Lecture 2: Probability and Statistics

February 13, 2008

2/7/08

Scribe: Jonathan "JP" Paranada

1 What is Probability?

1.1 Definition of Probability and Random Variables

Probability is the study of *random variables*, (a r .v. being any probabilistic outcome). Some examples of r.v.'s include:

- A coin toss. Assuming a fair coin, this is a completely random event.
- The number of visitors to a certain store in one day. This is not exactly random - if we knew at the beginning of the day how many people wanted to go to the store, it would not be a r.v. But since this information is unknown, this is a probabilistic outcome.
- The high temperature on 2/7/2013. Again, this is information we do not know.
- The high temperature on 3/4/1905. Even though we could look this information up, it is still probabilistic.

1.2 Sample Space

R.v.'s take up values in a *sample space*. This sample space can be *discrete* or *continuous*, and *finite* or *infinite*. For example:

- A coin flip has sample space {h, t}. This is discrete and finite.
- The number of visitors to a store has the sample space $\{0, 1, ..., \infty\}$. This is infinite and discrete.
- A temperature at a certain time has the sample space $\Re$. This is infinite and continuous.

The values in a sample space are called *atoms*.

2 Notation

- A random variable is denoted by a capital letter: X.
- The realization of a r.v. is lower case: x.

3 Discrete Distribution

- A *discrete distribution* assigns a probability p to every atom in the space. For example, an unfair coin could have p(X=h) = 0.7, p(X=t) = 0.3.
- The probabilities must sum to one, i.e. $\sum_x p(X = x) = 1$.

4 Events

- Consider a space of atoms, which we can represent with a box. Then an *event* is a subset of these atoms.
- The *probability of an event* is the sum of atomic probabilities in that subset, i.e. $\sum_{x \in a} p(X = x) = p(a)$.

5 Joint Distributions

- Typically, we are interested in collections of r.v.'s (e.g. visitors in a store *every* day).

A *joint distribution* is the distribution over a configuration of all r.v.'s in an ensemble. The *joint probability* is the probability that, for N events, those N events will occur together.

- For example: p(h, h, h, h) = .0625, p(t, h, h, h) = .0625, ..., p(t, t, t, t) = .0625

We read the joint probability p(X = x, Y = y) as "the probability of x and y".

6 Conditional Distributions

A *conditional distribution* is a distribution of a r.v. given some evidence/prior knowledge. This is denoted p(X = x | Y = y) (read: "the probability of x given y"). For example:

p(David Blei listens to Steely Dan) = 0.5
p(Dave listens to S.D. | Toni is home) = 0.1
p(Dave listens to S.D. | Toni is not home) = 0.7

Note that there is one distribution per value of y. In each distribution, all probabilities p(X = x) must sum to one. That is,

$\sum_x p(X = x \mid Y = y) = 1$ but
$\sum_y p(X = x \mid Y = y) \neq 1$ necessarily.
We define the *conditional probability* in this way:

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

where p(Y=y) > 0.

7 The Chain Rule

$$p(X, Y) = \frac{p(X, Y)p(Y)}{p(Y)} = p(X \mid Y)p(Y)$$

The chain rule gives us a relation between a joint distribution and a conditional distribution. It can also be generalized as:

$$p(X_1, ..., X_N) = p(X_1) \prod_{n=2}^{N} p(X_n \mid X_1, ..., X_{n-1})$$

8 Marginalization

Given a set of r.v.'s, we are often interested in a subset of them. That is, we fix some variables and let others vary. This can be expressed as:

$$p(X) = \sum_y \sum_z p(X, y, z)$$

Here we sum over fixed y and z while X is unknown.

9 Bayes' Rule

Bayes' rule gives us a relation between a conditional distribution and the "reverse" conditional distribution, i.e. a relationship between p(X|Y) and p(Y|X).

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

The denominator is p(X), so we can alternately write:

$$p(Y \mid X) = \frac{p(X \mid Y)p(Y)}{p(X)}$$

To derive Bayes' rule, note that the chain rule implies the latter equation (since p(X, Y) = p(X|Y)p(Y) = p(Y|X)p(X)), and marginalizing out y in the denominator combined with the definition of conditional probability yields the former equation.

10 Independence

10.1 Definition

R.v.'s are *independent* (notation: $\perp$, but with two vertical lines) if knowing one doesn't give us any information about the other(s). That is, p(X|Y = y) = p(X) for all y.

- This means that the joint factorizes as the product of the marginals: p(X, Y) = p(X|Y)p(Y) = p(X)p(Y).

Examples of r.v.'s that are not independent include:

- Whether it rains and whether you go to the beach
- A person's height and a person's sex

Examples of r.v.'s that are independent include:

- The result of rolling two dice
- Whether it rains tomorrow and who the next U.S. president is

10.2 Conditional Independence and the Two Coins Example

Say we have two coins, one fair and one unfair, with $p(C_1 = H) = .5$, $p(C_2 = H) = .7$. We will

1. Choose one coin at random, i.e. pick some $z \in \{1, 2\}$that determines our choice of coin C_z.
2. Flip $C_z$twice to get two results X, Y.

If we knew z, then X and Y would be independent (each with probabilities determined by the coin we had chosen). But say we did not know z and the first coin flip was heads. Then the second flip is more likely to be heads. Thus X and Y are not independent.

Formally, we can state that X and Y are *conditionally independent* if, when given information z, they become independent. That is, p(Y|X, Z = z) = p(Y|Z = z).

This also implies that p(Y, X|Z = z) = p(Y|Z=z)p(X|Z=z) (since the two are independent given z, the joint factorizes).