A Few Examples of Machine Learning Algorithms

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences

Learning Objectives

By the end of this topic, you should be able to:

- Describe machine learning.
- Describe the major classes of ML methods.
- Describe how rule-based decision trees are constructed.
- Describe how KNN works.

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

What is Machine Learning?

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

What is Machine Learning (ML)?

"Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed."

--- Arthur Samuel (1959)

"A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."

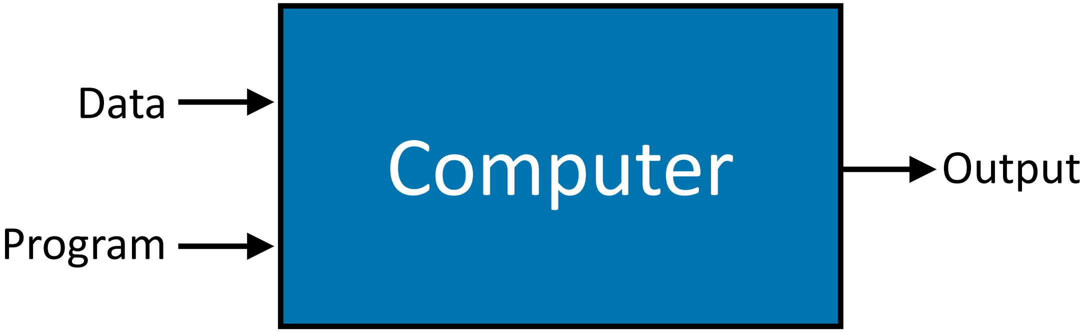
-- Tom Mitchell (1997)

ML solves complex problems that cannot be solved by numerical means alone.

What is Machine Learning (ML)?

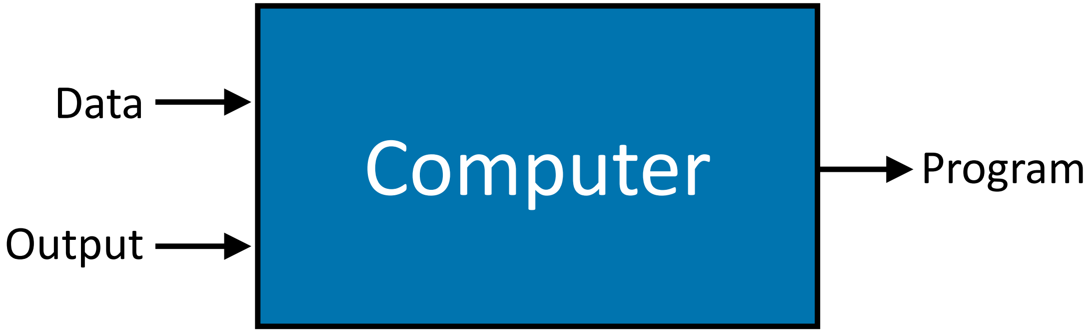
AKA Human-guided analysis

Traditional Programming



AKA Data/Machine-guided analysis

Machine Learning

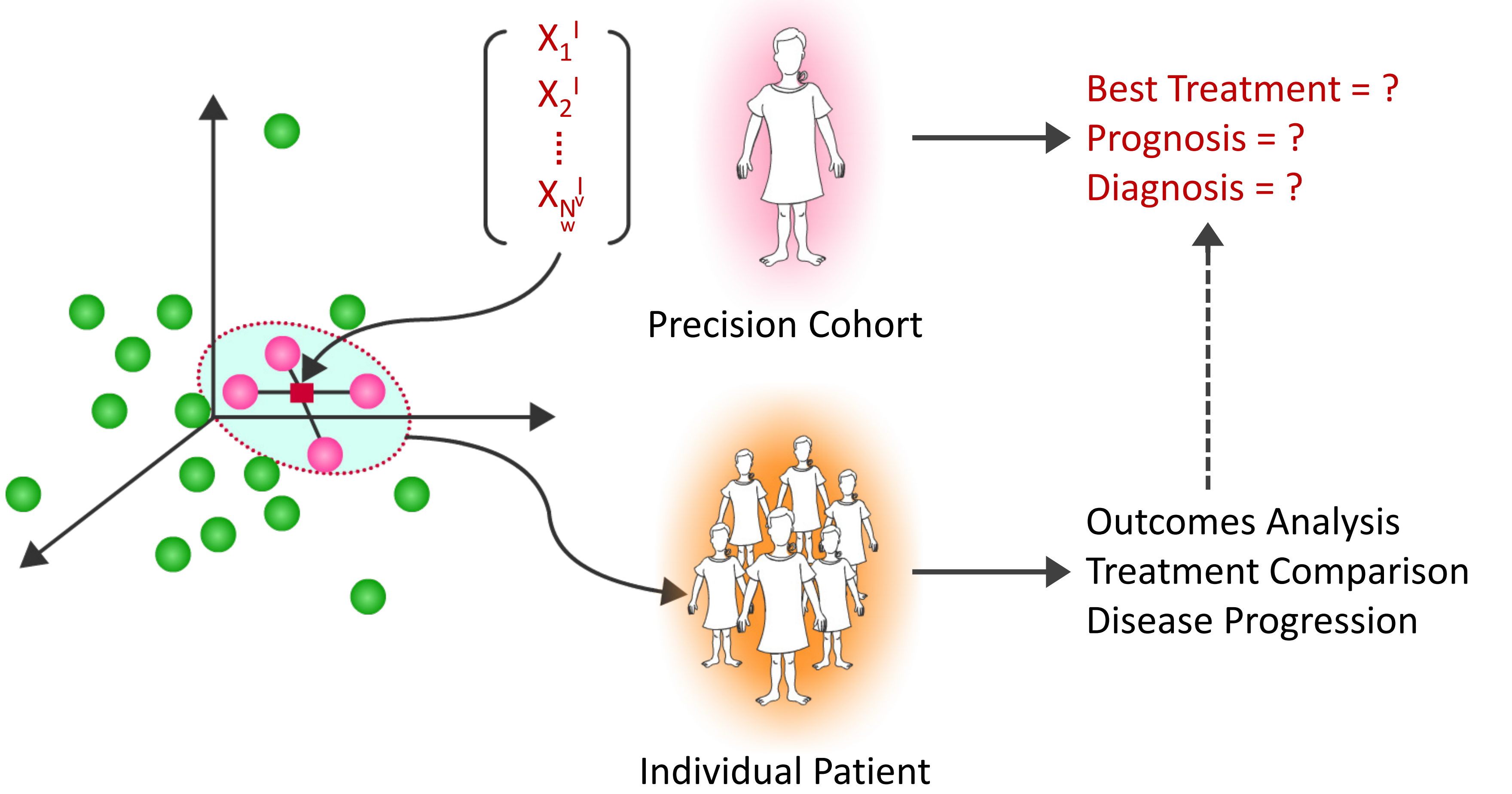


The output is the sample class or the variable we want to predict.

