Revision of Probability

Advanced Statistical Inference

Simone Rossi

Syntax of Probability

Random Variables

> A random variable [...] refers to a part of the world whose status is initially unknown. [...]

S.Russell, P.Norvig, "Artificial Intelligence. A Modern Approach", Prentice Hall (2003)

Probability is a mathematical framework to reason about uncertain events.

Some definitions

Probability space $(\Omega, \mathcal{F}, \mathbb{P})$:

- Ω is the *sample space*, the set of all possible outcomes of an experiment;
- $\mathcal{F}$ is the *event space*, a set of all possible subsets of Ω;
- $\mathbb{P}$ is the *probability measure*, a function that assigns probabilities to events (i.e., $\mathbb{P} : \mathcal{F} \to [0, 1]$).

Random variable X:

- A function that maps outcomes in Ω to a set of values in $\mathbb{R}$: $X : \Omega \to \mathbb{R}$.
- Assigns a numerical value to each outcome in Ω.

Probability axioms

The probability laws need to satisfy three axioms, also known as **Kolmogorov's axioms**:

1. **Non-negativity**: $\mathbb{P}(E) \geq 0$ for all $E \in \mathcal{F}$;
2. **Normalization**: $\mathbb{P}(\Omega) = 1$;
3. **Additivity**: For any sequence of mutually exclusive events $E_1, E_2, \ldots$ (i.e., $E_i \cap E_j = \emptyset$ for $i \neq j$), we have

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i).$$

Some properties

- **Complement**:

We define the complement of an event E as $E^c = \Omega \setminus E$. Then, $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$;

- **Joint**:

For any two events E and F, we define the joint probability of E and F both occurring as $\mathbb{P}(E \cap F) = \mathbb{P}(E, F)$; If E and F are independent, then $\mathbb{P}(E \cap F) = \mathbb{P}(E) \cdot \mathbb{P}(F)$.

- **Union**:

For any two events E and F, we define the probability of E or F occurring as $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$. If E and F are mutually exclusive, then $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$.

Discrete Random Variables

A random variable X is *discrete* if it takes on a finite number of values.

We define the probability of a random variable X taking on a value x as $\mathbb{P}(X = x)$.

We also define the **probability mass function** (PMF) as $p_X(x) = \mathbb{P}(X = x)$, or $p(x)$ for short.

> **Example**
>
> An example of a discrete random variable is the outcome of a die roll.

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$
$$\mathcal{F} = \{\emptyset, \{1\}, \{2\}, \dots, \{1, 2\}, \dots, \{1, 2, 3, 4, 5, 6\}\}$$
$$\mathbb{P}(X = i) = \frac{1}{\alpha_i}, \quad \text{with} \quad \sum_{i=1}^{6} \frac{1}{\alpha_i} = 1.$$

Joint Probability of Discrete Random Variables

For two discrete random variables X and Y, we define the **joint probability mass function** (PMF) as $p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$.

Example

- X is tomorrow's weather: $X \in \{\text{rainy}, \text{sunny}, \text{cloudy}, \text{snowy}\}$;
- Y is a binary variable indicating whether I will arrive at work on time: $Y \in \{\text{yes}, \text{no}\}$.
- For N days, update a table with the corresponding number of occurrences.

On-time/Weather	sunny	rainy	cloudy	snowy
yes	40	15	5	0
no	5	35	10	1

The probability of $X = \text{sunny}$ and $Y = \text{yes}$ is

$$\mathbb{P}(X = \text{sunny}, Y = \text{yes}) = p_{X,Y}(\text{sunny}, \text{yes}) = 40/N.$$

Sum Rule

Let's consider two random variables X with M possible values and Y with L possible values.

- **Sum rule**: The probability of X is the sum of the joint probabilities of X and Y over all possible values of Y:

$$\mathbb{P}(X = x) = \sum_{i=1}^{L} \mathbb{P}(X = x, Y = y_i) = \sum_{y} p(x, y).$$

This is also known as the **marginalization** rule.

Example

In the previous example, the probability of $X = \text{sunny}$ is

$$\mathbb{P}(X = \text{sunny}) = \sum_{y \in \{\text{yes,no}\}} \mathbb{P}(X = \text{sunny}, Y = y) = \frac{40 + 5}{N}.$$

Product Rule

Consider $X = x_i$, then the fraction for which $Y = y_j$ is called the **conditional probability** $\mathbb{P}(Y = y_j \mid X = x_i)$

- **Product rule**: The joint probability of X and Y is the product of the conditional probability of Y given X and the probability of X:

$$\begin{aligned}\mathbb{P}(X = x, Y = y) &= \mathbb{P}(Y = y \mid X = x)\mathbb{P}(X = x) = p(y \mid x)p(x) \\ &= \mathbb{P}(X = x \mid Y = y)\mathbb{P}(Y = y) = p(x \mid y)p(y)\end{aligned}$$

Generalization to N Random Variables

For N random variables $X_1, X_2, \ldots, X_N$, we can generalize the product rule as

$$\begin{aligned}\mathbb{P}(X_1, \ldots, X_N) &= \mathbb{P}(X_N \mid X_1, \ldots, X_{N-1})\mathbb{P}(X_1, \ldots, X_{N-1}) \\ &= \mathbb{P}(X_N \mid X_1, \ldots, X_{N-1})\mathbb{P}(X_{N-1} \mid X_1, \ldots, X_{N-2})\mathbb{P}(X_1, \ldots, X_{N-2}) \\ &= \prod_{i=1}^{N} \mathbb{P}(X_i \mid X_1, \ldots, X_{i-1}).\end{aligned}$$

Continuous Random Variables

Continuous Random Variables

A random variable X is *continuous* if it takes on an infinite number of values.

