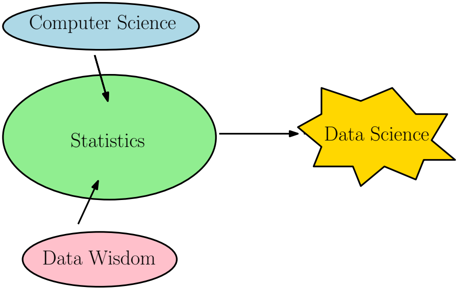
Housekeeping

- Lecture notes will be posted on edstem.
- One assignment due on 2 Nov, will be posted two weeks prior to the due date
- Any questions email rachel.wang@sydnedy.edu.au

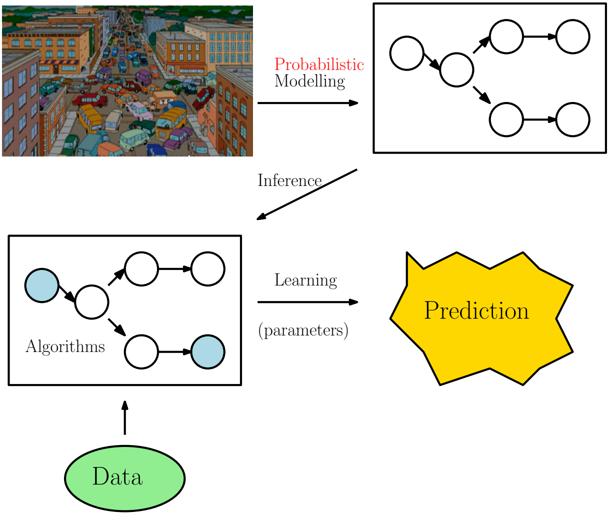
From Statistics to Data Science

- Statistics: distilling knowledge from data to solve real-world problems
  Models ⇒ Inference ⇒ Prediction
- Arrival of “Big Data”



Paradigm

The real world is complicated...

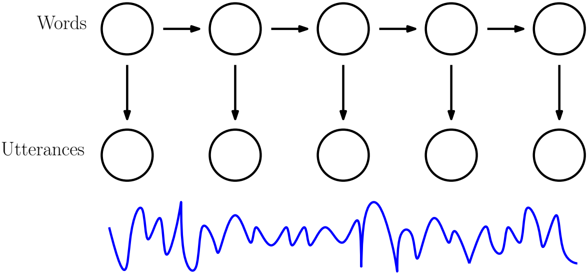


Why graphical models?

- Bridging two branches of mathematics: probability and graph theory
- Solving real world problems in bioinformatics, speech processing, image processing, artificial intelligence, and many others – modelling multiple random variables and their dependencies
- Connections to causality

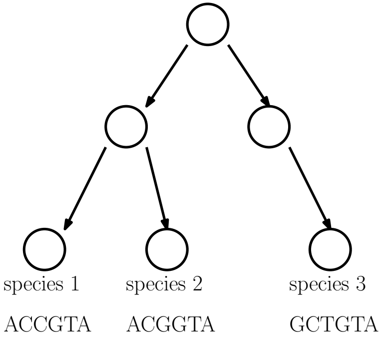
Some real world examples

Speech recognition



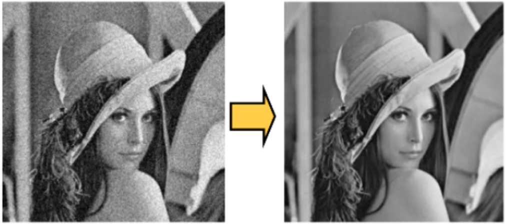
Some real world examples

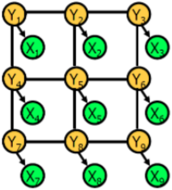
Phylogenetic tree



Some real world examples

Image denoising





X_i: noisy pixels
Y_i: “true” pixels

Plan of the class

- Some probability theory
- Directed graphs and joint probabilities
  - Representation
  - Conditional independence
- Undirected graphs and joint probabilities
  - Representation
  - Conditional independence
- Inference algorithms
  - Sum-product
  - Max-product
- More applications

Probability

Definitions

- A random variable X (discrete, finitely many values) takes values from a set $\{x_1, \ldots, x_r\} \subset \mathbb{R}$ depending on the outcomes of a random experiment. e.g.
  - number on a randomly thrown die
  - number of heads in 10 coin tosses

Probability

Definitions

- A random variable X (discrete, finitely many values) takes values from a set $\{x_1, \ldots, x_r\} \subset \mathbb{R}$ depending on the outcomes of a random experiment. e.g.
  - number on a randomly thrown die
  - number of heads in 10 coin tosses
- The distribution of X is characterised by the probability mass function (PMF) $p(x) := P(X = x)$. Note $0 \leq p(x) \leq 1$, and $\sum_x p(x) = 1$.