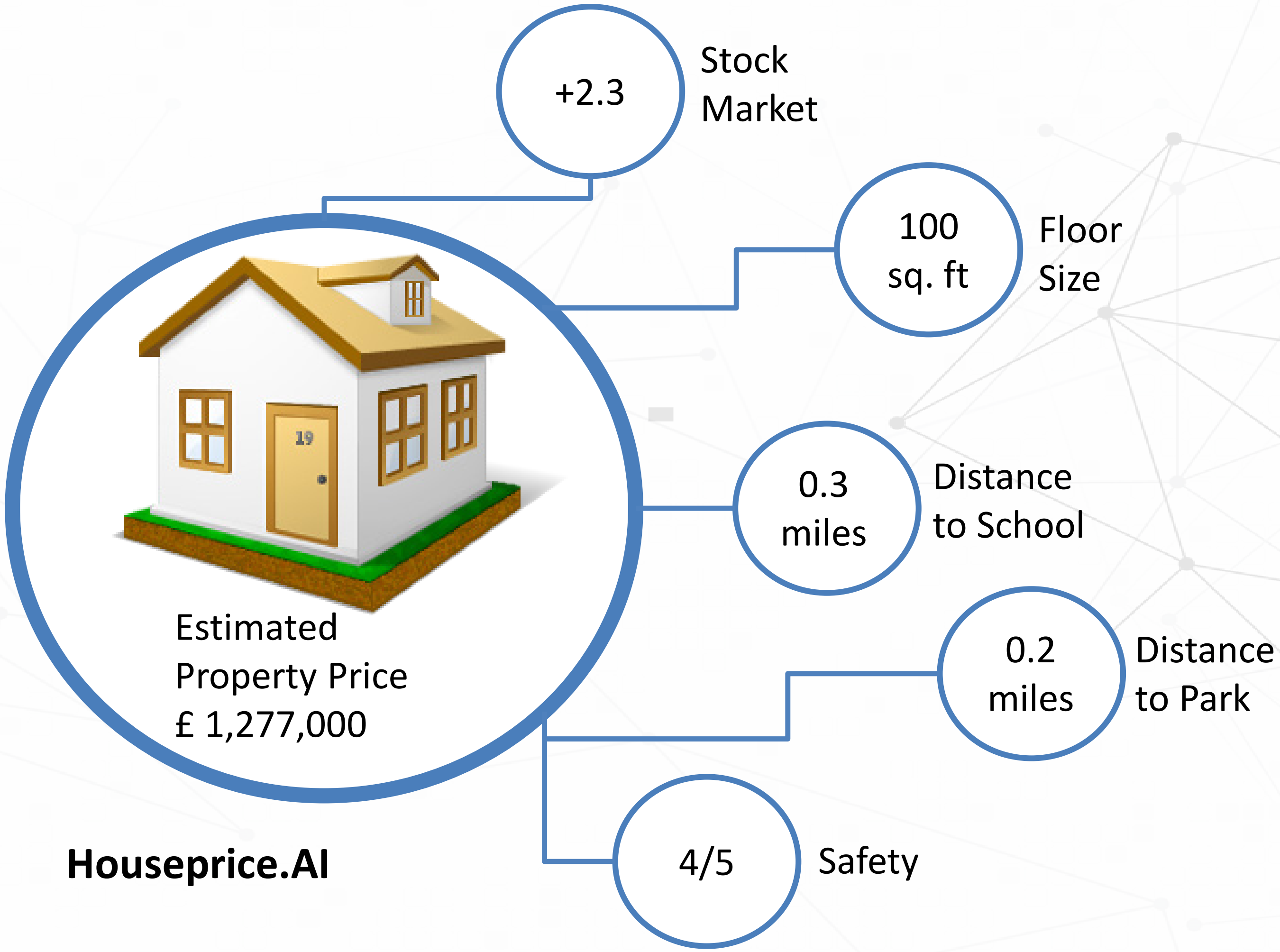
Suitable Problems for ML

- The highly complex nature of many real-world problems, though, often means that inventing specialised algorithms that will solve them perfectly every time is impractical, if not impossible.
- Examples of machine learning problems include, “Will this patient die from this cancer?”, “What is the market value of this house?”.

Will this patient die from this cancer?



What is the market value of this house?



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

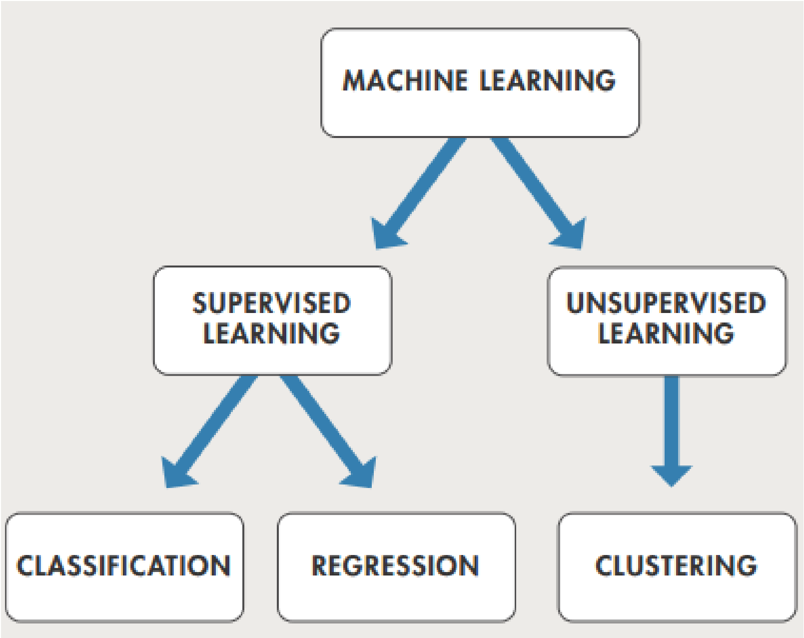
Overview of Machine Learning

BS0004 Introduction to Data Science

Dr Wilson Goh

School of Biological Sciences

Supervised and Unsupervised ML



Supervised machine learning:

- **Classification machine learning systems:** guess the class (e.g. survive or die).
- **Regression:** guess the value Y when $X_1..X_n$ is observed.

Unsupervised machine learning: The program is given data and must find patterns and relationships therein **without** explicitly using class information (output).

- **Clustering**: Group together samples that are more similar to one another (then check for corroboration with output/class).

