Probabilistic Modelling

Miguel Rios

Universiteit van Amsterdam

March 31, 2019

Content

Probability review

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

Probability review

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

For the first event space, the unique probability measure satisfying the requirements above 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 Ω 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 Ω 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 Ω 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 $X : \Omega \rightarrow \Re$.

Random variables

- We flip 10 coins, and we want to know the number of coins that come up heads.
  The sample space Ω 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

Random variables

- We flip 10 coins, and we want to know the number of coins that come up heads.
  The sample space Ω 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
- 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 $F_X : \Re \rightarrow [0, 1]$ 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 $F_X : \Re \to [0, 1]$ which specifies a probability measure as,

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

- Properties:

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

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:

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

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 $g : \Re \to \Re$ 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

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

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

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

- 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

- $X \sim$ 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 $0 \le p \le 1$):
  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 $0 \le p \le 1$):
  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

- $X \sim \text{Geometric}(p)$ (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

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

- $X \sim \mathsf{Uniform}(a, b)$ (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

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

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

- $X \sim \text{Exponential}(\lambda)$ (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

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

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

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

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

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

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

Sum rule and product rule

- $p(x, y)$ is the joint distribution of two random variables x, y. 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.
  With the corresponding marginal distributions $p(x)$ and $p(y)$
- $p(y|x)$ is the conditional distribution of y given x.

Sum rule and product rule

- $p(x, y)$ is the joint distribution of two random variables x, y. With the corresponding marginal distributions $p(x)$ and $p(y)$
- $p(y|x)$ is the conditional distribution of y given x.
- We denote the sum rule as (also known as the marginalization property):

$$p(x) = \begin{cases} \sum_{y \in Y} p(x, y), & \text{if} y \text{is discrete} \\ \int_Y p(x, y) dy, & \text{if} y \text{is continuous} \end{cases} \tag{20}$$

Sum rule and product rule

- $p(x, y)$ is the joint distribution of two random variables x, y. With the corresponding marginal distributions $p(x)$ and $p(y)$
- $p(y|x)$ is the conditional distribution of y given x.
- We denote the sum rule as (also known as the marginalization property):

$$p(x) = \begin{cases} \sum_{y \in Y} p(x, y), & \text{if} y \text{is discrete} \\ \int_Y p(x, y) dy, & \text{if} y \text{is continuous} \end{cases} \tag{20}$$

- 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

$$\underbrace{p(y \mid x)}_{\text{posterior}} = \frac{\overbrace{p(x \mid y)}^{\text{likelihood}} \overbrace{p(y)}^{prior}}{\underbrace{p(x)}_{evidence}} \tag{21}$$

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

$$p(y \mid x) = \frac{p(x, y)}{p(x)} = \frac{p(x \mid y)p(y)}{\sum_{y' \in Y} p(x \mid y')p(y')} \tag{22}$$

Bayes' rule

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

$$p(y \mid x) = \frac{p(x,y)}{p(x)} = \frac{p(x \mid y)p(y)}{\sum_{y' \in Y} p(x \mid y')p(y')} \tag{22}$$

- If the random variables X and Y are continuous

$$f(y \mid x) = \frac{f(x,y)}{f_X(x)} = \frac{f(x \mid y)f(y)}{\int_{-\infty}^{\infty} f(x \mid y')f(y')dy'} \tag{23}$$

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

-

Questions?

References I