CS 70 Discrete Mathematics and Probability Theory
Fall 2016 Walrand Probability: An Overview

Probability is a fascinating theory. It provides a precise, clean, and useful model of uncertainty. The successes of Probability Theory in Computer Science are remarkable: data science, machine learning, artificial intelligence, voice and image recognition, and communication theory are based on that theory.

The objective of these notes is to introduce the key ideas of Probability Theory on simple examples. Hopefully, this overview will help you see the forest as you explore its different trees in the course.

1 Pick a Marble

Setup

Imagine a bag with 100 marbles that are identical, except for their color. Among those, 10 are blue, 20 are red, 30 are green, and 40 are white. You shake the bag and pick a marble without looking.

Probability

You will certainly agree that the odds that you picked a green marble are 30 out of 100. Similarly, the odds that you picked a blue marble are 10 out of 100. We say that the probability that the marble is green is $30/100 = 0.3$. We write $Pr[\text{green}] = 0.3$.

Interpretation

What does this mean precisely? Well, this is not really that obvious. Two interpretations are useful. The first interpretation is a *subjective* willingness to bet on the outcome. Imagine the following game of chance. You bet some amount and you get \$100.00 if the marble is green. How much are you willing to bet? I would be willing to bet \$30.00. The second interpretation is *frequentist*. It says that if you were to repeat this experiment (shaking the bag with 100 marbles and pick a marble without looking), you would pick a green marble about 30% of the time. Note that this is an interpretation at this point, not a theorem.

Additivity

Consider the event 'the marble is blue or green.' The odds of that event are $40/100$. We write $Pr[\text{blue or green}] = 0.4$. Note that $Pr[\text{blue or green}] = Pr[\text{blue}] + Pr[\text{green}]$. This is not surprising since the number of marbles that are blue or green is the sum of the number of blue marbles plus the number of green marbles. We say that *probability is additive*.

Conditional Probability

Assume that the marble you picked is blue or green. What are the odds that it is blue or red? Well, since you picked one of the 40 marbles that are either blue or green, that marble is blue or red only if it is one of the 10 blue marbles. Since 10 out of the 40 blue or green marbles are blue, we see that the odds that you picked a blue or red marble, *given that you picked a blue or green marble*, is $10/40$. We say that the conditional probability of blue or red given blue or green is $10/40$. We write $Pr[\text{blue or red}|\text{blue or green}] = 10/40$.

Note that

$$Pr[\text{blue or red}|\text{blue or green}] = \frac{Pr[(\text{blue or red}) \text{ and } (\text{blue or green})]}{Pr[\text{blue or green}]} = \frac{Pr[\text{blue}]}{Pr[\text{blue or green}]}.$$

Bayes' Rule

Assume that we paint a black dot on half of the blue and half of the red marbles, and also on 20% of the green and 20% of the white marbles. You pick a marble at random and are told that the marble has a black dot. What are the odds that the marble is red? To answer this question, we note that there are $5+10+6+8=29$ marbles with a black dot, out of which 10 are red. Thus, the answer is $10/29$. This calculation is an example of *Bayes' Rule*. The idea is that one specified $Pr[\text{black dot}|\text{blue}] = 0.5$ and similarly for the other colors. One also knows $Pr[\text{blue}] = 0.1$, and similarly for the other colors. The calculation determines $Pr[\text{red}|\text{black dot}]$, which in a sense is the reverse of the specification. A similar calculation determines the likelihood of a disease (e.g., flu) given a symptom (e.g., fever). Here, the symptom is the back dot and the disease is the color of the marble.

Random Variable

Say that you get \$8.00 if you pick a blue marble, \$5.00 if it is red, \$2.00 if it is green, and \$2.00 if it is white. The amount you get is then a function of the color of the marble you picked. This function is fixed. Let us call the function $X(\cdot)$. Thus, $X(\text{blue}) = 10$ and $X(\text{white}) = 2$, and so on. We call X a *random variable*. Thus, we say that *a random variable is a real-valued function of the outcome of a random experiment*. Here, the random experiment is choosing a marble. The outcome is the color of the marble. We have specified all the possible outcomes: blue, red, green, white. Also, we know the probability of each outcome. For instance, $Pr[\text{blue}] = 0.1$. The set of outcomes and their probability specifies the random experiment. The function X assigns a real number to each outcome. Note that the values assigned to different outcomes do not have to be different. Here, $X(\text{green}) = X(\text{white}) = 2$.