Supervised Learning (Classification)

- **Learn from past experience, and use the learned knowledge to classify new data.**
- **Knowledge learned by intelligent algorithms.**
- **Examples:**
  - Clinical diagnosis for patients
  - Cell type classification

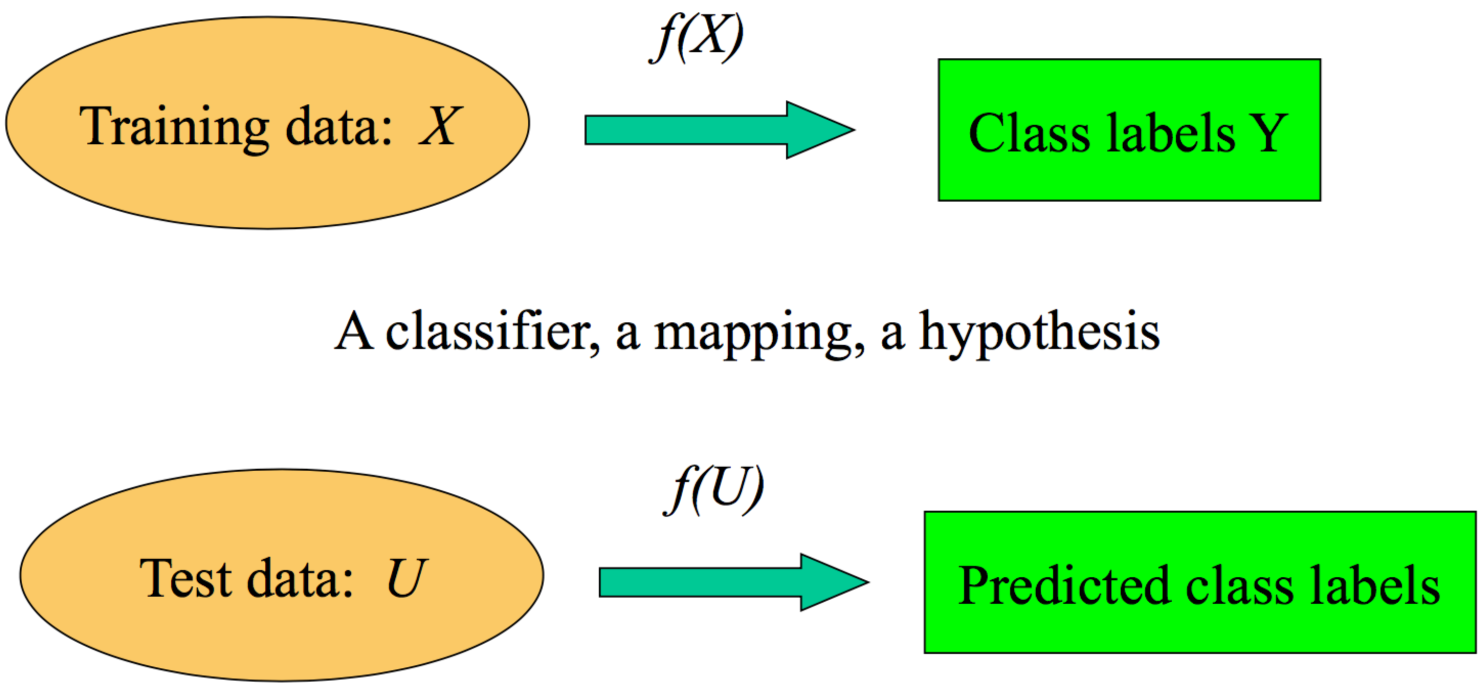
- **Classification involves > 1 class of data. E.g.,** Normal vs disease cells for a diagnosis problem.
- **Training data is a set of instances (samples, points, etc.) with known class labels.**
- **Test data is a set of instances whose class labels are to be predicted.**

Some Notation

- Training data:
  $\{<x_1, y_1>, <x_2, y_2>, \ldots, <x_m, y_m>\}$
  - where x_j are n-dimensional vectors and y_j are from a discrete space Y. E.g., Y = {normal, disease}.

- Test data:
  $\{<u_1, ?>, <u_2, ?>, \ldots, <u_k, ?>\}$
  - Where u_k is an n-dimensional vector and ? are the classes to be predicted.

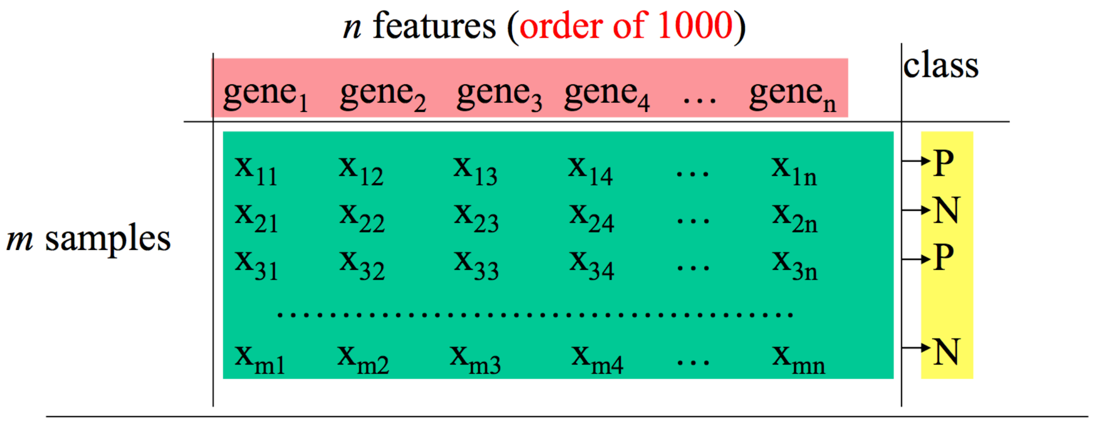
Process



f(X)
Training data: X
Class labels Y
A classifier, a mapping, a hypothesis
f(U)
Test data: U
Predicted class labels

Relational Data Representation (X and Y)

X is $gene_1 \ldots gene_n$



Class = Y

Which sources of big biological data are amendable to this? Genomics, Transcriptomics, RT-PCR, Proteomics or combinations of these.

