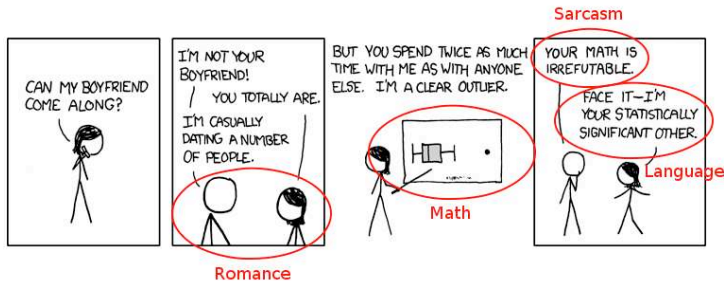


# Introduction to Machine Learning

Kee-Siong Ng  
keesiong.ng@nicta.com.au



## Outline

- Introduction
- Bayesian Probability Theory
- Sequence Prediction and Data Compression
- Decision Trees and Ensemble Learning
- Bayesian Networks

COMP3620 (2009)

1

## Outline

- Introduction
  1. What is machine learning?
  2. When is learning (not) possible?
- Bayesian Probability Theory
- Sequence Prediction and Data Compression
- Decision Trees and Ensemble Learning
- Bayesian Networks

COMP3620 (2009)

2

## What is Machine Learning?

From Wikipedia:

Machine learning is the subfield of AI that is concerned with the design and development of **algorithms** that allow computers to **improve their performance over time based on data**. A major focus of machine learning research is to automatically produce (induce) models, such as rules and patterns, from data.

In summary: It is about collecting data, analysing data, interpreting data, and acting on insights so obtained to achieve goals.

COMP3620 (2009)

3

## Inductive Learning from Observations

- Machine learning research covers a wide spectrum of activities.
- Focus on classification/prediction problems in this course.
- Informally, given a sequence of data

$$[(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)], \quad x_i \in X, y_i \in Y$$

generated by an (unknown) process, learn a function  $f : X \rightarrow Y$  that, given any  $x \in X$ , predicts the corresponding  $y$  value accurately.

## Example Problems


- Spam filtering

Data: [ (“Congratulations, you have won the jackpot! ...”, +),  
 (“Bankwest is conducting an account audit ...”, +),  
 (“Your COMP3620 assignment is late ...”, -), ... ]

Is “Congratulations on your COMP3620 assignment ...” a spam?

- Hand-written character recognition

Data: [ ( , 3), ( , 1), ( , 3), ( , 4), ( , 7), ... ]

What’s the label of ?

## Philosophical Concerns about Induction

- Hume (1748): How do you know the future will be like the past?
  - Inductive processes are inherently unjustifiable.
- Popper (1935): Scientific theories/conjectures are only falsifiable; they can never be confirmed for sure.
  - Competing theories cannot be compared.
- Need to define learning problems carefully.

## Learning (More Formally)

Given

- a sequence of data

$$D = [(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)], \quad x_i \in X, y_i \in Y,$$

where each  $(x_i, y_i)$  is drawn independently at random from an unknown probability distribution  $P$  on  $X \times Y$ ;

- a set  $F$  of functions from  $X \rightarrow Y$ ;

Learn an  $f \in F$  that minimises *expected error*

$$\mathbb{E}_{(x,y) \sim P} \mathbb{I}(y \neq f(x)). \quad \left( \sum_{(x,y)} P(x,y) \mathbb{I}(y \neq f(x)), \mathbb{I} : \text{indicator function} \right)$$

## Algorithmic Strategy

- We want  $f \in F$  that minimises  $\mathbb{E}_{(x,y) \sim P} \mathbb{I}(y \neq f(x))$ .
- But we don't know  $P$ !

## Algorithmic Strategy

- We want  $f \in F$  that minimises  $\mathbb{E}_{(x,y) \sim P} \mathbb{I}(y \neq f(x))$ .
- But we don't know  $P$ !
- Idea: Use data  $D$  as a proxy for  $P$ .
- Instead of seeking  $\arg \min_{f \in F} \mathbb{E}_{(x,y) \sim P} \mathbb{I}(y \neq f(x))$ , we seek  $f$  that minimises *empirical error*

$$\frac{1}{n} \sum_{(x,y) \in D} \mathbb{I}(y \neq f(x)). \quad (1)$$

- Intuitive justification: Law of large numbers

Fix  $f$ : (1)  $\longrightarrow \mathbb{E}_{(x,y) \sim P} \mathbb{I}(y \neq f(x))$  as  $|D| \longrightarrow \infty$

## Laws of Large Numbers

- A Bernoulli trial is an experiment that returns 1 with probability  $p$  and 0 with probability  $1 - p$ .
- If we do  $n$  such experiments, how many times  $k$  are we likely to get the result 1?
- Coin experiment

## Laws of Large Numbers

- A Bernoulli trial is an experiment that returns 1 with probability  $p$  and 0 with probability  $1 - p$ .
- If we do  $n$  such experiments, how many times  $k$  are we likely to get the result 1?
- (Bernoulli's Theorem) For any small error  $\epsilon$ , any small difference  $\delta$ , there is a number of trials  $N$  such that for any  $n > N$ ,

$$\Pr[(p - \epsilon) \leq k/n \leq (p + \epsilon)] > 1 - \delta$$

## Laws of Large Numbers

- Bernoulli's theorem doesn't tell us how fast  $k/n$  converges to  $p$ .
- Let  $b(k; n, p)$  be the probability of getting  $k$  occurrences of 1 in  $n$  Bernoulli trials.
- (Abraham De Moivre) A binomial distribution  $b(k; n, p)$  is approximated by a normal distribution with mean  $\mu = pn$  and standard deviation  $\sigma = \sqrt{(1-p)pn}$ .
- Thus, e.g., the probability that  $k$  is within  $2\sigma$  of  $pn$  is about 0.95.