Probability

Definitions

- A random variable X (discrete, finitely many values) takes values from a set $\{x_1, \ldots, x_r\} \subset \mathbb{R}$ depending on the outcomes of a random experiment. e.g.
  - number on a randomly thrown die
  - number of heads in 10 coin tosses
- The distribution of X is characterised by the probability mass function (PMF) $p(x) := P(X = x)$. Note $0 \leq p(x) \leq 1$, and $\sum_x p(x) = 1$.
- Joint PMF of $(X_1, \ldots, X_n)$

$$p(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n).$$

Probability

Independence

- Two random variables X_1 and X_2 are independent, written $X_1 \perp\!\!\!\perp X_2$, iff the joint PMF factorises, that is $P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)P(X_2 = x_2)$ for all x_1 and x_2, or shorthand $p(x_1, x_2) = p(x_1)p(x_2)$.
  Intuitively, what does independence mean?

Probability

Independence

- Two random variables X_1 and X_2 are independent, written $X_1 \perp\!\!\!\perp X_2$, iff the joint PMF factorises, that is $P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)P(X_2 = x_2)$ for all x_1 and x_2, or shorthand $p(x_1, x_2) = p(x_1)p(x_2)$.
  Intuitively, what does independence mean?
- Changing single indices to subsets of indices, e.g. $A = \{2, 4\}$, $B = \{3\}$. $X_A \perp\!\!\!\perp X_B$ iff

$$p(x_A, x_B) = p(x_A)p(x_B)$$

Probability

Conditional independence

- The conditional PMF of X_1 given $X_2 = x_2$ is

$$
\begin{aligned}
p(x_1|x_2) :=& P(X_1 = x_1 | X_2 = x_2) \\
=& \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_2 = x_2)} = \frac{p(x_1, x_2)}{p(x_2)},
\end{aligned}
$$

for x_2 such that $p(x_2) > 0$.
Note if $X_1 \perp\!\!\!\perp X_2$, $p(x_1|x_2) = p(x_1)$. In other words, the value of X_2 does not influence the value of X_1.

Probability

Conditional independence

- X_1 and X_2 are conditionally independent given X_3, written $X_1 \perp\!\!\!\perp X_2 | X_3$, iff
  - $p(x_1, x_2 | x_3) = p(x_1 | x_3) p(x_2 | x_3)$, or equivalently
  - $p(x_1 | x_2, x_3) = p(x_1 | x_3)$.

  for all x_3 such that $p(x_3) > 0$. Given X_3, there is no further relationship between X_1 and X_2.

Probability

Conditional independence

- X_1 and X_2 are conditionally independent given X_3, written $X_1 \perp\!\!\!\perp X_2 | X_3$, iff
  - $p(x_1, x_2 | x_3) = p(x_1 | x_3) p(x_2 | x_3)$, or equivalently
  - $p(x_1 | x_2, x_3) = p(x_1 | x_3)$.

  for all x_3 such that $p(x_3) > 0$. Given X_3, there is no further relationship between X_1 and X_2.
- Similarly, for sets of random variables, X_A and X_B are conditionally independent given X_C iff

$$p(x_A, x_B | x_C) = p(x_A | x_C) p(x_B | x_C)$$

  or

$$p(x_A | x_B, x_C) = p(x_A | x_C)$$

  for all x_C such that $p(x_C) > 0$.

Probability

Given the joint PMF, we can compute

- marginal PMF (marginalisation)
- conditional PMF (marginalisation and normalisation)

e.g. Given $p(x_1, x_2, x_3, x_4)$ for (X_1, X_2, X_3, X_4),

- $p(x_1, x_2) = \sum_{x_3, x_4} p(x_1, x_2, x_3, x_4)$
- $p(x_1 | x_2) = \frac{\sum_{x_3, x_4} p(x_1, x_2, x_3, x_4)}{\sum_{x_1, x_3, x_4} p(x_1, x_2, x_3, x_4)}$

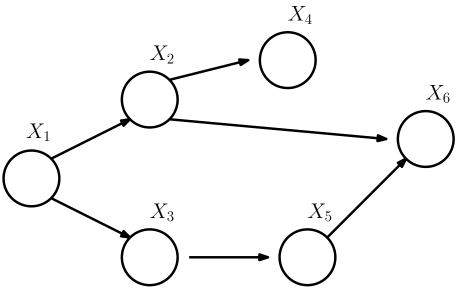
Goal: Construct models for probability distributions (joint PMF)

A toy example in modelling

X: a student's score in an exam $(0, 1, 2)$
Y: difficulty level of the exam $(0, 1)$
Z: the student's effort in general $(0, 1)$
L: quality of the reference letter from the professor who taught the course $(0, 1)$
W: the student's score in another course $(0, 1)$

Directed graphical models

- A directed graph $G(V, E)$, where V is a set of nodes and E is a set of oriented edges. For each $i \in V$, there is an associated random variable X_i.
- Further assume G is acyclic (DAG).
- For each $i \in V$, let π_i be the set of parent nodes. Then X_{π_i} is the "parents" of X_i.

