10424/624: Recitation 1

January 17, 2025

1 Probability Spaces, Random Variables, and Notation

Just as we can construct shapes using geometry or abstract vector spaces using linear algebra, probability allows us to create a mathematical construct known as a "probability space". This is composed of three elements:

1. The sample space Ω
2. The event space $\mathcal{A}$ or $\mathcal{F}$
3. The probability function $\mathcal{P}$

Probability Axioms This space requires the following axioms in order to be valid:

- $0 \leq \mathcal{P}(A) \leq 1$ for all $A \in \mathcal{A}$
- $P(\Omega) = 1$
- $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ for any collection of disjoint events in the event space $A_1, A_2, ...,$

A **random variable** is a function that maps from the sample space to the reals. Intuitively, this means that it assigns a real number to each outcome in the samples space. Symbolically, we can write this $X : \Omega \to \mathcal{R}$.

2 Discrete and Continuous Probabilities

The event space $\mathcal{A}$ may be either discrete or continuous, which will affect how we describe it and what a valid probability function $\mathcal{P}$ might look like.

Discrete Probabilities For a discrete probability space, the **probability mass function** (PMF) describes the likelihood that a random variable takes on a given value:

$$f(x) = P(X = x)$$

Another way to describe the probability of a discrete random variable is through the **cumulative distribution function** (CDF)

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} p(x_i)$$

Example 6.1
We assume that the reader is already familiar with computing probabilities of intersections and unions of sets of events. A gentler introduction to probability with many examples can be found in chapter 2 of Walpole et al. (2011).

Consider a statistical experiment where we model a funfair game consisting of drawing two coins from a bag (with replacement). There are coins from USA (denoted as \$) and UK (denoted as £) in the bag, and since we draw two coins from the bag, there are four outcomes in total. The state space or sample space Ω of this experiment is then (\$, \$), (\$, £), (£, \$), (£, £). Let us assume that the composition of the bag of coins is such that a draw returns at random a \$ with probability 0.3.

The event we are interested in is the total number of times the repeated draw returns \$. Let us define a random variable X that maps the sample space Ω to $\mathcal{T}$, which denotes the number of times we draw \$ out of the bag. We can see from the preceding sample space we can get zero \$, one \$, or two \$s, and therefore $\mathcal{T} = \{0, 1, 2\}$. The random variable X (a function or lookup table) can be represented as a table like the following:

$$X((\$, \$)) = 2 \tag{6.1}$$
$$X((\$, £)) = 1 \tag{6.2}$$
$$X((£, \$)) = 1 \tag{6.3}$$
$$X((£, £)) = 0\,. \tag{6.4}$$

Since we return the first coin we draw before drawing the second, this implies that the two draws are independent of each other, which we will discuss in Section 6.4.5. Note that there are two experimental outcomes, which map to the same event, where only one of the draws returns \$. Therefore, the probability mass function (Section 6.2.1) of X is given by

$$\begin{aligned} P(X = 2) &= P((\$, \$)) \\ &= P(\$) \cdot P(\$) \\ &= 0.3 \cdot 0.3 = 0.09 \end{aligned} \tag{6.5}$$
$$\begin{aligned} P(X = 1) &= P((\$, £) \cup (£, \$)) \\ &= P((\$, £)) + P((£, \$)) \\ &= 0.3 \cdot (1 - 0.3) + (1 - 0.3) \cdot 0.3 = 0.42 \end{aligned} \tag{6.6}$$
$$\begin{aligned} P(X = 0) &= P((£, £)) \\ &= P(£) \cdot P(£) \\ &= (1 - 0.3) \cdot (1 - 0.3) = 0.49\,. \end{aligned} \tag{6.7}$$

Figure 1: An example from Deisenroth et al. Chapter 6

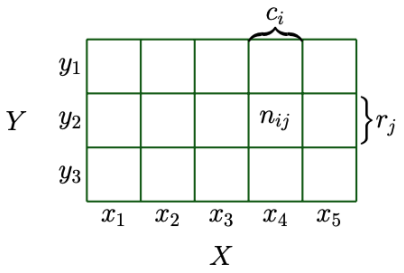

Figure 2: Visualization of a discrete joint distribution. Credit: Deisenroth et al, Figure 6.2

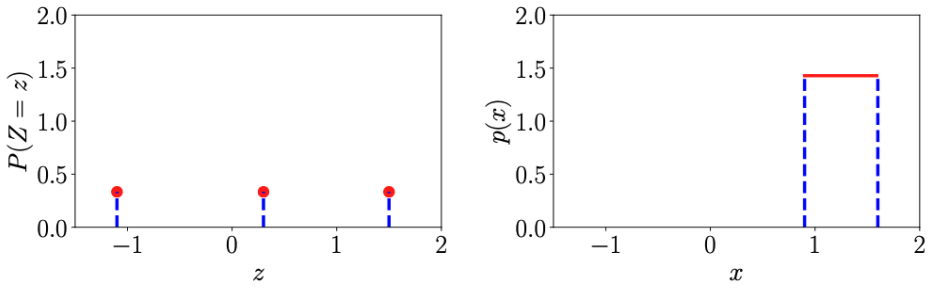

Figure 3: Examples of probability functions for continuous and discrete distributions. Credit: Deisenroth et al, Figure 6.3

