# Probabilistic Modelling

Miguel Rios

Universiteit van Amsterdam

April 4, 2019

# Content

❶ Introduction

❷ PGM

❸ Introduction word alignment

# Probability review

- The sample space is the set of all possible outcomes of the experiment denoted by $\Omega$.
  For example, two successive coin tosses the sample space of $\{$hh, tt, ht, th$\}$, where $h$ heads and $t$ tails.

# Probability review

- The sample space is the set of all possible outcomes of the experiment denoted by $\Omega$.
  For example, two successive coin tosses the sample space of {hh, tt, ht, th}, where $h$ heads and $t$ tails.
- A event space is a set whose elements $A \in F$ (called events) are subsets of $\Omega$ (i.e., $A \subseteq \Omega$ is a collection of possible outcomes of an experiment)

# Probability review

- The sample space is the set of all possible outcomes of the experiment denoted by $\Omega$.
  For example, two successive coin tosses the sample space of {hh, tt, ht, th}, where $h$ heads and $t$ tails.
- A event space is a set whose elements $A \in F$ (called events) are subsets of $\Omega$ (i.e., $A \subseteq \Omega$ is a collection of possible outcomes of an experiment)
- Probability measure is a function $P : F \to \Re$, we associate a number $P(A)$ that measures the probability or degree of belief that the event will occur.

# Probability review

- The sample space is the set of all possible outcomes of the experiment denoted by $\Omega$.
  For example, two successive coin tosses the sample space of {hh, tt, ht, th}, where $h$ heads and $t$ tails.
- A event space is a set whose elements $A \in F$ (called events) are subsets of $\Omega$ (i.e., $A \subseteq \Omega$ is a collection of possible outcomes of an experiment)
- Probability measure is a function $P : F \rightarrow \Re$, we associate a number $P(A)$ that measures the probability or degree of belief that the event will occur.
- satisfies the following properties:

# Probability review

- The sample space is the set of all possible outcomes of the experiment denoted by $\Omega$.
  For example, two successive coin tosses the sample space of $\{$hh, tt, ht, th$\}$, where $h$ heads and $t$ tails.
- A event space is a set whose elements $A \in F$ (called events) are subsets of $\Omega$ (i.e., $A \subseteq \Omega$ is a collection of possible outcomes of an experiment)
- Probability measure is a function $P : F \rightarrow \Re$, we associate a number $P(A)$ that measures the probability or degree of belief that the event will occur.
- satisfies the following properties:
  - $P(A) \geq 0$

# Probability review

- The sample space is the set of all possible outcomes of the experiment denoted by $\Omega$.
  For example, two successive coin tosses the sample space of {hh, tt, ht, th}, where $h$ heads and $t$ tails.
- A event space is a set whose elements $A \in F$ (called events) are subsets of $\Omega$ (i.e., $A \subseteq \Omega$ is a collection of possible outcomes of an experiment)
- Probability measure is a function $P : F \rightarrow \Re$, we associate a number $P(A)$ that measures the probability or degree of belief that the event will occur.
- satisfies the following properties:
  - $P(A) \geq 0$
  - $A_1, A_2, \ldots$ are disjoint events (i.e. $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then
    $P(\bigcup_i A_i) = \sum_i P(A_i)$

# Probability review

- The sample space is the set of all possible outcomes of the experiment denoted by $\Omega$.

  For example, two successive coin tosses the sample space of $\{$hh, tt, ht, th$\}$, where $h$ heads and $t$ tails.
- A event space is a set whose elements $A \in F$ (called events) are subsets of $\Omega$ (i.e., $A \subseteq \Omega$ is a collection of possible outcomes of an experiment)
- Probability measure is a function $P : F \rightarrow \Re$, we associate a number $P(A)$ that measures the probability or degree of belief that the event will occur.
- satisfies the following properties:
  - $P(A) \geq 0$
  - $A_1$, $A_2$, . . . are disjoint events (i.e. $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then

    $P(\bigcup_i A_i) = \sum_i P(A_i)$
  - $P(\Omega) = 1$

## Example

Consider the event of tossing a six-sided die. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$.

We can define the simplest event space $F = \{\emptyset, \Omega\}$. Another event space is the set of all subsets of $\Omega$.

For the first event space, the probability measure is given by $P(\emptyset) = 0$, $P(\Omega) = 1$.

