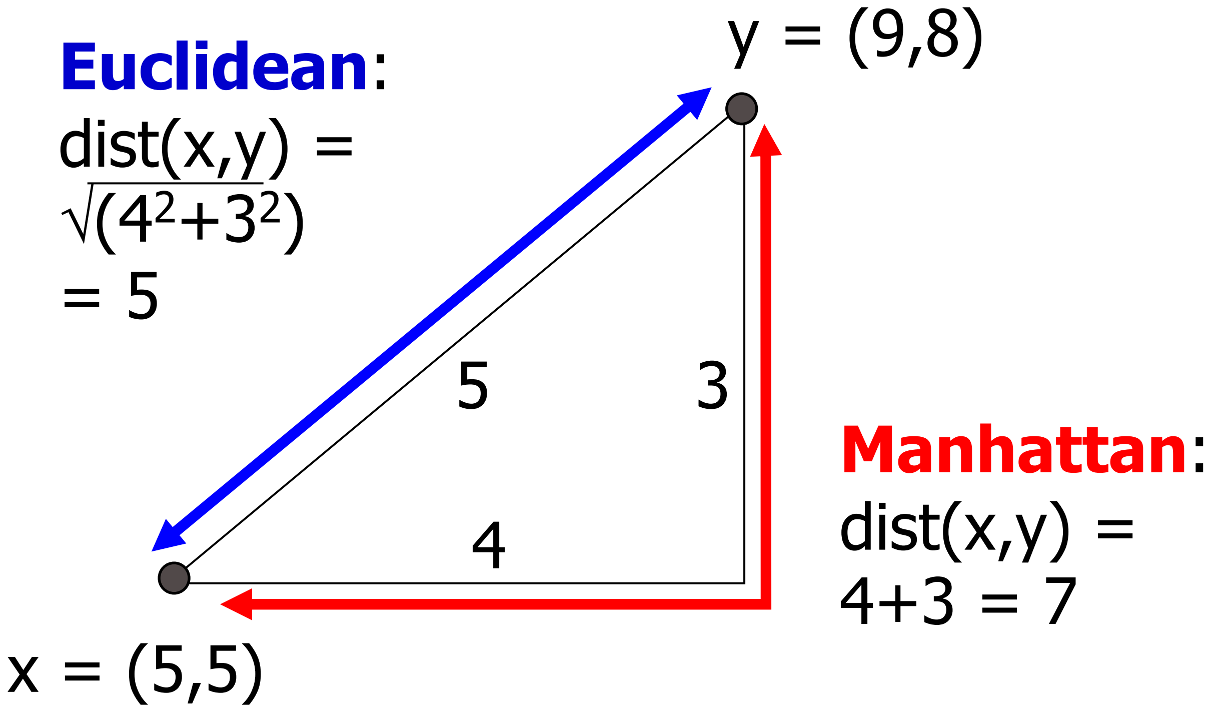
- The task of clustering is to partition the data so the instances are grouped in similar items by using distance/similarity measure

CLUSTERING
(UNSUPERVISED LEARNING)

- A set of data points, each with a set of attributes and a similarity measure, find clusters such that
  - Data points in one cluster are more similar
  - Data points in separate clusters are less similar to one another
- Key: measure of similarity between instances
  - Euclidean or Manhattan distance
  - Hamming distance
  - Other problem specific measures