Variables/ Features

- **Categorical features (Nominal/ Ordinal)**
  - Colour = {red, blue, green}
- **Continuous or numerical features (Interval/ Ratio)**
  - Gene Expression
  - Age
  - Blood Pressure

Data Example

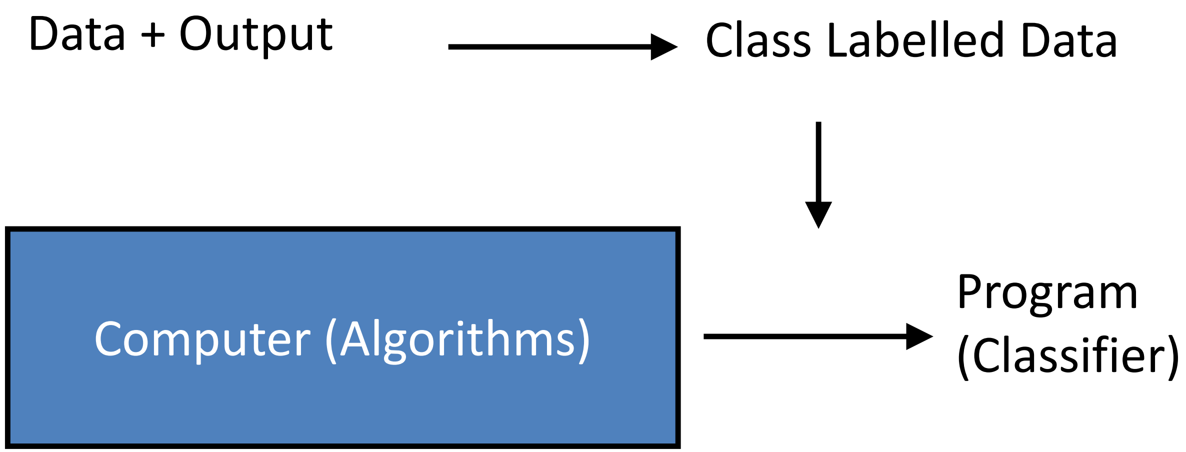
Each column is a variable

Each row is a Sample

Outlook	Temp	Humidity	Windy	Class
Sunny	75	70	True	Play
Sunny	80	90	True	Don’t
Sunny	85	85	False	Don’t
Sunny	72	95	True	Don’t
Sunny	69	70	False	Play
Overcast	72	90	True	Play
Overcast	83	78	False	Play
Overcast	64	65	True	Play
Overcast	81	75	False	Play
Rain	71	80	True	Don’t
Rain	65	70	True	Don’t
Rain	75	80	False	Play
Rain	68	80	False	Play
Rain	70	96	False	Play

Categorical Continuous Categorical

Supervised Learning (Global View)



How do you know if your predictions are good?

- **Many measures:**
  - Accuracy, error rate, false positive rate, false negative rate, sensitivity, specificity, precision.
- **K-fold cross validation:**
  - Given a dataset, divide it into k even parts, k-1 of them are used for training, and the rest one part treated as test data.
- **Independent validation (Performance on independent blind test data):**
  - Blind test data properly represent real world.

Requirements of a Good Classifier

- High accuracy, sensitivity, specificity and precision (Is this truly possible?).
- High comprehensibility.

What determines good performance?

Good Quality Data	Suitable Classification Algorithm	Meaningful Variables	Supervised or Unsupervised
Clean			Regression Classification

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Decision Trees

BS0004 Introduction to Data Science

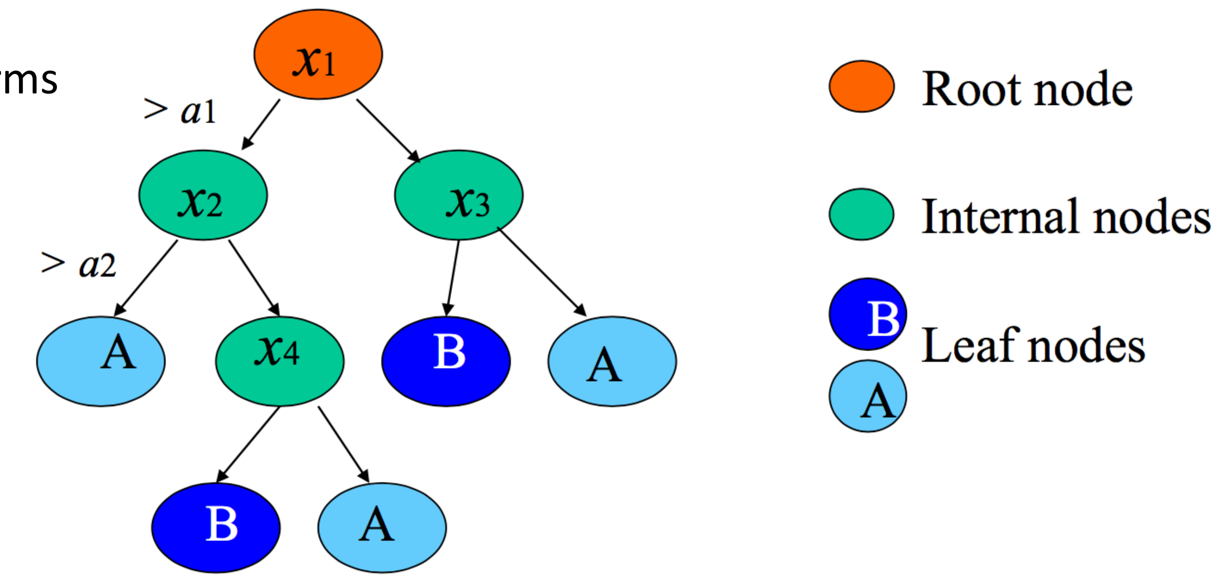
Dr Wilson Goh
School of Biological Sciences

Decision Trees

- A group of rule-based methods useful for classification.
- Systematic selection/ ordering of a small number of features used for the decision making.
- This increases comprehensibility of the knowledge patterns (tells us which variables are the most important).

Structure of Decision Trees

Every path from root to a leaf forms a **decision rule**.



- If $x_1 > a_1$ & $x_2 > a_2$, then it's class A.
- Easy interpretation, but accuracy may be unattractive.