To define the probability of a continuous random variable X taking on a value x, we use the **probability density function** (PDF) $p_X(x)$.

- Probability of X falling in the interval $[a, b]$ is

$$\mathbb{P}(X \in [a,b]) = \int_a^b p_X(x)dx.$$

- The PDF must satisfy the following properties:

$$p_X(x) \geq 0, \quad \text{for all} \quad x \in \mathbb{R},$$
$$\int_{-\infty}^{\infty} p_X(x)dx = 1.$$

Gaussian Distribution

The *Gaussion* distribution over $\mathbb{R}$ is defined by its probability density function (PDF):

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

where μ is the mean and σ^2 is the variance.

Multivariate Gaussian Distribution

The *multivariate Gaussian* distribution over $\mathbb{R}^d$ is defined by its PDF:

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} \det\{\boldsymbol{\Sigma}\}^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is a positive definite covariance matrix.

Properties of the Gaussian Distribution

- **Linear transformations**:

If $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ with $\boldsymbol{A} \in \mathbb{R}^{p \times d}$ and $\boldsymbol{b} \in \mathbb{R}^p$, then $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$;

- **Marginalization**:

If $\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix}$ with $\boldsymbol{x} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right)$, then $\boldsymbol{x}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$;

- **Conditioning**:

If $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix}$, then $\boldsymbol{x}_1 | \boldsymbol{x}_2 \sim \mathcal{N}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$.

Expectation and other properties

Expectation

Expectation of a function $f(x)$ with respect to a probability distribution $p(x)$:

$$\mathbb{E}_{p(x)}[f(x)] = \int f(x)p(x)\mathrm{d}x.$$

. . .

Properties of Expectation

- **Mean** of a random variable obtained by setting $f(x) = x$:

$$\mathbb{E}[x] = \int xp(x)\mathrm{d}x.$$

- **Linearity** with respect to constants a and b:

$$\mathbb{E}[af(x) + b] = a\mathbb{E}[f(x)] + b.$$

- **Variance** of a random variable obtained by setting $f(x) = (x - \mathbb{E}[x])^2$:

$$\mathbb{E}[(x - \mathbb{E}[x])^2] = \int (x - \mathbb{E}[x])^2 p(x)\mathrm{d}x.$$

Bayes' Theorem

Bayes' Theorem

Bayes' theorem is a fundamental result in probability theory that describes how to invert conditional probabilities.

Given two random variables X and Y, Bayes' theorem states that

$$\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(X = x \mid Y = y)\mathbb{P}(Y = y)}{\mathbb{P}(X = x)}$$

or, in terms of the PDFs,

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}$$

where $p(y \mid x)$ is the **posterior**, $p(x \mid y)$ is the **likelihood**, $p(y)$ is the **prior**, and $p(x)$ is the **evidence** (or **marginal likelihood**).

How to derive Bayes' theorem?

From **product rule** and **sum rule**:

$$\begin{aligned} p(x, y) &= p(y \mid x)p(x) = p(x \mid y)p(y) \\ p(x) &= \sum_y p(x, y) = \sum_y p(x \mid y)p(y). \end{aligned}$$

Then, dividing the first equation by the second, we obtain Bayes' theorem:

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}.$$

The denominator in Bayes' theorem is the **normalization constant** $p(x)$, which ensures that the posterior distribution integrates to 1.

Why is Bayes' theorem important?

- It provides a principled way to update beliefs in the light of new evidence;
- It is the foundation of Bayesian statistics and machine learning;

$$p(\text{hypothesis} \mid \text{data}) = \frac{p(\text{data} \mid \text{hypothesis})p(\text{hypothesis})}{p(\text{data})}$$

> **Example**
>
> - **Hypothesis** is the event that a patient has a disease;
> - **Data** is the event that the patient has a positive test result.
>
> Suppose: sensitivity $p(\text{positive} \mid \text{disease})$ of 99%, specificity $p(\text{negative} \mid \text{no disease})$ of 95% and a prevalence of 1% of the population, then the probability of the patient having the disease given a positive test result
>
> $$p(\text{disease} \mid \text{positive}) = \frac{p(\text{positive} \mid \text{disease})p(\text{disease})}{p(\text{positive} \mid \text{disease})p(\text{disease}) + p(\text{positive} \mid \text{no disease})p(\text{no disease})}$$