L1 Probability Review

Greg Ridgeway

2025-03-30

Table of contents

1 Sample space and events

Sample space is the set of all possible outcomes from a random process.

- $S = \{\text{dropout}, \text{graduate}\}$, simplified high school outcome
- $S = \{0, 1, 2, 3, ...\}$, number of workplace injuries
- $S = [0, 130]$, age… or collision speed
- $S = \{\text{nothing}, \text{restrain}, \text{strike}, \text{baton/taser/OC spray}, \text{firearm}\}$, ordered use-of-force

An **event** is a subset of the sample space

- $A = \{\text{graduate}\}$
- $A = \{0, 1, 2\}$
- $A = [16, 25]$
- $A = \{\text{nothing}, \text{restrain}\}$

2 What is probability?

Frequentist definition: The probability of an **event** is the **fraction** of times it is expected to happen when the basic **process** is done **over and over** under the **same conditions**.

Subjectivist definition: Probability is an individual's **belief** expressed as a measure of uncertainty.

What is the probability that the Phillies win next year's World Series? The frequentist would have to think about next year's World Series as being part of a larger sequence of games, perhaps depending on a complex set of factors (e.g. weather, injuries, opponents). Subjectivists are comfortable simply expressing their belief as a probability.

Example: Let X be the number of cities in the world with more than 1 million people that start with the letter M. What is the probability that $X \geq 20$? Frequentists will say that this has to be 0 or 1 since this is not a repeatable experiment. However, after reading this question you probably are already formulating an estimate of how sure you are that there are more than 20 such cities. You might be so certain that you would be willing to bet someone real money on whether $X \geq 20$ or not. de Finetti in 1937 defined subjective probability as essentially the price that you are willing to pay for a lottery ticket that yields 1 unit of money if the event occurs and nothing otherwise.

2.1 Axioms of probability

Let A be an event in a sample space S.

1. $P(A) \geq 0$, probability cannot be negative
2. $P(S) = 1$, some event in the sample space will happen
3. If $A_1 \cap A_2 = \emptyset$, then $P(A_1 \cup A_2) = P(A_1) + P(A_2)$, probabilities for mutual exclusive events are additive

Note: $\cup$ is like "or", $\cap$ is like "and"

2.2 Additional useful properties

1. $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$
2. $P(\bar{A}) = 1 - P(A)$, where $\bar{A}$ means "not A"
3. If A and B are independent, then $P(A \cap B) = P(A)P(B)$

Two events are **independent** if the chances for the second event are the same regardless of how the first event turns out. Otherwise, the two events are **dependent**.

3 Conditional probability

Conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

The conditional probability $P(A|B)$ is the probability that the event A occurs given that B occurs. Often we will be interested in computing conditional probabilities like

$$P(Y = 1|x_1, x_2, \ldots, x_d)$$

the probability that $Y = 1$ given a set of case features, $\mathbf{x}$.

Sometimes, for example when we study the naïve Bayes classifier, we will be interested in $P(x_j|Y = 1)$. Pay attention to which events are on which side of the vertical bar.

It is common (possibly intentional at times) to flip conditional probabilities when making arguments. The "prosecutor's fallacy" is to give $P(\text{evidence}|\text{innocence})$ but interpret it as $P(\text{innocence}|\text{evidence})$. In Philadelphia the District Attorney correctly claimed that of those arrested for illegal gun possession, very few of them go on to shoot someone ($P(\text{shoot}|\text{VUFA arrest})$ is small), but the police department argued that almost all the

shooters have been previously arrested for illegal gun possession ($P(\text{VUFA arrest}|\text{shooter})$ is large). Both are correct statements, but essentially they are arguing about which probability is more important for public safety.

Famously, Alan Dershowitz defended OJ Simpson claiming that one in 2500 men who beat their intimate partners eventually kill them. We want to know

$$P(\text{murdered by batterer}|\text{murdered})$$

but Dershowitz gave

$$P(\text{murdered by batterer}|\text{battered}) = \frac{1}{2500}$$

The useful probability, $P(\text{murdered by batterer}|\text{murdered})$, turns out to be 0.29.

3.1 Exercises

Table 1: Probability of cheating on taxes and mom's birthday month

	Mom's birthday	
	Jan-Mar	Apr-Dec
tax cheat	0.025	0.075
not tax cheat	0.225	0.675

$S = \{(\text{Jan-Mar}, \text{cheat}), (\text{Jan-Mar}, \text{no cheat}), (\text{Apr-Dec}, \text{cheat}), (\text{Apr-Dec}, \text{no cheat})\}$.