Continuous Probabilities The probability mass function of a continuous function is not defined. Instead, we discuss the **probability density function** (PDF):

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx \qquad \forall a \leq b$$

We can also discuss the CDF of a continuous function:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\,dx$$

Note that a continuous probability function need not be less than or equal to 1, only have an area under the curve of 1.

3 Rules of Probability

The "**addition rule**" and "**product rule**" allow us to speak of the likelihood of two events in the same sample space. If we want to define the likelihood that either event x OR event y occurs, we can write

$$P(x \cup y) = P(x) + P(y) - P(x \cap y) \qquad \textbf{addition rule}$$

If we want to write the likelihood that event x AND y both occur, we can write

$$p(x \cap y) = p(y \mid x)p(x) \qquad \textbf{product rule}$$

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

The notation $p(y|x)$ describes a **conditional distribution**. It describes the likelihood of an event y occurring once we already know that the random variable X takes on value x.

If $p(y|x) = p(y)$, the two events x and y are considered to be **independent**. This allows us to write $p(x \cap y) = p(x)p(y)$. Note that this is not generally true, only for independent events.

Discussing the likelihood that two events both occur is so common that we can also use the notation $p(x, y)$ instead of $p(x \cap y)$. This is known as the "**joint distribution**".

If we want to obtain a **marginal distribution** from a joint distribution we can "marginalize out" or "sum out" a variable:

$$p(x) = \begin{cases} \sum_{y \in \mathcal{Y}} p(x, y) & \text{if } \mathcal{Y} \text{ is discrete,} \\ \int_{\mathcal{Y}} p(x, y)\, \mathrm{d}y & \text{if } \mathcal{Y} \text{ is continuous.} \end{cases}$$

Finally, we can use the fact that $p(x, y) = p(x|y)p(y) = p(y|x)p(x)$ to derive **Bayes' rule**

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

4 Summary Statistics and Independence

A statistic S for a random variable $X \in \mathcal{X}$ is a deterministic function $S : \mathcal{X} \to \mathcal{Y}$ of the random variable.

Means and covariances The **expected value** of a function $g : \mathbb{R} \to \mathbb{R}$ of a continuous random variable $X \sim p$ is

$$\mathbb{E}_X[g(X)] = \int_{\mathcal{X}} g(x)p(x)dx$$

For a discrete random variable, the integral can be replaced by a sum:

$$\mathbb{E}_X[g(X)] = \sum_{x \in \mathcal{X}} g(x)p(x)$$

For a vector-valued random variable $X = (X_1, \ldots, X_d)^T$, the expectation is computed element-wise: $\mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])^T$.

Other statistics: The **median** of a univariate statistic is that value (or set of values) where the cumulative distribution function $P(x) = \mathbb{P}(X \leq x)$ is equal to $1/2$. Intuitively, it divides the domain into two sections which have equal weight. The **mode** is simply the most frequently occurring value – it is the point (or set of points) where $p(x)$ is the highest.

Linearity of expectation: If $f(\mathbf{x}) = ag(\mathbf{x}) + bh(\mathbf{x})$, where $a, b \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^d$, then

$$\mathbb{E}_X[f(X)] = a\mathbb{E}_X[g(X)] + b\mathbb{E}_X[h(X)]$$

Covariance: The covariance between two random variables $X, Y \in \mathbb{R}$ is given by

$$\text{Cov}[X, Y] = \mathbb{E}_{X,Y}\left[(X - \mathbb{E}_X[X])(Y - \mathbb{E}_Y[Y])\right] = \text{Cov}[Y, X]$$

We also have the useful relation $\text{Cov}[X, Y] = \mathbb{E}_{X,Y}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y]$. The variance of X is simply its covariance with itself, and $\text{Cov}[X, Y] \in \mathbb{R}^{(d\times d)}$.

