Review of Probability

David M. Blei
Columbia University

September 5, 2016

Before we discuss data and models, we will learn about graphical models. Graphical models are a way of using graphs to represent and compute about families of probability distributions.

In this lecture we review the concepts from probability important for understanding graphical models. The concepts are random variables, joint distributions, conditional distributions, the chain rule, marginalization, (conditional) independence, and Bayes rule. Later in the semester, we will discuss the core statistical concepts that are important for understanding how to use probability models to analyze data.

To begin, consider the "card problem" (via David MacKay)

- There are three cards: (R/R); (R/B); (B/B)
- I go through the following process.
  - Close my eyes and pick a card
  - Pick a side at random
  - Show you that side
- I show you R. What's the probability the other side is red too?

Random variables, atoms, and events. Probability is about *random variables.* A random variable is any "probabilistic" outcome. As two examples: (a) the flip of a coin and (b) the height of someone chosen randomly from a population.

It is often useful to think of quantities that are not strictly probabilistic as random variables. For example, (a) the temperature on on December 18 of next year; the temperature on 03/04/1905; (c) the number of times the term "streetlight" appears on the internet.

Random variables take on values in a *sample space*. They can be *discrete* or *continuous*. For example:

- Coin flip: $\{H, T\}$

- Height: positive real values $(0, \infty)$
- Temperature: real values $(-\infty, \infty)$
- Number of words in a document: Positive integers $\{1, 2, \ldots\}$

We call the values of random variables *atoms*. For now we will consider discrete random variables.

Notation: We denote the random variable with a capital letter; denote a realization of the random variable with a lower case letter. E.g., X is a coin flip, x is the value (H or T) of that coin flip.

A *discrete probability distribution* assigns probability to every atom in the sample space. For example, if X is an (unfair) coin, then

$$P(X = H) = 0.7$$
$$P(X = T) = 0.3$$

The probabilities over the space must sum to one

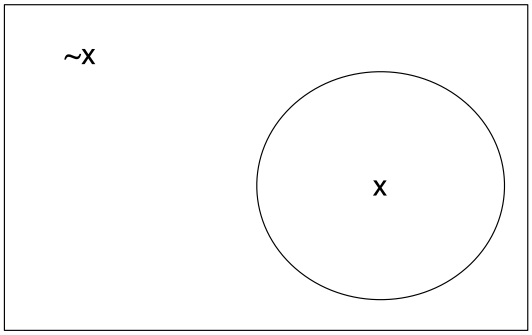
$$\sum_x P(X = x) = 1$$

(Note how we use summation. This is shorthand for "all possible values of x.")

A subset of atoms is called an *event*. (Note that atoms are events too.) The probability of a subset is a sum of probabilities of its atoms. E.g., the atoms of a die are the individual sides (1, 2, 3, 4, 5, 6). The probability that the die is bigger than 3 is

$$P(D > 3) = P(D = 4) + P(D = 5) + P(D = 6).$$

Here is a useful picture to understand atoms, events, and random variables:



- An *atom* is a point in the box; all the atoms are the *sample space*.
- The sample space is endowed with a probability function, $p(\omega)$; it sums to one, $\sum_{\omega \in \Omega} p(\omega) = 1$.
- For example, each atom might be each possible outcome of 2 dice. The probability of each atom is the probability of obtaining exactly that configuration.
- An *event* is a subset of atoms (e.g., the sum is equal to seven). E.g., two events in this picture are x and $\bar{x}$.
- The probability of an event is sum of probabilities of its atoms.

Random variables. Technically, a *random variable* is a function from the sample space to the reals. However, for our purposes now, a random variable is a partition of the sample space into unique values. The distribution of the random variable is determined by the underlying probability measure on the sample space.

(We will omit measure theory. A cartoon view of measure theory is that it generalizes this story to continuous spaces, and invokes the real analysis needed to do so.)

Joint distributions. The central object of this class is the *joint distribution*. Typically, we reason about collections of random variables. The joint distribution is a distribution over the configuration of all the random variables in the collection.

For example, imagine flipping 4 coins. The joint distribution is over the space of all possible outcomes of the four coins.

$$
\begin{gathered}
P(HHHH) = 1/16 \\
P(HHHT) = 1/16 \\
P(HHTH) = 1/16 \\
\ldots
\end{gathered}
$$

Notice you can think of this as a single random variable with 16 values.

Q **|** *What is an atom here? What is an event?*

Conditional distributions. A *conditional distribution* is the distribution of a random variable given the value of other random variables.