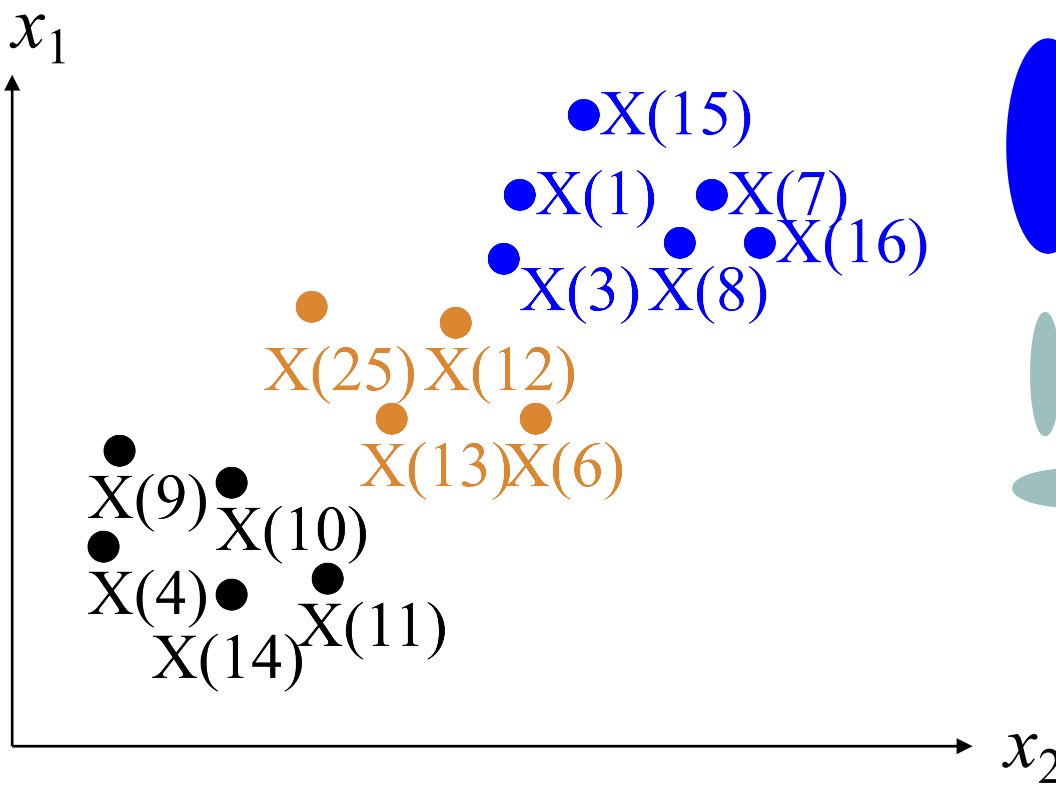


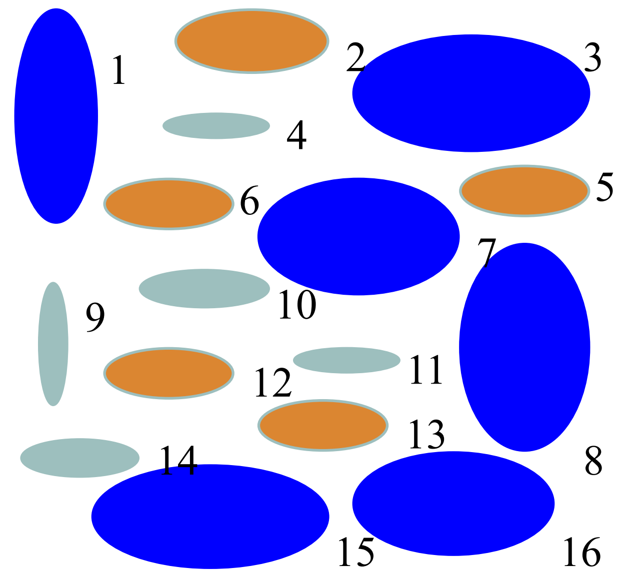
MANHATTAN & EUCLIDEAN DISTANCES



CLUSTERING

- Find “natural” grouping of instances given un-labeled data

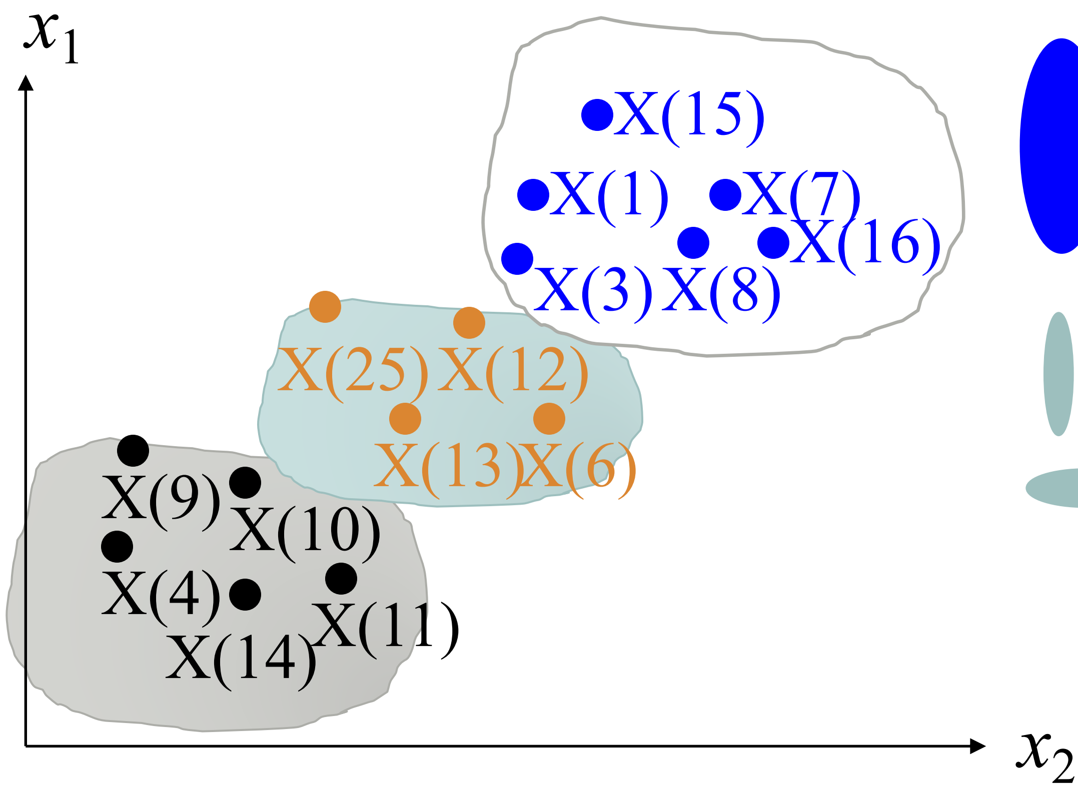


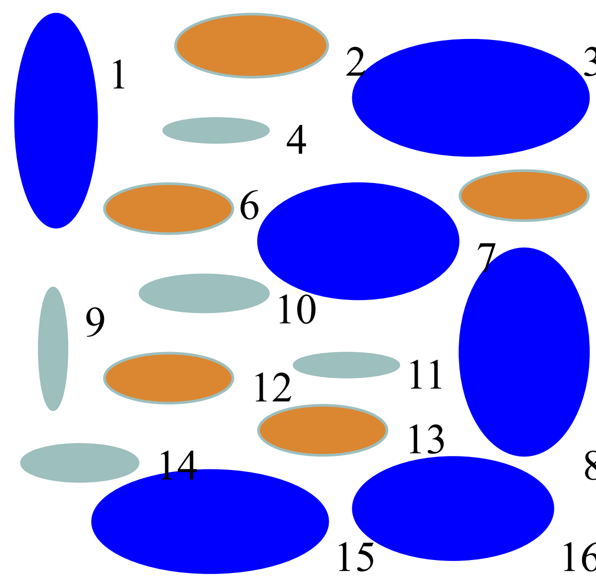