In multiple dimensions, we have the covariance between $X, Y \in \mathbb{R}^d$ as:

$$\text{Cov}[X, Y] = \mathbb{E}_{X,Y}\left[(X - \mathbb{E}_X[X])(Y - \mathbb{E}_Y[Y])^T\right] = \text{Cov}[Y, X]^T$$

The covariance matrix is the covariance of X with itself:

$$\mathbb{V}[X] = \text{Cov}[X, X] = \mathbb{E}_X\left[(X - \mathbb{E}_X[X])(X - \mathbb{E}_X[X])^T\right] = \mathbb{E}_X[XX^T] - \mathbb{E}_X[X]\mathbb{E}_X[X]^T$$

A quantity related to univariate covariance is the correlation (also called the correaltion coefficient):

$$\text{corr}[X, Y] = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Cov}[X, X] \cdot \text{Cov}[Y, Y]}} \in [-1, 1]$$

Empirical means and covariances The above statistics (mean and covariance) are defined over the underlying distribution $p(x)$; they are also called the *population* statistics. We also have the empirical versions of these statistics computed over data $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^d$:

$$\text{Empirical mean}: \quad \overline{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$$

$$\text{Empirical covariance matrix}: \quad \mathbf{\Sigma} = \frac{1}{N}\sum_{i=1}^{n} (\mathbf{x}_i - \overline{x})(\mathbf{x}_i - \overline{x})^T$$

We note that $\mathbf{\Sigma}$ is also symmetric and positive semi-definite.

Three expressions for the variance: The variance of a single univariate random variable X can be expressed in various ways:

1. Standard definition:

$$\mathbb{V}_X[X] := \mathbb{E}_X[(X - \mu)^2], \quad \text{where } \mu = E_X[X]$$

2. Raw score formula:

$$\mathbb{V}_X[X] := \mathbb{E}_X[X^2] - \mu^2, \quad \text{where } \mu = E_X[X]$$

3. Sum of squared differences: The empirical covariance can be written as

$$\begin{aligned}
\Sigma &= \frac{1}{N}\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} \left(X_i^2 + \overline{X}^2 - 2X_i\overline{X}\right) \\
&= \frac{1}{N}\left(\sum_{i=1}^{N} X_i^2 + \sum_{i=1}^{N} \overline{X}^2 - 2\sum_{i=1}^{N} X_i\overline{X}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{N}\left(\sum_{i=1}^{N} X_i^2 + N\overline{X}^2 - 2N\overline{X}^2\right) \\
&= \frac{1}{N}\left(\sum_{i=1}^{N} X_i^2 - N\overline{X}^2\right) \\
&= \frac{1}{N}\left(\sum_{i=1}^{N} X_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} X_i\right)^2\right) \\
&= \frac{1}{N}\left(\sum_{i=1}^{N} X_i^2 - \frac{1}{N}\sum_{i,j} X_i X_j\right) \\
&= \frac{1}{2N^2}\left(2N\sum_{i=1}^{N} X_i^2 - 2\sum_{i,j} X_i X_j\right) \\
&= \frac{1}{2N^2}\left(N\sum_{i=1}^{N} X_i^2 + N\sum_{j=1}^{N} X_j^2 - 2\sum_{i,j} X_i X_j\right) \\
&= \frac{1}{2N^2}\left(\sum_{i,j} X_i^2 + N\sum_{i,j} X_j^2 - 2\sum_{i,j} X_i X_j\right) \\
&= \frac{1}{2N^2}\sum_{i,j}\left(X_i - X_j\right)^2
\end{aligned}
$$

Statistical independence Two random variables X, Y are independent if and only if $p(x, y) = p(x)p(y)$. If X, Y are independent, then

- $p(y \mid x) = p(y)$, and $p(x \mid y) = p(x)$
- $\mathrm{Cov}[X, Y] = 0$
- $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y] = \mathbb{V}[X - Y]$

Conditional independence: Two random variables x, y are conditionally independent given z, if

$$p(x, y \mid z) = p(x \mid z)p(y \mid z)$$

More Useful Identities For two random variables X, Y, we have

$$\begin{aligned}\mathbb{V}[X+Y] &= \mathbb{V}[X] + \mathbb{V}[Y] + \text{Cov}[X,Y] + \text{Cov}[Y,X] \\ \mathbb{V}[X-Y] &= \mathbb{V}[X] + \mathbb{V}[Y] - \text{Cov}[X,Y] - \text{Cov}[Y,X]\end{aligned}$$

For a deterministic affine transformation $y = Ax + b$, we have

$$\begin{aligned}\mathbb{E}_X[y] &= A\mathbb{E}_X[X] + b \\ \mathbb{V}[X] &= \mathbb{E}_X\left[(AX - A\mathbb{E}_X[X])(AX - A\mathbb{E}_X[X])^T\right] \\ &= \mathbb{E}_X\left[A(X - \mathbb{E}_X[X])(X - \mathbb{E}_X[X])^T A^T\right] \\ &= A\mathbb{E}_X\left[(X - \mathbb{E}_X[X])(X - \mathbb{E}_X[X])^T\right] A^T = A\mathbb{V}[X]A^T\end{aligned}$$

5 Citations

Much of the material in this recitation was adapted from the following two resources:

1. Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). Mathematics for Machine Learning. Cambridge: Cambridge University Press. ’
2. 10605 Machine Learning with Large Datasets F24 Recitation 4 – Probability Review (link)