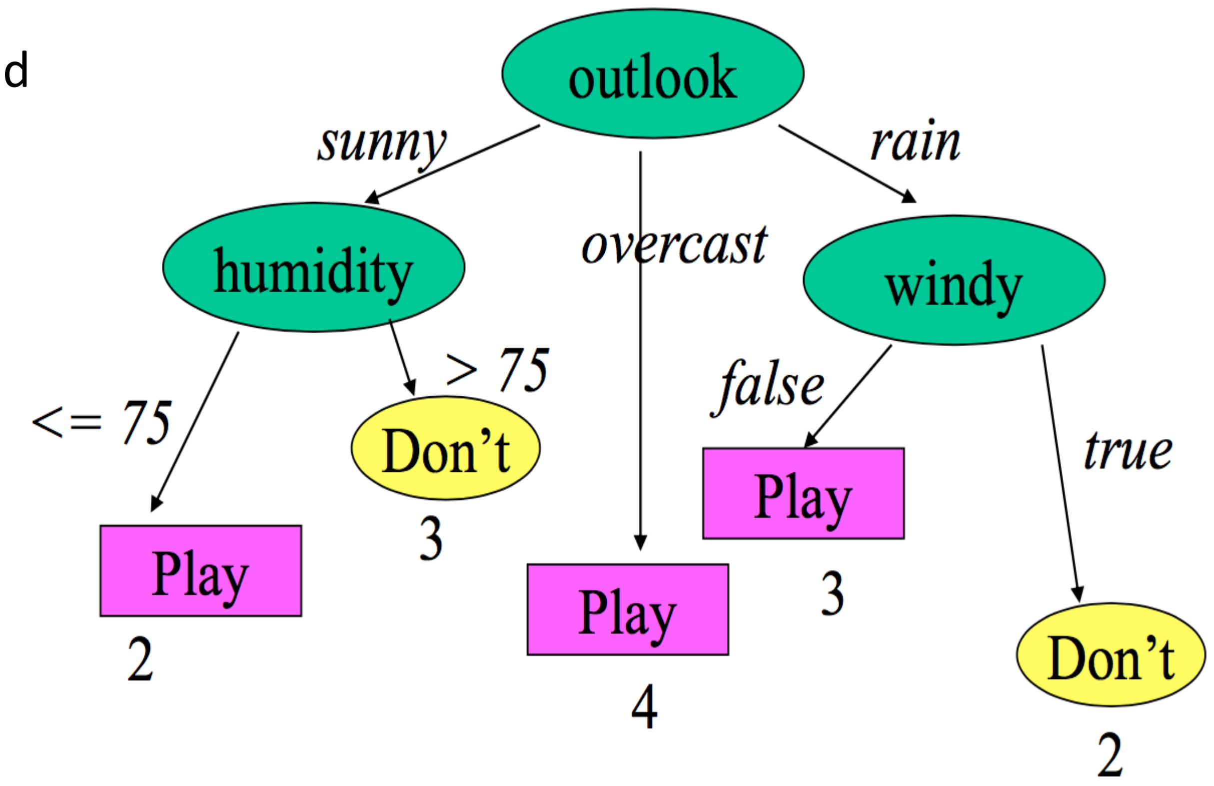
Decision Tree Example

Outlook	Temp	Humidity	Windy	Class
Sunny	75	70	True	Play
Sunny	80	90	True	Don’t
Sunny	85	85	False	Don’t
Sunny	72	95	True	Don’t
Sunny	69	70	False	Play
Overcast	72	90	True	Play
Overcast	83	78	False	Play
Overcast	64	65	True	Play
Overcast	81	75	False	Play
Rain	71	80	True	Don’t
Rain	65	70	True	Don’t
Rain	75	80	False	Play
Rain	68	80	False	Play
Rain	70	96	False	Play

A total of 14 outcomes:
9 Play
5 Don’t Play

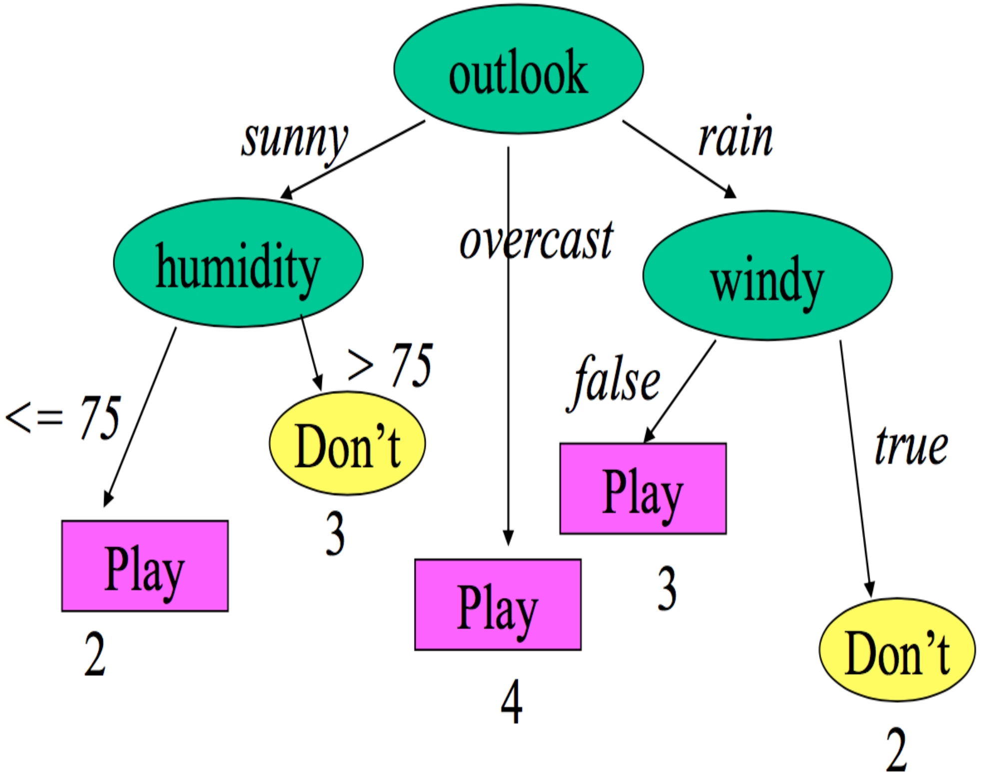
Decision Tree Example

Construction of a tree is equivalent to determination of root node of the tree and root nodes of its sub-trees.



Decision Tree Example

Outlook	Temp	Humidity	Windy	Class	Predicted	Verdict
Sunny	75	70	True	Play	Play	TP
Sunny	80	90	True	Don’t	Don’t	TN
Sunny	85	85	False	Don’t	Don’t	TN
Sunny	72	95	True	Don’t	Don’t	TN
Sunny	69	70	False	Play	Play	.
Overcast	72	90	True	Play	Play	.
Overcast	83	78	False	Play	Play	.
Overcast	64	65	True	Play	Play	.
Overcast	81	75	False	Play	Play	.
Rain	71	80	True	Don’t	Don’t	.
Rain	65	70	True	Don’t	Don’t	.
Rain	75	80	False	Play	Play	.
Rain	68	80	False	Play	Play	.
Rain	70	96	False	Play	Play	.



Most Discriminatory Variable

- Every variable can be used to partition the training data e.g., “Play and Don’t Play”.
- If the partitions contain at least 1 pure class of training instances, then this variable is most certainly discriminatory.