Elliptical blobs (objects)

CLUSTERING

- Find "natural" grouping of instances given un-labeled data





Elliptical blobs (objects)

CLUSTERING: METHODS

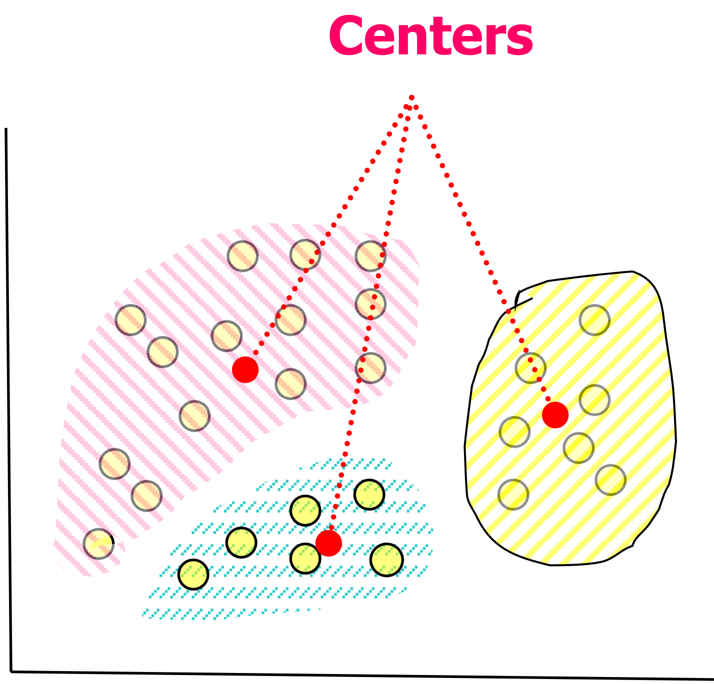
- Partitioning-based clustering
  - **K-means clustering**
  - K-medoids clustering