For the second event space, one valid probability measure is to assign the probability of each set in the event space to be $\frac{i}{6}$ where $i$ is the number of elements of that set; for example, $P(\{1, 2, 3, 4\}) = \frac{4}{6}$ and $P(\{1, 2, 3\}) = \frac{3}{6}$

# Conditional probability

- Let $B$ be an event with non-zero probability.
  The conditional probability of any event $A$ given $B$ is defined as:

$$P(A \mid B) = \frac{P(A, B)}{P(B)} \tag{1}$$

# Conditional probability

- Let $B$ be an event with non-zero probability.
  The conditional probability of any event $A$ given $B$ is defined as:

$$P(A \mid B) = \frac{P(A, B)}{P(B)} \tag{1}$$

- $P(A \mid B)$ is the probability measure of the event $A$ after observing the occurrence of event $B$.

# Chain rule

- Let $S_1, \cdots, S_k$ be events, $P(S_i) > 0$. Then the chain rule:

$$\begin{aligned} &P(S_1, S_2, \cdots, S_k) \\ =&P(S_1)P(S_2|S_1)P(S_3|S_2, S_1) \cdot P(S_k|S_1, S_2, \cdot S_{k-1}) \end{aligned} \tag{2}$$

# Chain rule

- Let $S_1, \cdots, S_k$ be events, $P(S_i) > 0$. Then the chain rule:

$$\begin{aligned} &P(S_1, S_2, \cdots, S_k) \\ =&P(S_1)P(S_2|S_1)P(S_3|S_2, S_1) \cdot P(S_k|S_1, S_2, \cdot S_{k-1}) \end{aligned} \tag{2}$$

- With $k = 2$ events, this is the definition of conditional probability:

$$P(S_1, S_2) = P(S_1)P(S_2|S_1) \tag{3}$$

# Chain rule

- Let $S_1, \cdots, S_k$ be events, $P(S_i) > 0$. Then the chain rule:

$$\begin{aligned} & P(S_1, S_2, \cdots, S_k) \\ = & P(S_1)P(S_2|S_1)P(S_3|S_2, S_1) \cdot P(S_k|S_1, S_2, \cdot S_{k-1}) \end{aligned} \tag{2}$$

- With $k = 2$ events, this is the definition of conditional probability:

$$P(S_1, S_2) = P(S_1)P(S_2|S_1) \tag{3}$$

- In general, the chain rule is derived by applying the definition of conditional probability multiple times, for example:

$$\begin{aligned} & P(S_1, S_2, S_3, S_4) \\ = & P(S_1, S_2, S_3)P(S_4 \mid S_1, S_2, S_3) \\ = & P(S_1, S_2)P(S_3 \mid S_1, S_2)P(S_4 \mid S_1, S_2, S_3) \\ = & P(S_1)P(S_2 \mid S_1)P(S_3 \mid S_1, S_2)P(S_4 \mid S_1, S_2, S_3) \end{aligned} \tag{4}$$

# Independence

- Two events are called independent if and only if $P(A, B) = P(A)P(B)$, or $P(A \mid B) = P(A)$

# Independence

- Two events are called independent if and only if $P(A, B) = P(A)P(B)$, or $P(A \mid B) = P(A)$
- Thus, independence is equivalent to saying that observing $B$ does not have any effect on the probability of $A$

# Random variables

- We flip 10 coins, and we want to know the number of coins that come up heads.
  The sample space $\Omega$ are 10-length sequences of heads and tails. For example, we might have $\omega_0 = \langle H, H, T, H, T, H, H, T, T, T \rangle \in \Omega$.

# Random variables

- We flip 10 coins, and we want to know the number of coins that come up heads.
  The sample space $\Omega$ are 10-length sequences of heads and tails. For example, we might have $\omega_0 = \langle H, H, T, H, T, H, H, T, T, T \rangle \in \Omega$.
- we care about real-valued functions of outcomes, the number of heads that appear among our 10 tosses.
  These functions are known as random variables.

# Random variables

- We flip 10 coins, and we want to know the number of coins that come up heads.
  The sample space $\Omega$ are 10-length sequences of heads and tails. For example, we might have $\omega_0 = \langle H, H, T, H, T, H, H, T, T, T \rangle \in \Omega$.
- we care about real-valued functions of outcomes, the number of heads that appear among our 10 tosses.
  These functions are known as random variables.
- A random variable $X$ is a function $X : \Omega \to \Re$.