Partitions

- Categorical feature:
  - Number of partitions of the training data is equal to the number of values of this feature e.g. Number of partitions {Play, Don’t Play} = 2.
- Numerical feature:
  - Two partitions based on some threshold e.g. A > 100 (splits into values which are greater than 100 or otherwise).

Data Example

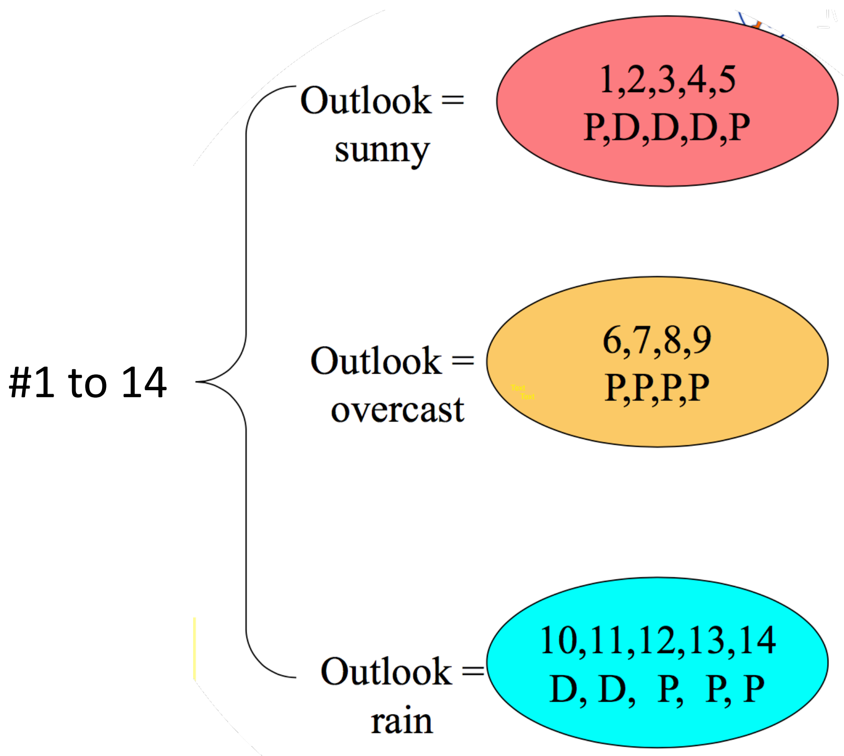
Each column is a variable

Each row is a Sample

Outlook	Temp	Humidity	Windy	Class
Sunny	75	70	True	Play
Sunny	80	90	True	Don’t
Sunny	85	85	False	Don’t
Sunny	72	95	True	Don’t
Sunny	69	70	False	Play
Overcast	72	90	True	Play
Overcast	83	78	False	Play
Overcast	64	65	True	Play
Overcast	81	75	False	Play
Rain	71	80	True	Don’t
Rain	65	70	True	Don’t
Rain	75	80	False	Play
Rain	68	80	False	Play
Rain	70	96	False	Play

Categorical　　Continuous　　Categorical

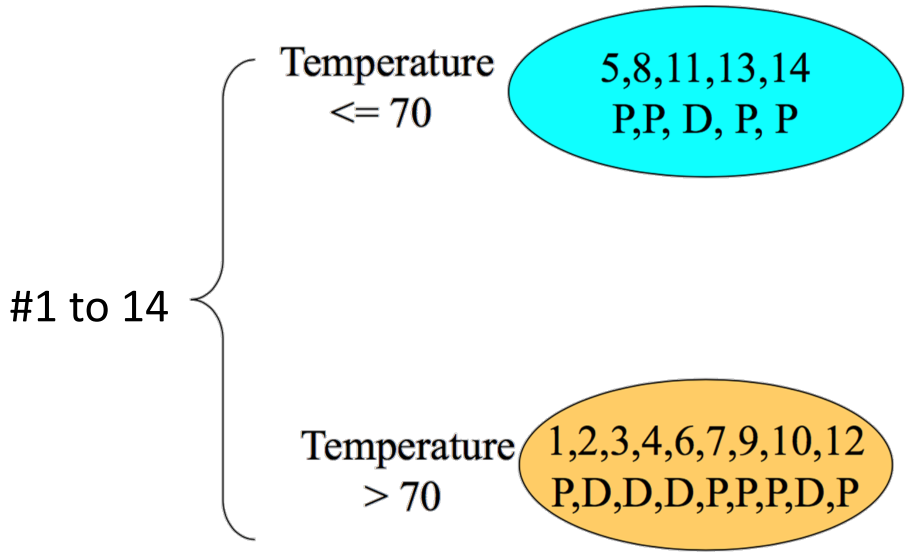
Partitioning Variables



A categorical feature is partitioned based on its number of possible values.

Partitioning Variables

A numerical feature is generally partitioned by choosing a “cutting point”.



Decision Tree Construction

1. Select the "best" feature as root node of the whole tree.
2. Partition dataset into subsets using this feature so that the subsets are as "pure" as possible.
3. After partition by this feature, select the best feature (with respect to the subset of training data) as root node of this sub-tree.
4. Recursively, until the partitions become pure or almost pure.

Let’s Construct a Decision Tree

Outlook	Temp	Humidity	Windy	Class
Sunny	75	70	True	Play
Sunny	80	90	True	Don’t
Sunny	85	85	False	Don’t
Sunny	72	95	True	Don’t
Sunny	69	70	False	Play
Overcast	72	90	True	Play
Overcast	83	78	False	Play
Overcast	64	65	True	Play
Overcast	81	75	False	Play
Rain	71	80	True	Don’t
Rain	65	70	True	Don’t
Rain	75	80	False	Play
Rain	68	80	False	Play
Rain	70	96	False	Play

Gini Coefficient

- Gini Index or coefficient can be used as an approximation of the power of a variable.
  - Split is completely pure, Gini index = 0
  - Split is impure, max Gini index = 1 – 1/k (where k = number of class levels)

$$Gini = \sum_{i \neq j} p(i)p(j)$$

i and j are levels of the target variable

- The sum of the joint probabilities of all impure combinations.
- Minimum value of Gini Index will be 0 when all observations belong to one class label.

Gini Coefficient

Suppose we have class label with 2 levels -> Normal (N) and Cancer (C). There are 4 possible permutations.

1	2	3	4
Normal	Cancer	Cancer	Normal
Normal	Cancer	Normal	Cancer

P(Class=N).P(Class=N) + P(Class=C).P(Class=C) + P(Class=C).P(Class=N) + P(Class=N).P(Class=C) = 1

P(Class=N).P(Class=C) + P(Class=C).P(Class=N) = 1 - P(Class=N).P(Class=N) - P(Class=C).P(Class=C)

P(Class=N).P(Class=C) + P(Class=C).P(Class=N) = 1 – P^2(Class=N) - P^2(Class=C)

Maximum value of Gini Index = 1 – (P^2(Class=N) + P^2(Class=C))

Maximum value of Gini Index = $1 - \sum_{t=0}^{t=k} P_t^2$

Where t is the class, and k are attributes of class (N and C).

