Disclaimer: These notes have only been lightly proofread.

2.1 Probability review

2.1.1 Motivation

Question: Why do we use probability in data science?

Answer: Probability theory is a principled framework to model uncertainty.

Question: Where does uncertainty come from?

Answer: There are several sources:

1. it can be intrinsic to certain phenomenon (e.g. quantum mechanics);
2. reasoning about future events;
3. we can only get partial information about some complex phenomenon:
   (a) e.g. throwing a dice, it is hard to fully observe the initial conditions;
   (b) for an object recognition model, a mapping from pixels to objects can be incredibly complex.

2.1.2 Notation

Note that probability theorists and the graphical models community both use a lot of notational shorthands. The meaning of notations often has to be inferred from the context. Therefore, let’s recall a few standard notations.

Random variables will be noted $X_1, X_2, X_3, \ldots$, or sometimes $X, Y, Z$. Usually, they will be real-valued.

$x_1, x_2, x_3, \ldots$ (or $x, y, z$), will denote the realizations of the former random variables (the values the $X$s can take).
Formally

Let us define $\Omega$, a sample space of elementary events, $\{\omega_1, \omega_2, \omega_3, \ldots\}$.

Then a random variable is a (measurable\(^2\)) mapping $X : \Omega \mapsto \mathbb{R}$.

Then, a probability distribution $P$ is a mapping $P : \mathcal{E} \mapsto [0, 1]$, where $\mathcal{E}$ is the set of all subsets of $\Omega$, i.e. the set of events (i.e. $2^\Omega$, i.e. a $\sigma$-field\(^3\)) ; such that

$$
\begin{align*}
-P(E) &\geq 0 \quad \forall E \in \mathcal{E} \\
-P(\Omega) &= 1 \\
-P(\bigcup_{i=1}^{\infty} E_i) &= \sum_{i=1}^{\infty} (E_i) \quad \text{when } E_1, E_2, \ldots \text{ are disjoint.}
\end{align*}
$$

Kolmogorov axioms

Therefore, a probability distribution on $\Omega$ induces a probability distribution on the image of $X$\(^4\) : $\Omega_X \overset{\Delta}{=} X(\Omega)$. An event $\{x\}$ for $x \in \Omega_X$ thus gets the probability

$$
Px(\{x\}) = P(\{\omega : X(\omega) = x\})
= P(X^{-1}(\{x\}))
= P\{X = x\} \quad \text{(shorthand)}
= p(x) \quad \text{actually used shorthand, even more ambiguous}
$$

where $X^{-1}(A) \overset{\Delta}{=} \{\omega : X(\omega) \in A\}$.

Example

In the case of a dice roll, $\Omega = \{1, 2, \ldots, 6\}$. Let’s consider two random variables : $X$ measures whether the dice result is even. $Y$ measures whether the dice result is odd.

Formally, $X = 1_{\{2,4,6\}}$, and $Y = 1_{\{1,3,5\}}$ where

$$
1_A(\omega) \overset{\Delta}{=} \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}
$$

is the indicator function on $A$.

\(^1\)temporarily assumed to be a countable set
\(^2\)Wikipedia
\(^3\)the $\sigma$-field formalism is necessary when $\Omega$ is uncountable, which happens as soon as we consider a continuous random variable.
\(^4\)The image of $X$ is the set of the possible outputs of $X : X(\Omega) = \{x : \exists \omega \in \Omega \text{ s.t. } X(\omega) = x\}$
We can now define the joint distribution on \((X, Y) \in \Omega_X \times \Omega_Y\).

\[
P_{X,Y}(\{X = x, Y = y\}) = P\left(X^{-1}(\{x\}) \cap Y^{-1}(\{y\})\right)
\]

\((X, Y)\) can be called a random vector, or a vector-valued random variable, with “random variable” meant in a generalized sense.

We can represent the joint distribution as a table, such as in our running example:

<table>
<thead>
<tr>
<th>(Y)</th>
<th>(X = 0)</th>
<th>(X = 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Y = 0)</td>
<td>0</td>
<td>(\frac{1}{2})</td>
</tr>
<tr>
<td>(Y = 1)</td>
<td>(\frac{1}{2})</td>
<td>0</td>
</tr>
</tbody>
</table>

For instance: \(P(\{X = 1, Y = 0\}) = P(\{2, 4, 6\}) = \sum_{\omega \in \{2, 4, 6\}} p(\omega) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}\).

Let’s also define, in the context of a joint distribution, the marginal distribution, i.e. the distribution on components of the random vector:

\[
P\{X = x\} = \sum_{y \in \Omega_Y} P\{X = x, Y = y\}
\]

This rule is a property, deriving it from the axioms is left as an exercise for the reader.

### 2.1.3 Types of random variables

**Discrete random variables**

For a discrete random variable, \(\Omega_X\) is countable. Its probability distribution on \(\Omega_X\), \(P_X\), is fully defined by its probability mass function (aka pmf), \(P_X(\{X = x\})\), for \(x \in \Omega_X\). This notation is shortened as \(P_X(x)\), and even as \(p(x)\), “typing” \(x\) as only denoting values of the \(X\) variable. Thereby, it is possible that \(p(x) \neq p(y)\) even if \(x = y\), in the sense that \(p(x)\) means \(P_X(x)\) and \(p(y)\) means \(P_Y(y)\).