# Random variables

- We flip 10 coins, and we want to know the number of coins that come up heads.
  The sample space $\Omega$ are 10-length sequences of heads and tails. For example, we might have $\omega_0 = \langle H, H, T, H, T, H, H, T, T, T \rangle \in \Omega$.
- we care about real-valued functions of outcomes, the number of heads that appear among our 10 tosses.
  These functions are known as random variables.
- A random variable $X$ is a function $X : \Omega \to \Re$.
- We will denote random variables using upper case letters $X$

# Random variables

- We flip 10 coins, and we want to know the number of coins that come up heads.
  The sample space $\Omega$ are 10-length sequences of heads and tails. For example, we might have $\omega_0 = \langle H, H, T, H, T, H, H, T, T, T \rangle \in \Omega$.
- we care about real-valued functions of outcomes, the number of heads that appear among our 10 tosses.
  These functions are known as random variables.
- A random variable $X$ is a function $X : \Omega \rightarrow \Re$.
- We will denote random variables using upper case letters $X$
- We will denote the value that a random variable may take on using lower case letters $x$.
  Thus, $X = x$ means that we are assigning the value $x \in \Re$ to the random variable $X$

# Cumulative distribution functions

- To specify the probability measures used with random variables, it is convenient to specify alternative functions (CDFs, PDFs, and PMFs).

# Cumulative distribution functions

- To specify the probability measures used with random variables, it is convenient to specify alternative functions (CDFs, PDFs, and PMFs).
- A cumulative distribution function (CDF) is a function $F_X : \Re \to [0, 1]$ which specifies a probability measure as,

$$F_X(x) = P(X \leq x) \tag{5}$$

# Cumulative distribution functions

- To specify the probability measures used with random variables, it is convenient to specify alternative functions (CDFs, PDFs, and PMFs).
- A cumulative distribution function (CDF) is a function $F_X : \Re \rightarrow [0, 1]$ which specifies a probability measure as,

$$F_X(x) = P(X \leq x) \tag{5}$$

- Properties:

$$\begin{aligned} &0 \leq F_X(x) \leq 1 \\ &\lim_{x \rightarrow -\infty} F_X(x) = 0 \\ &\lim_{x \rightarrow +\infty} F_X(x) = 1 \\ &x \leq y \rightarrow F_X(x) \leq F_X(y) \end{aligned} \tag{6}$$

# Probability mass functions

- When a random variable $X$ takes on a finite set of possible values is a discrete random variable

# Probability mass functions

- When a random variable $X$ takes on a finite set of possible values is a discrete random variable
- A way to represent the probability measure associated with a random variable is to directly specify the probability of each value that the random variable can assume a probability mass function PMF is a function

# Probability mass functions

- When a random variable $X$ takes on a finite set of possible values is a discrete random variable
- A way to represent the probability measure associated with a random variable is to directly specify the probability of each value that the random variable can assume a probability mass function PMF is a function
- $p_X : \Omega \rightarrow \Re$ such that $p_X(x) = P(X = x)$

# Probability mass functions

- When a random variable $X$ takes on a finite set of possible values is a discrete random variable
- A way to represent the probability measure associated with a random variable is to directly specify the probability of each value that the random variable can assume a probability mass function PMF is a function
- $p_X : \Omega \rightarrow \Re$ such that $p_X(x) = P(X = x)$
- Properties:

$$
\begin{aligned}
0 \leq p_X(x) \leq 1 \\
\sum_{x \in X} p_X(x) = 1 \\
\sum_{x \in A} p_X(x) = P(X \in A)
\end{aligned}
\tag{7}
$$

# Probability density functions

- For some continuous random variables, the cumulative distribution function $F_X(x)$ is differentiable everywhere.
  In these cases, we define the Probability Density Function or PDF as the derivative of the CDF

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{8}$$

# Probability density functions

- For some continuous random variables, the cumulative distribution function $F_X(x)$ is differentiable everywhere.
  In these cases, we define the Probability Density Function or PDF as the derivative of the CDF

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{8}$$

- Properties:

$$\begin{gathered} f_X(x) \geq 0 \\ \int_{-\infty}^{\infty} f_X(x) = 1 \\ \int_{x \in A} f_X(x) dx = P(X \in A) \end{gathered} \tag{9}$$

# Expectation

- $X$ is a discrete random variable with PMF $p_X(x)$ and $g : \Re \rightarrow \Re$ is an arbitrary function.

# Expectation

- $X$ is a discrete random variable with PMF $p_X(x)$ and $g : \Re \to \Re$ is an arbitrary function.
- In this case, $g(X)$ can be considered a random variable, and we define the expectation of $g(X)$ as

$$\mathbb{E}[g(X)] = \sum_{x \in X} g(x) p_X(x) \tag{10}$$

# Expectation

- $X$ is a discrete random variable with PMF $p_X(x)$ and $g : \Re \rightarrow \Re$ is an arbitrary function.
- In this case, $g(X)$ can be considered a random variable, and we define the expectation of $g(X)$ as

$$\mathbb{E}[g(X)] = \sum_{x \in X} g(x) p_X(x) \tag{10}$$

- If $X$ is a continuous random variable with PDF $f_X(x)$, then the expected value of $g(X)$ is defined as:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \tag{11}$$