Gini Coefficient

1	2	3	4
Normal	Cancer	Cancer	Normal
Normal	Cancer	Normal	Cancer

- Max Gini Index value = $1 - (1/2)^2 - (1/2)^2 = 1 - 2*(1/2)^2 = 1 - 2*(1/4) = 1 - 0.5 = 0.5$
- Similarly for Nominal variable with k level, the maximum value Gini Index is= 1 - 1/k.
- Since the play data has 2 levels (play and don't play), its max Gini index is also 0.5.
- However, knowing the min and max Gini coefficients don't tell us what is the quality of a split given a variable.

Gini Coefficient of a Split

$GINI\ (s,t) = GINI\ (t) - P_L\ GINI\ (t_L) - P_R\ GINI\ (t_R)$

where

s: split

t: node

GINI (t): Gini Index of input node t

P_L: Proportion of observation in Left Node after split, s

GINI (t_L): Gini of Left Node after split, s

P_R: Proportion of observation in Right Node after split, s

GINI (t_R): Gini of Right Node after split, s

Gini Coefficient of Outlook

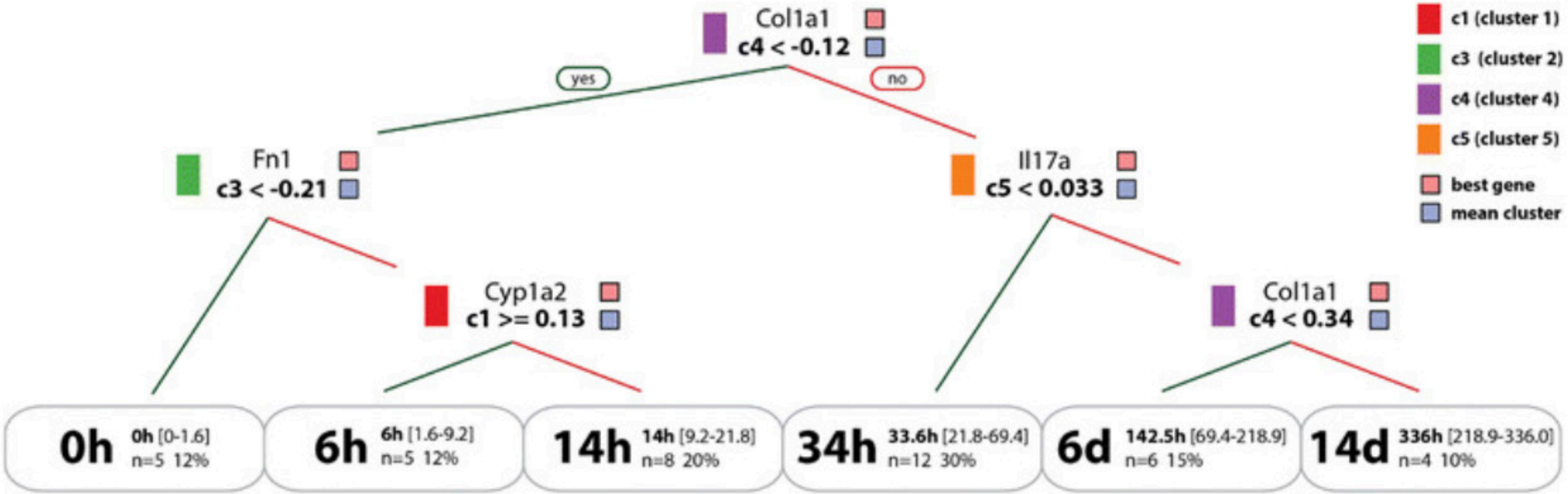
- GINI (t) ≠ 0.5
- GINI (t) = $1 - (5/14)^2 - (9/14)^2 = 0.46$ [the distribution between classes is not equal!]
- Gini (Sunny) = $1 - (2/5)^2 - (3/5)^2 = 0.48$
- Gini (Overcast) = $1 - (4/4)^2 - (0/4)^2 = 0$
- Gini (Rain) = $1 - (3/5)^2 - (2/5)^2 = 0.48$
- Gini (Outlook) = $0.46 - (5/14 * 0.48 + 4/14 * 0 + 5/14 * 0.48) = 0.46 - 0.34 = 0.12$

#note that Gini (overcast) is a pure sub-cluster

#Try doing Gini (Windy) and Gini (Humidity <= 75) yourself

Outlook	Temp	Humidity	Windy	Class
Sunny	75	70	True	Play
Sunny	80	90	True	Don't
Sunny	85	85	False	Don't
Sunny	72	95	True	Don't
Sunny	69	70	False	Play
Overcast	72	90	True	Play
Overcast	83	78	False	Play
Overcast	64	65	True	Play
Overcast	81	75	False	Play
Rain	71	80	True	Don't
Rain	65	70	True	Don't
Rain	75	80	False	Play
Rain	68	80	False	Play
Rain	70	96	False	Play

Decision Tree in Action



- When considering high-throughput data with thousands of variables, the split rules are often not so clear. In this case.
- A "representative best gene" is shown at the top but these are by no means exhaustive (there can be many equivalent best genes at each level) nor does the selection of best genes necessarily mean anything biologically beyond prediction value.

Decision Trees

Advantages

- Single coverage of training data (elegance).
- Divide-and-conquer splitting strategy (simple).
- Rules are obvious (understandable).

Disadvantages

- Fragmentation problem => Locally reliable but globally insignificant rules.
- Miss many globally significant rules.
- Mislead system.

Some Examples of Use of Decision Trees in Biological data

- In prostate and bladder cancers (Adam et al. Proteomics, 2001).
- In serum samples to detect breast cancer (Zhang et al. Clinical Chemistry, 2002).
- In serum samples to detect ovarian cancer (Petricoin et al. Lancet; Li & Rao, PAKDD 2004).

