A Quick Review of Probability

David M. Blei
Columbia University

October 3, 2015

¶ Before we discuss data and models, we will review probability and learn about graphical models. Our review of probability will be too elementary for many of you, but it is important that we are all up to speed.

¶ The card problem (MacKay, 2003)

- There are three cards
  - Red/Red
  - Red/Black
  - Black/Black
- 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 red. What's the probability the other side is red too?

¶ The random variable

- Probability is about *random variables.*
- A random variable is any "probabilistic" outcome.
- For example,
  - The flip of a coin
  - The height of someone chosen randomly from a population
- We'll see that it's sometimes useful to think of quantities that are not strictly probabilistic as random variables.
  - The temperature on on December 18 of next year.
  - The temperature on 03/04/1905
  - The number of times "streetlight" appears in a document

¶ Sample spaces & atoms

- Random variables take on values in a *sample space*.

- They can be *discrete* or *continuous*:
  - 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 *atoms*.
- 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.

¶ Discrete distributions

- A discrete distribution assigns a probability to every atom in the sample space. For now, we will focus on discrete random variables.
- For example, if X is an (unfair) coin, then

$$\begin{aligned} P(X = H) &= 0.7 \\ P(X = T) &= 0.3 \end{aligned}$$

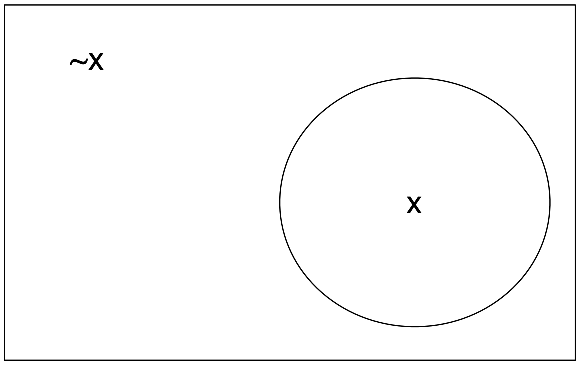
- The probabilities over the entire space must sum to one

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

  (Note how we use the summation notation.)
- Probabilities of disjunctions are sums over part of the space. E.g., the probability that a die is bigger than 3:

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

¶ A useful picture



- An *atom* is a point in the box
- An *event* is a subset of atoms (e.g., $d > 3$)
- The probability of an event is sum of probabilities of its atoms.

¶ Joint distributions

- This is the central object of this class.
- Typically, we consider collections of random variables.
- The joint distribution is a distribution over the configuration of all the random variables in the ensemble.
- For example, imagine flipping 4 coins. The joint distribution is over the space of all possible outcomes of the four coins.

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

- You can think of it as a single random variable with 16 values.
- Q: What is an atom here? What is an event?

¶ Conditional distributions

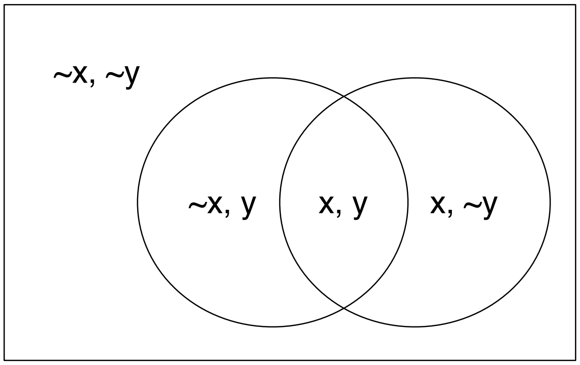
- A *conditional distribution* is the distribution of a random variable given the value of other random variables.
- $P(X = x \mid Y = y)$ is the probability that $X = x$ when $Y = y$.
- For example,

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

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

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

¶ Definition of conditional probability



- Conditional probability is defined 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 (the space of X coupled with the space of Y).
- The conditional probability is the relative probability of $X = x$ in the space where $Y = y$.

¶ Return to the card problem

- 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*, which let's us define 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

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

¶ 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 $P(X, Y, Z)$
- Can do this with *marginalization*

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

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

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

¶ Bayes rule

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

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

- Let Y be a disease and X be a symptom. From the distribution of the symptom given the disease $P(X \,|\, Y)$ and the probability of the disease $P(Y)$, we can compute the (more useful) distribution of the disease given the symptom $P(Y \,|\, X)$.
- 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.

$$P(Y \,|\, X) = P(Y)$$

- This means that their joint distribution factorizes,

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

- Why? The chain rule

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

¶ Examples of independence

- 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 Bush in the last election
  - The color of the sky / The time of day

¶ 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

¶ Two coins

- 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).
- Question: Are X and Y independent?
- What if we knew which coin was flipped Z?

¶ Conditional independence

- X and Y are *conditionally independent* given Z.

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

  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.

¶ Next: We will discuss graphical models, which are a formalism for computing with and reasoning about families of joint distributions.

¶ The concepts that are important for learning about graphical models:

- A random variable
- A joint distribution of a collection of random variables
- A conditional distribution
- The chain rule
- 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.

References

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.