# Expectation

- Intuitively, the expectation of $g(X)$ can be thought of as a weighted average of the values that $g(x)$ can taken on for different values of $x$, where the weights are given by $p_X(x)$

# Expectation

- Intuitively, the expectation of $g(X)$ can be thought of as a weighted average of the values that $g(x)$ can taken on for different values of $x$, where the weights are given by $p_X(x)$
- Properties:

$$
\begin{aligned}
&\mathbb{E}[a] = a \text{for any constant} a \in \Re \\
&\mathbb{E}[af(X)] = a\,\mathbb{E}[f(X)] \text{for any constant} a \in \Re \\
&\text{Linearity of Expectation}\, \mathbb{E}[f(X) + g(X)] = \mathbb{E}[f(X)] + \mathbb{E}[g(X)]
\end{aligned}
\tag{12}
$$

# Discrete random variables

- $X \sim \text{Bernoulli}(p)$ (where $0 \leq p \leq 1$):
  one if a coin with heads probability $p$ comes up heads, zero otherwise

$$p(x) = \begin{cases} p, & \text{if } x = 1. \\ 1 - p, & \text{if } x = 0. \end{cases} \tag{13}$$

# Discrete random variables

- $X \sim \text{Bernoulli}(p)$ (where $0 \leq p \leq 1$):
  one if a coin with heads probability $p$ comes up heads, zero otherwise

$$p(x) = \begin{cases} p, & \text{if } x = 1. \\ 1 - p, & \text{if } x = 0. \end{cases} \tag{13}$$

- $X \sim \text{Binomial}(n, p)$ (where $0 \leq p \leq 1$):
  the number of heads in $n$ independent flips of a coin with heads probability $p$

$$p = \binom{n}{x} \cdot p^x (1 - p)^{n-x} \tag{14}$$

# Discrete random variables

- $X \sim$ Geometric$(p)$ (where $p > 0$):
  the number of flips of a coin with heads probability $p$ until the first heads.
  $$p(x) = p(1-p)^{x-1} \tag{15}$$

# Discrete random variables

- $X \sim$ Geometric$(p)$ (where $p > 0$):
  the number of flips of a coin with heads probability $p$ until the first heads.
  $$p(x) = p(1-p)^{x-1} \tag{15}$$
- $X \sim$ Poisson$(\lambda)$ (where $\lambda > 0$):
  a probability distribution over the non-negative integers used for modelling the frequency of rare events.
  $$p(x) = e^{-\lambda}\frac{\lambda^x}{x!} \tag{16}$$

# Continuous random variables

- $X \sim$ Uniform$(a, b)$ (where $a < b$): equal probability density to every value between $a$ and $b$ on the real line

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq b \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

# Continuous random variables

- $X \sim$ Uniform$(a, b)$ (where $a < b$):
  equal probability density to every value between $a$ and $b$ on the real line

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq b \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

- $X \sim$ Exponential$(\lambda)$ (where $\lambda > 0$):
  decaying probability density over the non-negative real

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

# Continuous random variables

- $X \sim$ Uniform$(a, b)$ (where $a < b$):
  equal probability density to every value between $a$ and $b$ on the real line

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq b \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

- $X \sim$ Exponential$(\lambda)$ (where $\lambda > 0$):
  decaying probability density over the non-negative real

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

- $X \sim$ Normal$(\mu, \sigma^2)$: also known as the Gaussian distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{19}$$

# Random variable example

- We cannot talk about the exact value of the random variable but we can reason about it's possible values

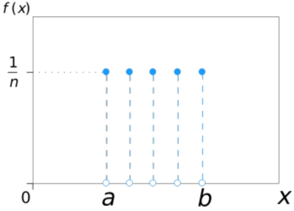
# Random variable example

- We cannot talk about the exact value of the random variable but we can reason about it's possible values
- We quantify the degree of belief we have in each outcome

# Random variable example

- We cannot talk about the exact value of the random variable but we can reason about it's possible values
- We quantify the degree of belief we have in each outcome
- Uniform distribution: every outcome is equally likely
  if $n$ is the size of the set of possible outcomes the probability that $x$ takes on any value (e.g. a) is $\frac{1}{n}$

$$p(x) = \frac{1}{n}\text{for all}x \in [a, b] \quad (20)$$



# Random variable example

- A random variable is a function that maps from a sample space $\Omega$ to $\Re$
  $x : \Omega \rightarrow \Re$

# Random variable example

- A random variable is a function that maps from a sample space $\Omega$ to $\Re$
  $x : \Omega \rightarrow \Re$
- Example: which pet do kids love the most?
  Sample space: $\Omega = \{\text{bird, cat, dog}\}$

$$x(\omega) = \begin{cases} 1 & \omega = bird \\ 2 & \omega = cat \\ 3 & \omega = dog \end{cases} \tag{21}$$