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

K-Nearest Neighbours

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

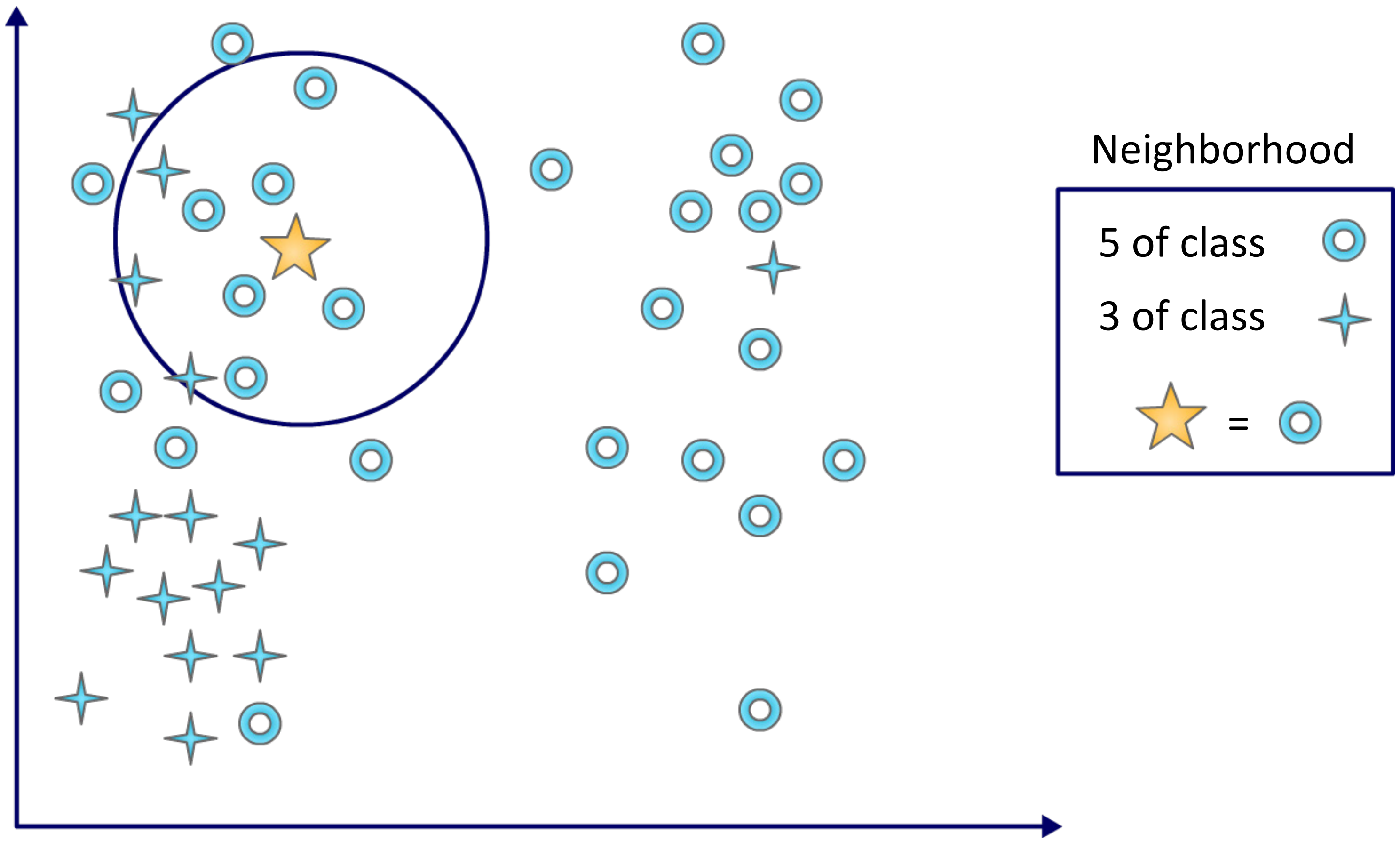
K-Nearest Neighbours (kNN)

Given a new case:

- Find k "nearest" neighbours, i.e., k most similar points in the training data set.
- Assign new case to the same class to which most of these neighbours belong.
- A common "distance" measure between samples x and y is $\sqrt{\sum_f (x[f] - y[f])^2}$.
- Where f ranges over variables of the samples.

Illustration of kNN (k=8)

What should the class of ☆ be?

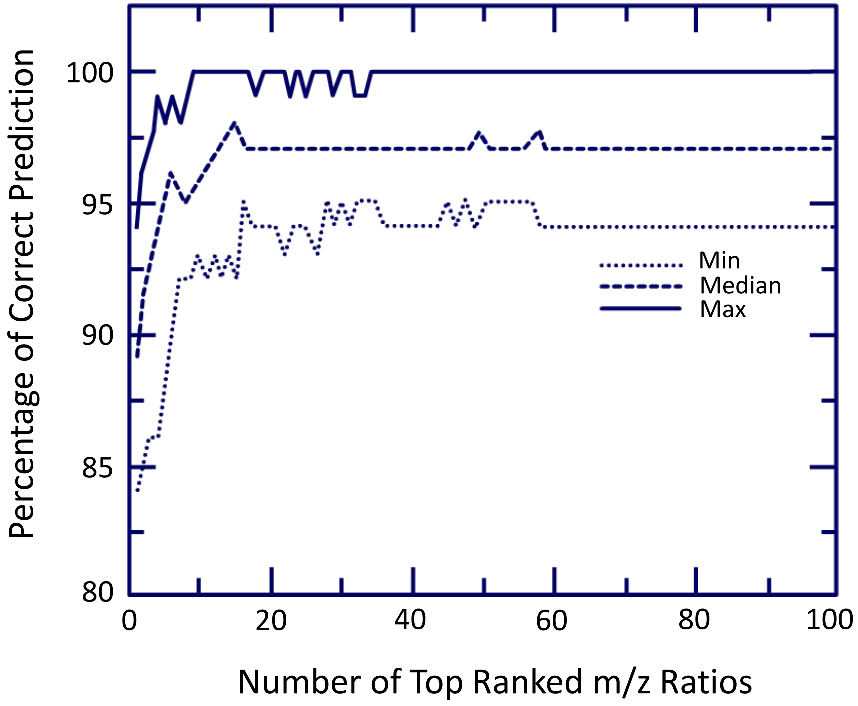


Some Issues

- Simple to implement.
- Must compare new case against all training cases.
  - May be slow during prediction.
- No need to train.
- But need to design distance measure properly.
  - May need expert for this.
- Can't explain prediction outcome.
  - Can't provide a model of the data.

Example Use of kNN

- Li et al, *Bioinformatics* 20:1638-1640, 2004.
  - Use kNN to diagnose ovarian cancers using proteomic spectra.
  - Data set is from Petricoin et al., *Lancet* 359:572-577, 2002.



Minimum, median and maximum of percentages of correct prediction as a function of the number of top-ranked *m/z* ratios on 50 independent partitions into learning and validation sets.

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Summary

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Key Takeaways from this Topic

1. Machine learning methods can be broadly divided into supervised and unsupervised.

2. Decision trees are very comprehensive when variable size is small.

3. KNN is a simple machine learning approach.

4. Designing a machine learning analysis pipeline is very complex.