$$\Pr[(pn - 2\sqrt{(1-p)pn}) \leq k \leq (pn + 2\sqrt{(1-p)pn})] = 0.95$$

## Laws of Large Numbers

- Bernoulli's theorem doesn't tell us how fast  $k/n$  converges to  $p$ .
- Let  $b(k; n, p)$  be the probability of getting  $k$  occurrences of 1 in  $n$  Bernoulli trials.
- (Abraham De Moivre) A binomial distribution  $b(k; n, p)$  is approximated by a normal distribution with mean  $\mu = pn$  and standard deviation  $\sigma = \sqrt{(1-p)pn}$ .
- Thus, e.g., the probability that  $k$  is within  $2\sigma$  of  $pn$  is about 0.95.

$$\begin{aligned} \Pr[(pn - 2\sqrt{(1-p)pn}) \leq k \leq (pn + 2\sqrt{(1-p)pn})] &= 0.95 \\ &= \Pr[(p - 2\sqrt{\frac{(1-p)p}{n}}) \leq k/n \leq (p + 2\sqrt{\frac{(1-p)p}{n}})] = 0.95 \end{aligned}$$

## Minimisation Angst

- We still have a problem if the function class  $F$  is too large.
- Can have many functions that all minimise empirical error.
- Consider underlying distribution  $P$  defined as follows:

$P$  : Density  $([0, 1] \times \{0, 1\})$

$$P(x, y) = \begin{cases} 1 & \text{if } (x \geq 0.5 \wedge y = 1) \text{ or } (x < 0.5 \wedge y = 0) \\ 0 & \text{otherwise} \end{cases}$$

- We can't learn if  $F$  is the set of *all* functions.
- Functions that *remember* the data achieve zero empirical error but incur large expected error!

## A Sufficient Condition for Successful Learning

- A sufficient (but far from necessary) condition for successful learning is  $|F| < \infty$  and  $|F| \ll |X|$ .
- Let's look at a simple proof.
- Further assumptions:
  - There is an unknown distribution  $Q$  on  $X$ .
  - There is an underlying true function  $t^* \in F$ .
  - Distribution  $P$  on  $X \times Y$  is defined by

$$P(x, y) = \begin{cases} Q(x) & \text{if } y = t^*(x) \\ 0 & \text{otherwise} \end{cases}$$

## Learnability Result

Define

$$\begin{aligned} F_{bad} &= \{h \in F \mid \mathbb{E}_{(x,y) \sim P} \mathbb{I}(h(x) \neq y) - \mathbb{E}_{(x,y) \sim P} \mathbb{I}(t^*(x) \neq y) \geq \varepsilon\} \\ &= \{h \in F \mid \mathbb{E}_{(x,y) \sim P} \mathbb{I}(h(x) \neq y) \geq \varepsilon\}. \end{aligned}$$

Given  $n$  randomly drawn examples  $D = \{(x_i, y_i)\}_{i=1}^n$  from  $P$ ,

$$\begin{aligned} &\Pr(\exists h \in F_{bad} \text{ consistent with } D) \\ &= \Pr(\bigvee_{i=1}^{|F_{bad}|} h_i \text{ consistent with } D) \\ &= \sum_{i=1}^{|F_{bad}|} \Pr(h_i \text{ consistent with } D) \\ &= |F_{bad}|(1 - \varepsilon)^n \leq |F|(1 - \varepsilon)^n. \end{aligned}$$

## Learnability Result

We have

$$\Pr(\exists h \in F_{bad} \text{ consistent with } D) \leq |F|(1 - \varepsilon)^n.$$

We want  $|F|(1 - \varepsilon)^n \leq \delta$  for some small  $\delta$ .

Solving for  $n$  yields

$$n \geq \frac{1}{\varepsilon} \left( \ln \frac{1}{\delta} + \ln |F| \right). \quad (2)$$

This means if we have seen  $n$  examples satisfying (2) and pick a hypothesis  $h^* \in F$  that is consistent with the examples, then with probability  $1 - \delta$ ,

$$\mathbb{E}_{(x,y) \sim P} \mathbb{I}(h^*(x) \neq y) < \varepsilon.$$

## Learnability Result

- If we drop the assumption that  $t^* \in F$ , picking the  $h^* \in F$  that minimises error on data is still a good strategy.
- Examples needed:

$$n \geq \frac{1}{2\varepsilon^2} \left( \ln \frac{1}{\delta} + \ln |F| \right)$$

- Given data  $D = \{(x_i, y_i)\}_{i=1}^n$ , with probability  $1 - \delta$ ,

$$\mathbb{E}_{(x,y) \sim P} \mathbb{I}(h^*(x) \neq y) < \frac{1}{|D|} \sum_{i=1}^n \mathbb{I}(h^*(x_i) \neq y_i) + \varepsilon.$$

## Problems in Practice

- We may not be able to get the number of examples required.
- We may not be able to efficiently find the function that minimises empirical error.
- There could be multiple empirical-error minimising functions.
- The first and third problems can be solved using Bayesian probability theory, which we look at next.

## Lecture Review

- Learning is not always possible
- When learning is possible, minimising empirical error is a good strategy justifiable using laws of large numbers
- Questions?