# Random variable example

- A random variable is a function that maps from a sample space $\Omega$ to $\Re$
  $x : \Omega \rightarrow \Re$
- Example: which pet do kids love the most?
  Sample space: $\Omega = \{$bird, cat, dog$\}$

$$x(\omega) = \begin{cases} 1 & \omega = bird \\ 2 & \omega = cat \\ 3 & \omega = dog \end{cases} \tag{21}$$

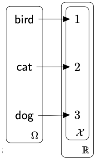
- if say $x$ we mean the set of outcomes
  $\omega : x(\omega) = x$ which is called an event

# Random variable example

- A random variable is a function that maps from a sample space $\Omega$ to $\Re$
  $x : \Omega \rightarrow \Re$
- Example: which pet do kids love the most?
  Sample space: $\Omega = \{$bird, cat, dog$\}$

$$x(\omega) = \begin{cases} 1 & \omega = bird \\ 2 & \omega = cat \\ 3 & \omega = dog \end{cases} \tag{21}$$

- if say $x$ we mean the set of outcomes
  $\omega : x(\omega) = x$ which is called an event
- we call $\mathcal{X}$ the support of $X$



# Random variable example

- A Categorical variable can model $1$ of $k$ categories
  $x \sim \mathsf{Cat}(\theta_1, ..., \theta_k)$

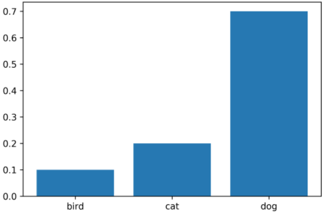
# Random variable example

- A Categorical variable can model $1$ of $k$ categories
  $x \sim \mathsf{Cat}(\theta_1, ..., \theta_k)$
- $x = 1, ..., k$

# Random variable example

- A Categorical variable can model $1$ of $k$ categories $x \sim \text{Cat}(\theta_1, ..., \theta_k)$
- $x = 1, ..., k$
- the categorical parameter is a probability vector

$$\begin{aligned} 0 \leq \theta_x \leq 1 \text{for} x \in [1, k] \\ \sum_{x=1}^{k} \theta_x = 1 \end{aligned} \tag{22}$$



# Sum rule and product rule

- $p(x, y)$ is the joint distribution of two random variables $x, y$.

# Sum rule and product rule

- $p(x, y)$ is the joint distribution of two random variables $x, y$.
- product rule: $p(x, y) = p(y \mid x)p(x)$

# Sum rule and product rule

- $p(x, y)$ is the joint distribution of two random variables $x, y$.
- product rule: $p(x, y) = p(y \mid x)p(x)$
- How does the joint PMF over two variables relate to the PMF for each variable separately?
  With the corresponding marginal distributions $p(x)$ and $p(y)$

# Sum rule and product rule

- $p(x, y)$ is the joint distribution of two random variables $x, y$.
- product rule: $p(x, y) = p(y \mid x)p(x)$
- How does the joint PMF over two variables relate to the PMF for each variable separately?
  With the corresponding marginal distributions $p(x)$ and $p(y)$
- We denote the sum rule as (also known as the marginalization property):

$$p(x) = \begin{cases} \sum_{y \in Y} p(x, y), & \text{if} y \text{is discrete} \\ \int_{Y} p(x, y) dy, & \text{if} y \text{is continuous} \end{cases} \tag{23}$$

# Sum rule and product rule

- $p(x, y)$ is the joint distribution of two random variables $x, y$.
- product rule: $p(x, y) = p(y \mid x)p(x)$
- How does the joint PMF over two variables relate to the PMF for each variable separately?
  With the corresponding marginal distributions $p(x)$ and $p(y)$
- We denote the sum rule as (also known as the marginalization property):

$$p(x) = \begin{cases} \sum_{y \in Y} p(x, y), & \text{if} y \text{is discrete} \\ \int_Y p(x, y) dy, & \text{if} y \text{is continuous} \end{cases} \tag{23}$$

- We sum out (or integrate out) the set of states $y$ of the random variable $Y$.