Distribution

Assume that we are interested only in how much you get, not in the details of the experiment that produces that gain. In that case, we can describe X by saying that $X = 8$ with probability 0.1 (which is the probability you pick a blue marble), $X = 5$ with probability 0.2, and $X = 2$ with probability 0.7 (the probability that you pick a green or white marble). Thus, the possible values of X are $8, 5, 2$ and their probability is $0.1, 0.2, 0.7$, respectively. These values and their probability are called the *distribution* of the random variable X.

Expectation

Imagine that you repeat the experiment (shake, pick, collect X) a very large number N of times. The frequentist interpretation suggests that the fraction of the times that you collect 8 is 0.1, that you collect 5 is 0.2 and that you collect 2 is 0.7. Thus, you collect 8 about $0.1N$ times, 5 about $0.2N$ times, and 2 about $0.7N$ times. Hence, the total amount you collect over the N experiments is about $8 \times 0.1N + 5 \times 0.2N + 2 \times 0.7N = (8 \times 0.1 + 5 \times 0.2 + 2 \times 0.7)N$. Accordingly, the average amount you collect *per experiment* is $8 \times 0.1 + 5 \times 0.2 + 2 \times 0.7$. We call this value the *expectation* of X and we write it as $E[X]$. We also call $E[X]$ the *mean value* or the *expected value* of X. Thus,

$$E[X] = 8 \times 0.1 + 5 \times 0.2 + 2 \times 0.7 = 3.2.$$

That is, $E[X]$ is the sum of the values of X multiplied by their probability.

Function

Would you rather play the game (pick a marble and get X) or get \$3.20 without playing the game? The answer depends on a key factor that the economists call the *utility* that you have for money. To make the situation a bit more dramatic, say that you can either get \$1.00 or play a game and get \$100.00 with probability 0.01 or \$0.00 otherwise. What do you prefer? Many people tend to choose to play the game. In fact, may people play the California Lottery where the odds of winning \$100M are much less that 10^{-8}.

Let $h(x)$ be the utility that you have for $\$x$. Say that (this is a silly example, but it will illustrate a point) $h(8) = 10$ and $h(5) = h(3.2) = h(2) = 0$. For instance, for \$8.00, you can by a ticket to go see the latest Pokemon movie you crave and that you cannot do anything of comparable value with less than \$8.00. Then we find that, after playing the marble game, $h(X) = 10$ with probability 0.1 and $h(X) = 0$ with probability 0.9. Hence, $E[h(X)] = 10 \times 0.1 + 0 \times 0.9 = 1$. On the other hand, if you don't play the game and get 3.2, then $h(3.2) = 0$. Thus, you would rather play the game. Similarly, people play the lottery beause winning would change their life, presumably for the better, whereas loosing \$1.00 does not affect their life.

Note that we calculated $E[h(X)]$ by finding the distribution (recall that this means the set of possible values and their probability) of $h(X)$. We could have calculated $E[h(X)]$ directly from the distribution of X:

$$E[h(X)] = h(8)Pr[X = 8] + h(5)Pr[X = 5] + h(2)Pr[X = 2] = 10 \times 0.1 + 0 \times 0.2 + 0 \times 0.7.$$

This is simple observation, but it is convenient.

In a similar way, we could have computed $E[h(X)]$ by looking at the outcomes of the marble picking game:

$$\begin{aligned} E[h(X)] &= h(X(\text{blue}))Pr[\text{blue}] + h(X(\text{red}))Pr[\text{red}] + h(X(\text{green}))Pr[\text{green}] + h(X(\text{white}))Pr[\text{white}] \\ &= h(8)0.1 + h(5)0.2 + h(2)0.3 + h(2)0.5. \end{aligned}$$

Indeed, these three different ways of calculating $E[h(X)]$ correspond to different ways of summing the possible ways of getting the values of $h(X)$: summing over the values of $h(X)$, or the values of X, or the outcomes.