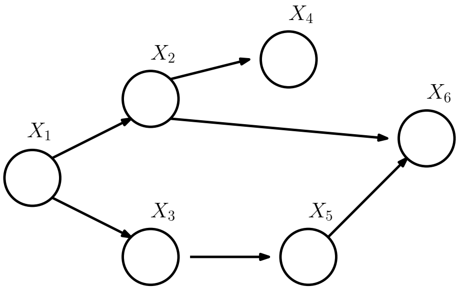


Directed graphical models

- To each node $i \in V$ associate a function $f_i(x_i, x_{\pi_i})$, where f_i is a nonnegative function satisfying $\sum_{x_i} f_i(x_i, x_{\pi_i}) = 1$. (compare $f_i(x_i, x_{\pi_i})$ vs $p(x_i|x_{\pi_i})$?)
- Define a joint PMF as

$$p(x_1, x_2, \ldots, x_n) := \prod_{i=1}^{n} f_i(x_i, x_{\pi_i}).$$

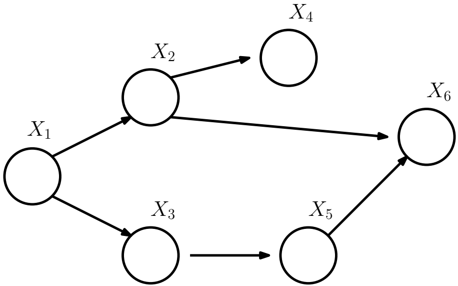
Is this a valid PMF?



Directed graphical models

- We can prove $f_i(x_i, x_{\pi_i})$ are in fact conditional probabilities $p(x_i|x_{\pi_i})$. It follows

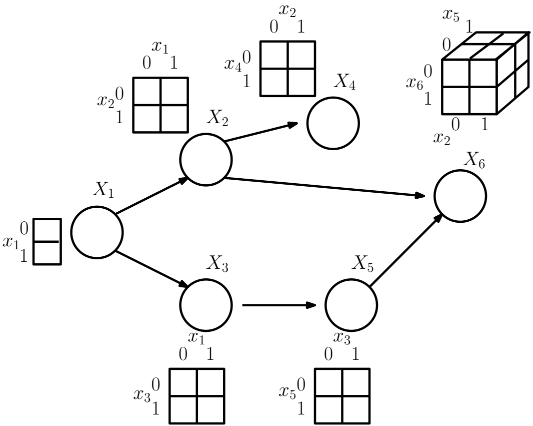
$$p(x_1, x_2, \ldots, x_n) := \prod_{i=1}^{n} p(x_i|x_{\pi_i}).$$



$$p(x_1, \ldots, x_6) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5)$$

Directed graphical models

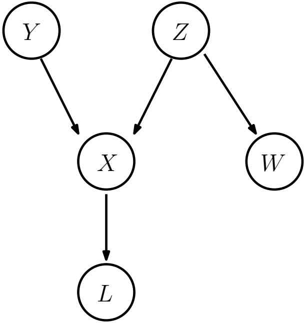
Representation economy - Suppose X_i's are all binary, $p(x_1, \ldots, x_6)$ needs a 6-dimensional table with 2^6 entries.



Directed graphical models

A toy example in modelling

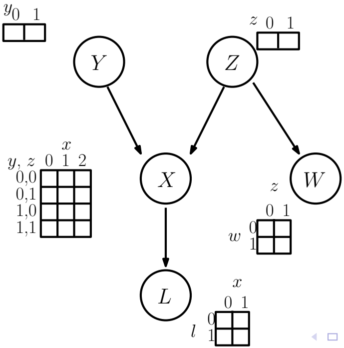
X: a student's score in an exam $(0, 1, 2)$
Y: difficulty level of the exam $(0, 1)$
Z: the student's effort in general $(0, 1)$
L: quality of the reference letter from the professor who taught the course $(0, 1)$
W: the student's score in another course $(0, 1)$



Directed graphical models

A toy example in modelling

X: a student's score in an exam $(0, 1, 2)$
Y: difficulty level of the exam $(0, 1)$
Z: the student's effort in general $(0, 1)$
L: quality of the reference letter from the professor who taught the course $(0, 1)$
W: the student's score in another course $(0, 1)$



DAG and conditional independence

Which independence assumptions are we exactly making by using a DAG model with a structure described by G? Important because

- we should know exactly what model assumptions we are making;
- this information will be helpful in designing inference algorithms later on.