# Bayes' rule

- To derive expressions for conditional probability Bayes' rule

$$\underbrace{p(y \mid x)}_{\text{posterior}} = \frac{\overbrace{p(x \mid y)}^{\text{likelihood}} \overbrace{p(y)}^{prior}}{\underbrace{p(x)}_{evidence}} \tag{24}$$

# Bayes' rule

- To derive expressions for conditional probability Bayes' rule

# Bayes' rule

- To derive expressions for conditional probability Bayes' rule
- In the case of discrete random variables $X$ and $Y$

$$p(y \mid x) = \frac{p(x,y)}{p(x)} = \frac{p(x \mid y)p(y)}{\sum_{y' \in Y} p(x \mid y')p(y')} \tag{25}$$

# Bayes' rule

- To derive expressions for conditional probability Bayes' rule
- In the case of discrete random variables $X$ and $Y$

$$p(y \mid x) = \frac{p(x,y)}{p(x)} = \frac{p(x \mid y)p(y)}{\sum_{y' \in Y} p(x \mid y')p(y')} \tag{25}$$

- If the random variables $X$ and $Y$ are continuous

$$f(y \mid x) = \frac{f(x,y)}{f_X(x)} = \frac{f(x \mid y)f(y)}{\int_{-\infty}^{\infty} f(x \mid y')f(y')dy'} \tag{26}$$

# Probabilistic modelling

- Representation
  How to express a probability distribution that models some real-world phenomenon?

# Probabilistic modelling

- Representation
  How to express a probability distribution that models some real-world phenomenon?
- Inference
  Given a probabilistic model, how do we obtain answers to relevant questions about the world?
  Querying the marginal or conditional probabilities of certain events of interest.

# Probabilistic modelling

- Representation
  How to express a probability distribution that models some real-world phenomenon?
- Inference
  Given a probabilistic model, how do we obtain answers to relevant questions about the world?
  Querying the marginal or conditional probabilities of certain events of interest.
- Learning
  Goal of fitting a model given a dataset. The model can be then use to make predictions about the future.

# Bayesian networks

- Directed graphical models are a family of probability distributions that admit a compact parameterisation that can be described using a directed graph.

# Bayesian networks

- Directed graphical models are a family of probability distributions that admit a compact parameterisation that can be described using a directed graph.
- By the chain rule we can write any probability as:

$$p(x_1, x_2, ..., x_n) = p(x_1)p(x_2 \mid x_1) \cdots p(x_n \mid x_{n-1}, ..., x_2, x_1). \quad (27)$$

# Bayesian networks

- Directed graphical models are a family of probability distributions that admit a compact parameterisation that can be described using a directed graph.
- By the chain rule we can write any probability as:

$$p(x_1, x_2, ..., x_n) = p(x_1)p(x_2 \mid x_1) \cdots p(x_n \mid x_{n-1}, ..., x_2, x_1). \quad (27)$$

- A Bayesian network is a distribution in which each factor on the right hand side depends only on a small number of ancestor variables $x_{A_i}$:

$$p(x_i \mid x_{i-1}, ..., x_1) = p(x_i \mid x_{A_i}) \quad (28)$$

# Bayesian networks

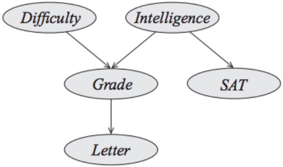
- Distributions of this form can be naturally expressed as directed acyclic graphs (DAG), in which vertices correspond to variables $x_i$ and edges indicate dependency relationships.

# Bayesian networks

- Distributions of this form can be naturally expressed as directed acyclic graphs (DAG), in which vertices correspond to variables $x_i$ and edges indicate dependency relationships.

Model of a student's grade $g$ on an exam. This grade depends on the exam's difficulty $d$ and the student's intelligence $i$ it also affects the quality $l$ of the reference letter from the professor who taught the course. The student's intelligence $i$ affects his SAT score $s$ in addition to $g$. Each variable is binary, except for $g$, which takes 3 possible values.

# Bayesian networks



$$p(l, g, i, d, s) = p(l \mid g)p(g \mid i, d)p(i)p(d)p(s \mid i) \tag{29}$$

# Bayesian networks

- Bayesian network is a directed graph $G = (V, E)$

# Bayesian networks

- Bayesian network is a directed graph $G = (V, E)$
- Together with a random variable $x_i$ for each node $i \in V$

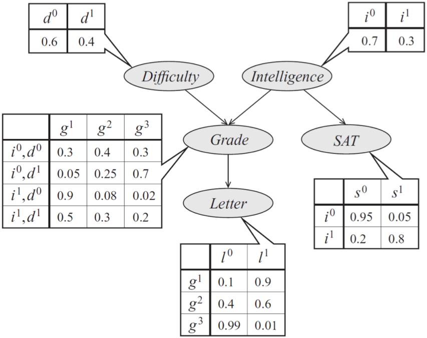
# Bayesian networks

- Bayesian network is a directed graph $G = (V, E)$
- Together with a random variable $x_i$ for each node $i \in V$
- One conditional probability distribution (CPD) conditioned on its parents
  $p(x_i \mid x_{A_i})$