More generally, for \(\Omega_X \in \mathbb{R}\), the probability distribution \(P_X\) is fully characterized by its cumulative distribution function (aka cdf): \(F_X(x) \triangleq P_X\{X \leq x\}\).

---

\(^5\)This comma means and, the intersection of both events.
It has the following properties:

1. $F_X$ is non-decreasing;
2. $\lim_{x \to -\infty} F_X(x) = 0$;
3. $\lim_{x \to +\infty} F_X(x) = 1$.

For discrete random variables, the cumulative distribution function is piecewise constant, and has jumps.

**Continuous random variables**

For a continuous random variable, the cumulative distribution function is “absolutely continuous”, i.e. is differentiable almost everywhere, and $\exists f(x)$ s.t. $F_X(x) = \int_{-\infty}^{x} f(u)du$. Said $f$ is called the probability density function of the random variable (aka pdf). Where $f$ is continuous, $\frac{d}{dx}F_X(x) = f(x)$.

The probability density function is the continuous analog of the probability mass function of a discrete random variable (with sums becoming integrals). Hence:

<table>
<thead>
<tr>
<th></th>
<th>discrete</th>
<th>continuous</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sum_{x \in \Omega_X} p(x) = 1$</td>
<td>$\int_{\Omega_X} p(x) = 1$</td>
<td></td>
</tr>
<tr>
<td>$p = \text{prob. mass function}$</td>
<td>$p = \text{prob. density function}$</td>
<td></td>
</tr>
</tbody>
</table>

Note in the continuous case, as a density function, $p(x)$ can be greater than 1, on a sufficiently narrow interval. For instance, the uniform distribution on $[0, \frac{1}{2}]$:

$$p(x) = \begin{cases} 
2 & \text{for } x \in [0, \frac{1}{2}] \\
0 & \text{otherwise}
\end{cases}$$

### 2.1.4 Other random variable basics

**Expectation/mean**

The expectation of a random variable is

$$\mathbb{E}[X] \triangleq \sum_{x \in \Omega_X} x \cdot p(x) \quad \text{or} \quad \int_{\Omega_X} x \cdot p(x) \, dx \quad \text{(in the continuous case)}$$

**Variance**

$$Var[X] \triangleq \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Variance is a measure of the dispersion of values around the mean.
Indepandance

$X$ is independant from $Y$, noted $X \perp Y$, iff $p(x, y) = p(x)p(y) \ \forall x, y \in \Omega_X \times \Omega_Y$.
Random variables $X_1, \ldots X_n$ are mutually independant iff $p(x_1, \ldots x_n) = \prod_{i=1}^n p(x_i)$.

Conditioning

For events $A$ and $B$, suppose that $p(B) \neq 0$. We define the probability of $A$ given $B$,

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)}$$

In terms of sample space, that means we look at the subspace where $B$ happens, and in that space, we look at the subspace where $A$ also happens.

For random variables $X$ and $Y$, thus :

$$P(X = x|Y = y) \triangleq \frac{P(X = x, Y = y)}{P(Y = y)}$$

$P(Y = y) = \sum_x P(X = x, Y = y)$ is a normalization constant, necessary in order to get a real probability distribution.

By definition, we get the product rule :

$$p(x, y) = p(x|y)p(y) \quad \text{(product rule)}$$

It is always true, with the subtle point that $p(x|y)$ is undefined if $p(y) = 0$.$^6$

Bayes rule

Bayes rule is about inverting the conditioning of the variables.

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\sum_{x'} p(x', y)} \quad \text{(Bayes rule)}$$

Chain rule

By successive application of the product rule, it is always true that :

$$p(x_1, \ldots, x_n) \begin{array}{c} = p(x_{1:n-1})p(x_n|x_{1:n-1}) \\
= \cdots \\
= \prod_{i=1}^n p(x_i|x_1, \ldots, x_{i-1}) \end{array} \quad \text{(Chain rule)}$$

The last part can be simplified using the conditional independance asumptions we make, like in the case of directed graphical models.

$^6$In probability theory, we usually do not care what happens on sets with probability zero; so we are free to define $p(x|y)$ to be any value we want when $p(y) = 0$. 2-5
Conditional independance

$X$ is conditionally independant of $Y$ given $Z$, noted $X \perp Y | Z$, iff

$$p(x, y | z) = p(x | z)p(y | z) \quad \forall x, y, z \in \Omega_x \times \Omega_y \times \Omega_z \text{ s.t. } p(z) \neq 0$$

For instance, with $Z$ the probability that a mother carries a genetic disease on chromosome X, $X$ the probability for her first child to carry the disease, and $Y$ the same probability for her second child, we can say that $X$ is independant of $Y$ given $Z$ (because only the status of the mother impacts directly each child: once that is known, children’s probabilities of carrying the disease are independant from each other).

As an exercise to the reader, prove that $p(x | y, z) = p(x | z)$ when $X \perp Y | Z$. 