- Density-based clustering
  - Separate regions of dense points by sparser regions of relatively low density

PARTITIONING-BASED CLUSTERING: K-MEANS

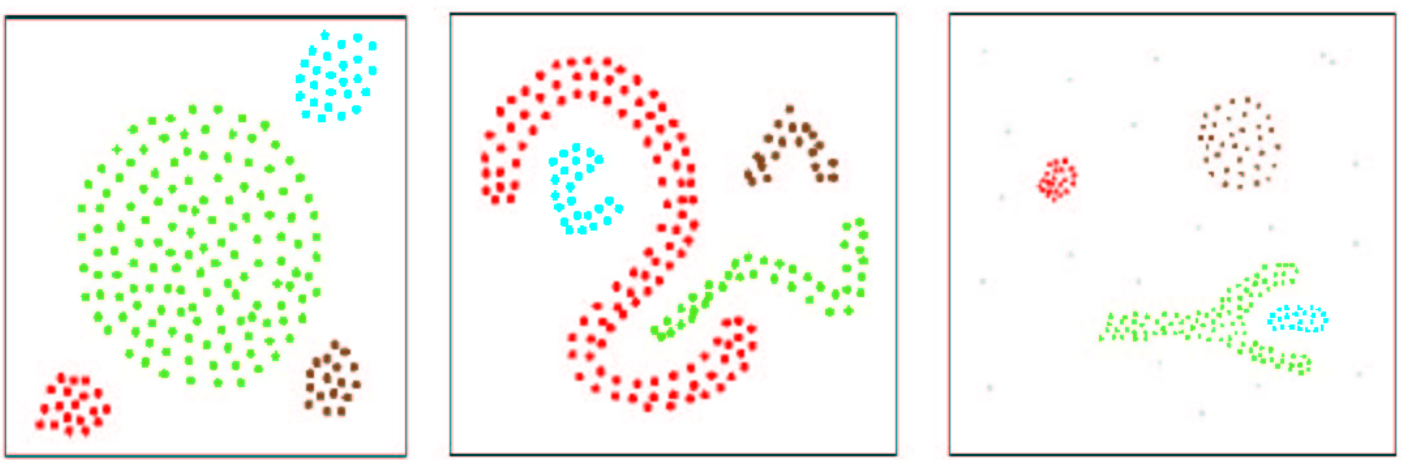
- Goal: minimise sum of square of distance
  - Between each point and centers of the cluster.
  - Between each pair of points in the cluster

- Algorithm:
  - Initialize K cluster centers
    - random, first K, K separated points
  - Repeat until stabilization:
    - Assign each point to closest cluster center
    - Generate new cluster centers
    - Adjust clusters by merging or splitting



DENSITY-BASED CLUSTERING

- **A cluster:** a connected dense component
- **Density:** the number of neighbors of a point
- Can find clusters of arbitrary shape



CLUSTERING: APPLICATIONS

Market Segmentation

- **Goal**: divide a market into distinct subsets of customers, any subset may be a market target
- **Approach**
  - Collect different attributes of customers, based on their related information (lifestyle, etc.)
  - Find clusters of similar customers
  - Evaluate buying patterns in the same cluster vs. those from different clusters

FURTHER READING

- Chapter 18.1-18.4 AIMA (Learning from examples)
- Introduction to Machine Learning

http://studentnet.cs.manchester.ac.uk/ugt/COMP24111