# Bayesian networks

- Bayesian network is a directed graph $G = (V, E)$
- Together with a random variable $x_i$ for each node $i \in V$
- One conditional probability distribution (CPD) conditioned on its parents
  $p(x_i \mid x_{A_i})$
- probability $p$ factorizes over a DAG $G$ if it can be decomposed into a product of factors

# Bayesian networks



| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

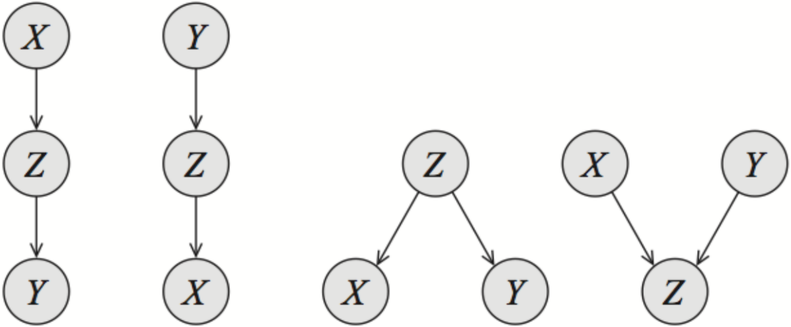
|  | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0, d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0, d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1, d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1, d^1$ | 0.5 | 0.3 | 0.2 |

|  | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

|  | $l^0$ | $l^1$ |
|---|---|---|
| $g^1$ | 0.1 | 0.9 |
| $g^2$ | 0.4 | 0.6 |
| $g^3$ | 0.99 | 0.01 |

$$p(l, g, i, d, s) = p(l \mid g)p(g \mid i, d)p(i)p(d)p(s \mid i) \tag{30}$$

# Bayesian networks



# Probabilistic modelling

- Inference
  Given a probabilistic model, how do we obtain answers to relevant questions about the world?
  Querying the marginal or conditional probabilities of certain events of interest.

$$p(x_1) = \sum_{x_2} \sum_{x_3} ... \sum_{x_n} p(x_1, x_2, x3, ..., x_n) \tag{31}$$

# Word alignment

- IBM models assume that each word in the French sentence is a translation of exactly zero or one word of the English sentence.

# Word alignment

- IBM models assume that each word in the French sentence is a translation of exactly zero or one word of the English sentence.
- The notation to refer to each word.
  Let a French sentence $f$ be represented by an array of $m$ words,
  $\langle f_1, ..., f_m \rangle$,
  and English sentence $e$ be represented by an array of $l$ words,
  $\langle e_1, ..., e_l \rangle$

# Word alignment

- IBM models assume that each word in the French sentence is a translation of exactly zero or one word of the English sentence.
- The notation to refer to each word.
  Let a French sentence $f$ be represented by an array of $m$ words, $\langle f_1, ..., f_m \rangle$,
  and English sentence $e$ be represented by an array of $l$ words, $\langle e_1, ..., e_l \rangle$
- IBM models decompose the joint probability of a sentence pair with the chain rule as:

$$p(e_1^l, f_1^m) = \underbrace{p(e_1^l)}_{\text{language model}} \times \underbrace{p(f_1^m \mid e_1^l)}_{\text{translation model}} \tag{32}$$

# Word alignment

- IBM models assume that each word in the French sentence is a translation of exactly zero or one word of the English sentence.
- The notation to refer to each word.
  Let a French sentence $f$ be represented by an array of $m$ words,
  $\langle f_1, ..., f_m \rangle$,
  and English sentence $e$ be represented by an array of $l$ words,
  $\langle e_1, ..., e_l \rangle$
- IBM models decompose the joint probability of a sentence pair with the chain rule as:

$$p(e_1^l, f_1^m) = \underbrace{p(e_1^l)}_{\text{language model}} \times \underbrace{p(f_1^m \mid e_1^l)}_{\text{translation model}} \tag{32}$$

- French words are conditionally independent given the English sentence.

# Word alignment

- IBM models assume that each word in the French sentence is a translation of exactly zero or one word of the English sentence.
- The notation to refer to each word.
  Let a French sentence $f$ be represented by an array of $m$ words, $\langle f_1, ..., f_m \rangle$,
  and English sentence $e$ be represented by an array of $l$ words, $\langle e_1, ..., e_l \rangle$
- IBM models decompose the joint probability of a sentence pair with the chain rule as:

$$p(e_1^l, f_1^m) = \underbrace{p(e_1^l)}_{\text{language model}} \times \underbrace{p(f_1^m \mid e_1^l)}_{\text{translation model}} \tag{32}$$

- French words are conditionally independent given the English sentence.
- Inference can be performed exactly.

# Mixture models

- A mixture model consist of $c$ mixture components, each defines a distribution over the space $X$.