Variance

We saw that one can describe a random variable X by its distribution. A summary of that distribution is the mean value $E[X]$. However, our discussion of the utility shows that this description is a bit crude and may not suffice to decide whether to play a game of chance. For instance, the expected gain of playing the lottery is negative. You would not paly a game where you are certain to loose.

The mean value does not say anything about the uncertainty of X, i.e., its variability. Here, by variability we mean that if we play the game many times, we observe a variety of values of X. The *variance* is a one-number summary of variability. The variance of X is defined by

$$var[X] = E[(X - E[X])^2].$$

The intuition is that if X is almost always close to $E[X]$, then the variance is small; otherwise, it is large.

In our marble example, $E[X] = 3.2$. Since $X = 8, 5$, or 2 with probability $0.1, 0.2, 0.7$, respectively, we see that

$$\begin{aligned} var[X] &= E[(X - E[X])^2] = (8 - 3.2)^2 Pr[X = 8] + (5 - 3.2)^2 Pr[X = 5] + (2 - 3.2)^2 Pr[X = 2] \\ &= (8 - 3.2)^2 0.1 + (5 - 3.2)^2 0.2 + (2 - 3.2)^2 0.7 = 23.04 \times 0.1 + 3.24 \times 0.2 + 1.44 \times 0.7 = 3.96. \end{aligned}$$

The square root of the variance is called the *standard deviation* and we denote it by σ_X. Here, $\sigma_X = \sqrt{3.96} \approx 2$.

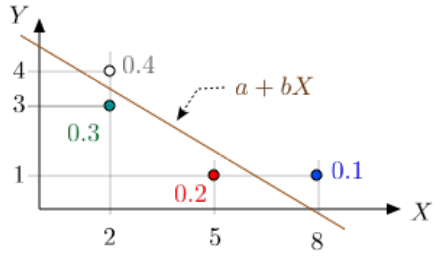

Figure 1: Linear Regression of Y over X (brown)

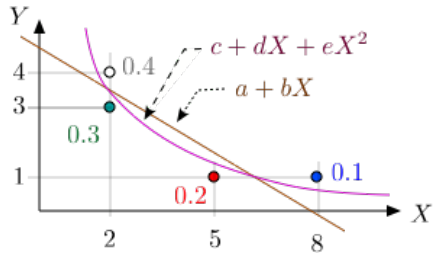

Figure 2: Quadratic Regression of Y over X (purple)

Linear Regression

Consider once again our bag of marbles. Define another random variable Y by $Y(\text{blue}) = 1, Y(\text{red}) = 1, Y(\text{green}) = 3$ and $Y(\text{white}) = 4$. Thus, each outcome (i.e., color) is assigned two numbers: X and Y. In another context, each person is associated with a height and a weight. Say that you want to guess the weight of a person from his/her height. How do you do it? Here, we want to guess Y from the value of X.

Here, a picture helps. Figure 1 shows the values of X and Y associated with the four possible outcomes. For instance, the blue outcome is associated with $X(\text{blue}) = 8$ and $Y(\text{blue}) = 1$. The figure also shows the probability of the different outcomes. We want a simple formula to provide a guess of Y based on X. In fact, we want a formula of the form $\hat{Y} = a + bX$. Here, $\hat{Y}$ is our guess for Y based on the value of X. Also, a and b are some constants. This formula corresponds to the line shown in the figure. We choose a and b so that the guess $\hat{Y}$ tends to be close to Y. This means that the line should be close to the actual points (X,Y) in the figure. Thus, $\hat{Y} - Y$ should be small. We make this precise by requiring that $E[(\hat{Y} - Y)^2]$ be as small as possible. That is, we choose a and b to minimize

$$E[(\hat{Y} - Y)^2] = E[(a + bX - Y)^2].$$

We explain in the lectures that the best choice of a and b is such that

$$\hat{Y} = E[Y] + \frac{cov(X,Y)}{var[X]}(X - E[X])$$

where $cov(X,Y) = E[XY] - E[X]E[Y]$.

