Lisa Yan and Jerry Cain
CS109

Lecture Notes #11
October 7, 2020

Joint Distributions

Based on a chapter by Chris Piech

Pre-recorded lecture: Binomial approximation from Lecture 10, Section 1 from this lecture.
In-lecture: More Binomial approximation, Linearity of Expectation proof (in Lecture 13, Joint statistics).

1 Joint Distributions

Often you will work on problems where there are several random variables (often interacting with one another). We are going to start to formally look at how those interactions play out.

For now we will think of joint probabilities with two events $X = a$ and $Y = b$. For this week, we will assume both X and Y are discrete random variables, and we will tackle the continuous case next week.

Discrete Case

In the discrete case, a joint probability mass function tells you the probability of any combination of events $X = a$ and $Y = b$:

$$p_{X,Y}(a, b) = P(X = a, Y = b)$$

This function tells you the probability of all combinations of events (the "," means "and"). If you want to back calculate the probability of an event only for one variable you can calculate a "marginal" from the joint probability mass function:

$$p_X(a) = P(X = a) = \sum_y p_{X,Y}(a, y)$$
$$p_Y(b) = P(Y = b) = \sum_x p_{X,Y}(x, b)$$

In the continuous case a joint probability density function tells you the relative probability of any combination of events $X = a$ and $Y = y$.

In the discrete case, we can define the function $p_{X,Y}$ non-parametrically. Instead of using a formula for p we simply state the probability of each possible outcome.

2 Multinomial Distribution

Say you perform n independent trials of an experiment where each trial results in one of m outcomes, with respective probabilities: $p_1, p_2, \ldots, p_m$ (constrained so that $\sum_i p_i = 1$). Define X_i to be the number of trials with outcome i. A multinomial distribution is a closed form function that answers the question: What is the probability that there are c_i trials with outcome i. Mathematically:

$$P(X_1 = c_1, X_2 = c_2, \ldots, X_m = c_m) = \binom{n}{c_1, c_2, \ldots, c_m} p_1^{c_1} p_2^{c_2} \ldots p_m^{c_m}$$

Example: Dice

A 6-sided die is rolled 7 times. What is the probability that you roll: 1 one, 1 two, 0 threes, 2 fours, 0 fives, 3 sixes (disregarding order).

$$P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 2, X_5 = 0, X_6 = 3) = \frac{7!}{2!3!}\left(\frac{1}{6}\right)^1\left(\frac{1}{6}\right)^1\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^3$$
$$= 420\left(\frac{1}{6}\right)^7$$

3 Federalist Papers

We can use Multinomial random variables to model word counts of an unknown document, and thereby determine authorship. Read on for more!

The Federalist Papers is a body of essays written between 1787 and 1788 by Alexander Hamilton, James Madison, and John Jay under the collective pseudonym "Publius." Suppose that James Madison and Alexander Hamilton both claim to have written Federalist Paper No. 53 (we do not consider John Jay as a possible author in this problem). We can use the contents of this document (with unknown author) along with historical evidence of the two possible authors' penmanship to determine who the author is probabilistically:

(1) *Collect data:* Estimate p_i, the probability that Hamilton writes word i, as the *frequency* of word i appearing in Hamilton's previous work (`hamilton.txt`). Use the same process to estimate q_i, the probability that Madison writes word i, from Madison's previous work (`madison.txt`).

(2) *Model unknown document using the Bag of Words model:* The *Bag of Words* model is the natural language processing model we discussed in class, which models the word counts in an unknown document using a *Multinomial random variable*, where the probability of writing word i is either p_i (given that the author is Hamilton) or q_i (given that the author is Madison).

(3) *Make a prediction:* Use Bayes' Theorem to determine the probability that Hamilton wrote the unknown document (Federalist Paper No. 53, in `unknown.txt`). If Hamilton was more likely than Madison to be the author given the unknown document's word counts, then predict that Hamilton is the author. Otherwise, Madison is the author.

Define three events: H is the event that Hamilton wrote the document, M is the event that Madison wrote the document, and D is the event that a document has the collection of words (i.e., bag of words) observed in Fedaralist Paper 53. We would like to know whether $P(H|D)$ is larger than $P(M|D)$:

$$P(H|D) > P(M|D) \tag{1}$$

If Equation 1 holds, then we predict Hamilton (otherwise we predict Madison).

Suppose that there are *n total* words in the unknown document; further suppose that there are *m unique* words in the document, where c_i is the number of times word i appears and $\sum_{i=1}^{m} c_i = n$. It can be shown that Equation 1 is equivalent to Equation 2 below, with a few assumptions:

$$\frac{\prod_{i=1}^{m} p_i^{c_i}}{\prod_{i=1}^{m} q_i^{c_i}} > 1 \tag{2}$$

If Equation 2 holds, then Equation 1 holds—and therefore we predict Hamilton. Otherwise, we predict Madison. This equivalence relies on a few assumptions:

- Prior to seeing the unknown document, Hamilton and Madison are equally likely to have written the document.
- Given the author of a document, the document's collection of words can be modeled as a Multinomial random variable, where the probability of writing word i is p_i (given that the author is Hamilton) or q_i (given that the author is Madison), where $i = 1, \ldots, m$.

It turns out that it is computationally intractable to compute Equation 2 because of an issue called arithmetic underflow, which we explore more on the problem set. :-) Instead, we often use the properties of logarithms to translate Equation 2 to Equation 3:

$$\sum_{i=1}^{m} c_i \log p_i - \sum_{i=1}^{m} c_i \log q_i > 0 \tag{3}$$