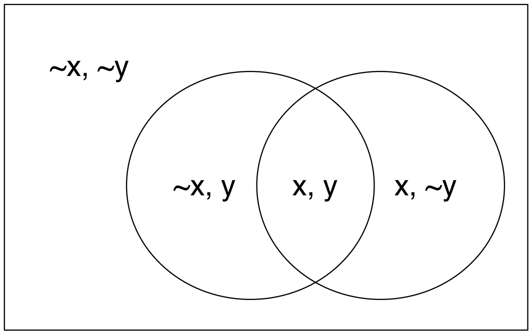
The quantity $P(X = x \mid Y = y)$ is the probability that $X = x$ when $Y = y$. For example,

$$\begin{aligned} P(\text{I listen to Steely Dan}) &= 0.5 \\ P(\text{I listen to Steely Dan} \mid \text{Toni is home}) &= 0.1 \\ P(\text{I listen to Steely Dan} \mid \text{Toni is not home}) &= 0.7 \end{aligned}$$

Note that $P(X = x|Y = y)$ is a different distribution for each value of y,

$$\begin{aligned} &\sum_x P(X = x \mid Y = y) = 1 \\ &\sum_y P(X = x \mid Y = y) \neq 1 \quad \text{(at least, not necessarily).} \end{aligned}$$

To define the conditional probability, we use this picture:



We define the conditional probability as

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

which holds when $P(Y) > 0$. Here the Venn diagram represents the space of cross-product atoms, i.e., the space of X coupled with the space of Y. The conditional probability is the relative probability of $X = x$ in the space where $Y = y$.

Now we can solve the card problem. Let X_1 be the random side of the random card I chose. Let X_2 be the other side of that card. Compute $P(X_2 = \text{red} \mid X_1 = \text{red})$

$$P(X_2 = \text{red} \mid X_1 = \text{red}) = \frac{P(X_1 = \text{red}, X_2 = \text{red})}{P(X_1 = \text{red})} \tag{1}$$

- The numerator is 1/3: Only one card has two red sides.

- The denominator is 1/2: Three sides of the six are red.
- The solution is 2/3.

The chain rule. Conditional probability lets us derive the *chain rule*. The chain rule defines the joint distribution as a product of conditionals,

$$\begin{aligned} P(X, Y) &= P(X, Y)\frac{P(Y)}{P(Y)} \\ &= P(X \mid Y)P(Y) \end{aligned}$$

In general, for any set of N variables

$$P(X_1, \ldots, X_N) = \prod_{n=1}^{N} P(X_n \mid X_1, \ldots, X_{n-1}).$$

Marginalization. Given a collection of random variables, we are often only interested in a subset of them. For example, compute $P(X)$ from a joint distribution $P(X, Y, Z)$. This is called the *marginal* of X. We compute it with *marginalization*,

$$P(X) = \sum_y \sum_z P(X, Y = y, Z = z)$$

We can see how this works from the Venn diagram. Atoms are configurations of $\{x, y, z\}$ and so $P(X = x)$ sums the distribution over the set of configurations where $X = x$.

We can also derive marginalization from the chain rule,

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

Bayes rule. From the chain rule and marginalization, we obtain *Bayes rule*,

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{\sum_y P(X \mid Y = y)P(Y = y)}.$$

Example: let Y be a disease and X be a symptom. From the distribution of the symptom given the disease $P(X \mid Y)$ and the probability of the disease $P(Y)$, we can compute the (more useful) distribution of the disease given the symptom $P(Y \mid X)$.

Bayes rule is important in *Bayesian statistics*, where Y is a parameter that controls the distribution of X.

Independence. Random variables X and Y are *independent* if knowing about X tells us nothing about Y,

$$P(Y \mid X) = P(Y).$$

This means that their joint distribution factorizes,

$$X \perp\!\!\!\perp Y \iff P(X, Y) = P(X)P(Y).$$

Why? It is because of the chain rule,

$$\begin{aligned} P(X, Y) &= P(X)P(Y \mid X) \\ &= P(X)P(Y). \end{aligned}$$

Notice this is a property of the probability function on the sample space. This perspective is important to understanding how graphical models work.

Some examples of independent random variables,

- Flipping a coin once / flipping the same coin a second time
- You use an electric toothbrush / blue is your favorite color

Examples of not independent random variables:

- Registered as a Republican / voted for Trump
- The color of the sky / The time of day

Q | *Are these independent?*

- Two twenty-sided dice
- Rolling three dice and computing $(D_1 + D_2, D_2 + D_3)$
- # enrolled students and the temperature outside today
- # attending students and the temperature outside today

Suppose we have two coins, one biased and one fair,

$$P(C_1 = H) = 0.5 \quad P(C_2 = H) = 0.7.$$

We choose one of the coins at random $Z \in \{1, 2\}$, flip C_Z twice, and record the outcome (X, Y). *Q | Are X and Y independent?* What if we knew the value of Z, which coin was flipped?

Conditional independence. Variables X and Y are *conditionally independent* given Z if

$$P(Y \mid X, Z = z) = P(Y \mid Z = z).$$

for all possible values of z. Again, this implies a factorization

$$X \perp\!\!\!\perp Y \mid Z \iff P(X, Y \mid Z = z) = P(X \mid Z = z)P(Y \mid Z = z),$$

for all possible values of z.

Summary. Next we will discuss graphical models, which are a formalism for computing with and reasoning about families of joint distributions. The concepts from probability important for learning about graphical models are

- A random variable
- A joint distribution of a collection of random variables
- A conditional distribution
- The chain rule
- Marginalization
- Independence & conditional independence

Later, we will review statistical concepts and models—such as maximum likelihood estimation and Bayesian statistics—to relate graphical models to algorithms for making estimates from data. At that point, will review the fundamental concepts of expectation, conditional expectation, and continuous random variables.