Quadratic Regression

In the previous section, we estimated Y by using a linear function $a + bX$ of X, as shown in Figure 1. Figure 2 suggests that a quadratic estimate $c + dX + eX^2$ is better than a linear estimate, i.e., that it is closer to the pairs (X,Y). In the lectures, we explain how to find the best values of c, d, e.

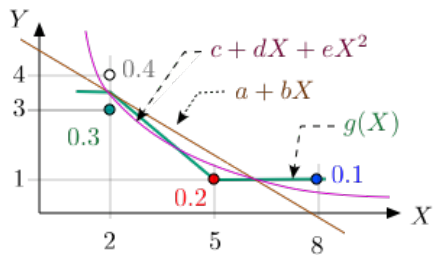

Figure 3: Conditional Expectation of Y given X (green)

Conditional Expectation

What if we could choose any function of X instead of being limited to linear or quadratic functions? Figure 3 shows the best possible function $g(X)$ of X to estimate Y. We explain in the lectures how to calculate that function called the *conditional expectation* of Y given X.

2 Flip Coins

So far, we looked at one or two random variables. In this section, we explore many random variables.

Setup

You have a coin. When you flip it, there are two possible outcomes: 'heads' (H) and 'tails' (T). Let $p = Pr[H]$, so that $Pr[T] = 1 - p$. For instance, the coin could be biased with $p = 0.6$, so that heads is more likely than tails.

Independence

Say that you flip the coin twice. There are four possible outcomes for this experiment: HH, HT, TH, and TT. Here, HT means that the first flip produces H and the second T, and similarly for the other outcomes. If we recall the definition of conditional probability, we have

$$\begin{aligned} &Pr[\text{first fip yields } H \mid \text{second flip yields } H] \\ &= \frac{Pr[(\text{first flip yields } H) \text{ and } (\text{second flip yields } H)]}{Pr[\text{second flip yields } H]} \\ &= \frac{Pr[HH]}{p}. \end{aligned}$$

In the last step, we used the fact that the probability that the second flip yields H is p.

Now, it is reasonable to assume that the likelihood that the first flip yields H does not depend on the fact that the second flip yields H and that this likelihood is then p. Hence, we are led to the conclusion that $p = Pr[HH]/p$, so that $Pr[HH] = p^2$. This assumption is called the *independence* of the coin flips. A similar reasoning yield to the conclusion that

$$Pr[HT] = p(1-p), Pr[TH] = (1-p)p, Pr[TT] = (1-p)^2.$$

Let $X = 1$ when the first flip is H and $X = 0$ when it is T. Also, let $Y = 1$ when the second flip is H and $Y = 0$ when it is T. Then we see that $Pr[X = 1] = Pr[Y = 1] = p$ and $Pr[X = 1, Y = 1] = Pr[X = 1]Pr[Y = 1]$. Also, $Pr[X = 1, Y = 0] = Pr[X = 1]Pr[Y = 0]$. More generally, $Pr[X = a, Y = b] = Pr[X = a]Pr[Y = b]$ for all a, b. Two random variables with that property are said to be *independent.*

Variance of Sum

Let X and Y be independent random variables. We show in the lectures that $var[X+Y] = var[X] + var[Y]$. More generally, if $X_1, \ldots, X_n$ are random variables such that any two of them are independent, then $var[X_1 + \cdots + X_n] = var[X_1] + \cdots + var[X_n]$. Moreover, we will see that $var[aX] = a^2 var[X]$ for any random variable X and any constant a. Consequently, we see that

$$var[\frac{X_1 + \cdots + X_n}{n}] = \frac{var[X_1] + \cdots + var[X_n]}{n^2}.$$

In particular, if $var[X_m] = \sigma^2$ for $m = 1, \ldots, n$, we have

$$var[\frac{X_1 + \cdots + X_n}{n}] = \frac{var[X_1] + \cdots + var[X_n]}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Chebyshev's Inequality

Flip a coin n times and let $X_m = 1$ if flip m yields H and $X_m = 0$ otherwise. Then

$$\begin{aligned} var[X_m] &= E[(X_m - E[X_m])^2] = E[(X_m - p)^2] = (1-p)^2 Pr[X_m = 1] + (0-p)^2 Pr[X_m = 0] \\ &= (1-p)^2 p + p^2(1-p) = p(1-p). \end{aligned}$$

Accordingly, in view of the previous section,

$$var[\frac{X_1 + \cdots + X_n}{n}] = \frac{p(1-p)}{n}.$$

Thus, when n is large, the variance of $A_n := (X_1 + \cdots + X_n)/n$ is very small. This suggests that the random variable A_n tends to be very close to its mean value, which happens to be p. Thus, we expect the fraction of heads A_n in n coin flips to be close to p.