# Mixture models

- A mixture model consist of $c$ mixture components, each defines a distribution over the space $X$.
- Each component can specialise its distribution on a subset of the data.

# Mixture models

- A mixture model consist of $c$ mixture components, each defines a distribution over the space $X$.
- Each component can specialise its distribution on a subset of the data.
- The probability of a mixture model with $c$ components assigns to $n$ data point is denoted by:

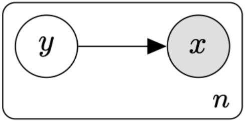
$$
\begin{aligned}
p(x_1^n) &= \prod_{i=1}^{n} \sum_{j=1}^{c} p(x_i, y_i = j) \\
&= \prod_{i=1}^{n} \sum_{j=1}^{c} p(y_i = j) p(x_i \mid y_i = j)
\end{aligned}
\tag{33}
$$

# Mixture models

- A mixture model consist of $c$ mixture components, each defines a distribution over the space $X$.
- Each component can specialise its distribution on a subset of the data.
- The probability of a mixture model with $c$ components assigns to $n$ data point is denoted by:

$$\begin{aligned} p(x_1^n) &= \prod_{i=1}^{n} \sum_{j=1}^{c} p(x_i, y_i = j) \\ &= \prod_{i=1}^{n} \sum_{j=1}^{c} p(y_i = j) p(x_i \mid y_i = j) \end{aligned} \tag{33}$$

- We introduced the random variable $y_i$ that ranges over the mixture components



# Word Alignment

- Learn a conditional probabilistic model of a French sentence $f$ given an English sentence $e$,
  which we denote as $p(f|e)$.

# Word Alignment

- Learn a conditional probabilistic model of a French sentence $f$ given an English sentence $e$,
  which we denote as $p(f|e)$.
- A dataset $D$ of $N$ sentence pairs that are known to be translations of each other,
  $D = (f^{(1)}, e^{(1)})...(f^{(N)}, e^{(N)})$

# Word Alignment

- Learn a conditional probabilistic model of a French sentence $f$ given an English sentence $e$,
  which we denote as $p(f|e)$.
- A dataset $D$ of $N$ sentence pairs that are known to be translations of each other,
  $D = (f^{(1)}, e^{(1)})...(f^{(N)}, e^{(N)})$
- Goal of our model will be to uncover the hidden word-to-word correspondences in these translation pairs.

# Word Alignment

- Learn a conditional probabilistic model of a French sentence $f$ given an English sentence $e$,
  which we denote as $p(f|e)$.
- A dataset $D$ of $N$ sentence pairs that are known to be translations of each other,
  $D = (f^{(1)}, e^{(1)})...(f^{(N)}, e^{(N)})$
- Goal of our model will be to uncover the hidden word-to-word correspondences in these translation pairs.
- We will learn the model from data, and use it to predict the existence of the missing word alignments

# Generative Process

- Generative process for the French sentence conditioned on the English sentence:

# Generative Process

- Generative process for the French sentence conditioned on the English sentence:
  1. Choose French sentence length $m$ based on the English sentence length $l$

# Generative Process

- Generative process for the French sentence conditioned on the English sentence:
  1. Choose French sentence length $m$ based on the English sentence length $l$
  2. For each French position $j$, choose the English position $a_j$ that it is generated from

# Generative Process

- Generative process for the French sentence conditioned on the English sentence:
    1. Choose French sentence length $m$ based on the English sentence length $l$
    2. For each French position $j$, choose the English position $a_j$ that it is generated from
    3. For each French position $j$, choose a French word based on the English word in position $a_j$

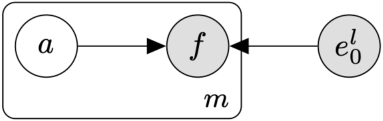
# Generative Process

- Generative process for the French sentence conditioned on the English sentence:
  1. Choose French sentence length $m$ based on the English sentence length $l$
  2. For each French position $j$, choose the English position $a_j$ that it is generated from
  3. For each French position $j$, choose a French word based on the English word in position $a_j$
- The generative story introduces the alignment variable $a_j$
  It is an indicator for the mixture component that the French word in position $j$ is generated from

# Generative Process

- Generative process for the French sentence conditioned on the English sentence:
  1. Choose French sentence length $m$ based on the English sentence length $l$
  2. For each French position $j$, choose the English position $a_j$ that it is generated from
  3. For each French position $j$, choose a French word based on the English word in position $a_j$
- The generative story introduces the alignment variable $a_j$
  It is an indicator for the mixture component that the French word in position $j$ is generated from
- The mixture components are English words

# IBM graphical model



# Questions?

# References I