To make this idea precise, Chebyshev developed an inequality which says that

$$Pr[|X - E[X]|^2 > \varepsilon] \leq \frac{var[X]}{\varepsilon^2}.$$

We prove this inequality in the lectures.

Thus, the likelihood that a random variable X differs from its mean $E[X]$ by at least ε is small if $var[X]$ is small. If we apply this inequality to A_n, we find that

$$Pr[|A_n - p| \geq \varepsilon] \leq \frac{p(1-p)}{n\varepsilon^2}.$$

Note that $p(1-p) \leq 1/4$ for any value of p. Consequently, we see that

$$Pr[|A_n - p| \geq \varepsilon] \leq \frac{1}{4n\varepsilon^2}.$$

Confidence Interval

Say that you do not know the value of $p = Pr[H]$. To estimate it, you flip the coin n times and note the fraction A_n of heads. The last inequality holds. Let us choose ε so that the right-hand side of the inequality

is $0.05 = 1/20$. That is, we choose ε so that $4n\varepsilon^2 = 20$, i.e., $\varepsilon^2 = 5/n$ or $\varepsilon = \sqrt{5}/\sqrt{n} \approx 2.25/\sqrt{n}$. Hence, the previous inequality with that value of ε implies that

$$Pr[|A_n - p| \geq \frac{2.25}{\sqrt{n}}] \leq 0.05,$$

so that

$$Pr[|A_n - p| \leq \frac{2.25}{\sqrt{n}}] \geq 1 - 0.05 = 95\%.$$

Now, since $|A_n - p| \leq \delta$ if and only if $p \in [A_n - \delta, A_n + \delta]$, we conclude that

$$Pr[p \in [A_n - \frac{2.25}{\sqrt{n}}, A_n + \frac{2.25}{\sqrt{n}}]] \geq 95\%.$$

For instance, say that $n = 10^4$ and $A_n = 0.31$. We then conclude that

$$Pr[p \in [0.31 - \frac{2.25}{100}, 0.31 + \frac{2.25}{100}]] \geq 95\%,$$

so that

$$Pr[p \in [0.2875, 0.3325]] \geq 95\%.$$

We say that $[0.2875, 0.3325]$ is a 95%-*confidence interval* for p. As you can see, the width of the confidence interval decreases like $1/\sqrt{n}$.

This example is the basis for the estimates in public opinion surveys.

Time until first H

We flip the coin until we get the first H. How many times do we need to flip the coin, on average? Let β be that average number of flips. That number of flips is 1 if the first flip is H, which occurs with probability p. If the first coin is T, which occurs with probability $1 - p$, then the process starts afresh and we need to flip the coin β more times, on average. Thus, $\beta = p \times 1 + (1 - p) \times (1 + \beta)$. Solving, we find $\beta = 1/p$.

Time until two consecutive Hs

We flip the coin until we get two consecutive Hs. How many times do we need to flip the coin, on average? Let β be that average number of flips. Let also $\beta(H)$ be the average number of additional flips until two consecutive Hs, given that the last flip is H. Then we claim that

$$\begin{aligned}\beta &= p(1 + \beta(H)) + (1 - p)(1 + \beta)\\ \beta(H) &= p \times 1 + (1 - p)(1 + \beta).\end{aligned}$$

The first identity can be seen by noting that if the first flip is H, then after that first flip one needs $\beta(H)$ additional flips, on average, since the last flip was H. However, if the second flip is T, then after the first flip one needs β additional flips, on average. The second identity can be justified similarly. Solving, one finds $\beta = 1/p + 1/p^2$.