Let $A =$ tax cheat and $B =$ Jan-Mar birthday. Compute the following

1. $P(S)$
2. $P(A)$
3. $P(\bar{A})$
4. $P(B)$
5. $\frac{P(A)}{1-P(A)}$ (odds of A)
6. $P(A|B)$
7. $P(A|\bar{B})$
8. $\frac{P(A|B)}{P(A|\bar{B})}$ (risk ratio)
9. $\frac{\frac{P(A|B)}{1-P(A|B)}}{\frac{P(A|\bar{B})}{1-P(A|\bar{B})}}$ (odds ratio)
10. $\log \frac{\frac{P(A|B)}{1-P(A|B)}}{\frac{P(A|\bar{B})}{1-P(A|\bar{B})}}$ (log odds ratio)

3.2 Randomized response design

Assume we do not know the probability that a randomly selected person cheats on their taxes. Let $p = P(\text{tax cheat})$.

Table 2: Probability of cheating on taxes and mom's birthday month

	Mom's birthday	
	Jan-Mar	Apr-Dec
tax cheat	$0.25p$	$0.75p$
not tax cheat	$0.25(1-p)$	$0.75(1-p)$

If we ask a large number of people "Say 'yes' if you cheat on your taxes *or* your mom's birthday in January, February, or March," what fraction would say "yes"?

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= (0.25p + 0.75p) + (0.25p + 0.25(1-p)) - 0.25p \\ &= p + 0.25p + 0.25 - 0.25p - 0.25p \\ &= 0.25 + 0.75p \end{aligned}$$

If you surveyed 10,000 people with this question and 34% said "yes," what would be your best guess for p, the probability that someone cheats on their taxes?

4 Bayes' Theorem

Bayes' Theorem

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{2}$$

Note that in order to compute $P(A)$ we often have to sum over all the options for B, for example

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap \bar{B}) \\ &= P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) \end{aligned}$$

4.1 Example 1. ELISA test

The ELISA test was introduced in the mid-1980s to screen blood for the AIDS virus. It detects antibodies, substances that the body produces when the virus is present.

- When antibodies are present, ELISA is positive with probability 0.98 and negative with probability 0.02
- When blood is not contaminated with antibodies, ELISA gives a positive result with probability about 0.07 and a negative result with probability 0.93
- Suppose that 1% of a large population carry the AIDS antibody in their blood

What is the probability that someone who tests positive actually has antibodies?

$$
\begin{aligned}
&P(\text{AB}|\text{ELISA+}) \\
&= \frac{P(\text{ELISA+}|\text{AB})P(\text{AB})}{P(\text{ELISA+})} \\
&= \frac{0.98 \times 0.01}{P(\text{ELISA+}|\text{AB})P(\text{AB}) + P(\text{ELISA+}|\text{no AB})P(\text{no AB})} \\
&= \frac{0.98 \times 0.01}{0.98 \times 0.01 + 0.07 \times 0.99} \\
&= 0.124
\end{aligned}
$$

4.2 Exercise 1: Are older parolees risky?

A report noted that in 2002, 0.8% of those rearrested within one year of release on parole were 50 years or older. Advocates of sentencing reform suggested that this statistic indicates that releasing prisoners age 50 and older could alleviate prison overcrowding without risking harm to the public.

50% of parolees released in 2002 were rearrested within one year.

1% of those released on parole in 2002 were 50 or older

Compute the probability of rearrest within one year for a parolee aged 50 or older who was released in 2002. What do you think of the proposed prisoner release program given the 0.8% statistic?

4.3 Exercise 2: How did Uncle Basil die?

Bart is trying to kill Uncle Basil for money. Knowing Uncle Basil likes dessert, Bart puts

- rat poison in the cherries flambé (fatal 60% of the time)
- cyanide in the chocolate mousse (fatal 90% of the time)

Uncle Basil eats cherries 50% of the time, chocolate mousse 40% of the time, and something else 10% of the time.

If Uncle Basil dies, what's the probability that it was the mousse that killed him?

4.4 Exercise 3: Beyond a reasonable doubt?

Police recover a knife at the murder scene on the Orient Express. There are 13 passengers on the train suspected of the murder and police believe that each of these 13 have the same chance of being the murderer.

A bloody fingerprint lifted from the knife used as the murder weapon best matches Mrs. Hubbard. However, forensic examiners note that the match is not perfect and calculate that there is a 0.008 probability of a fingerprint match even if Mrs. Hubbard was innocent.

What is the probability that Mrs. Hubbard is the murderer?

In 1769, William Blackstone said "It is better that ten guilty escape than one innocent suffer." Blackstone's Ratio convicts those with probability of guilt greater than or equal to $10/(10+1) = 0.909$. Would you convict Mrs. Hubbard based on the evidence?

Ben Franklin in 1785 said, "It is better 100 guilty Persons should escape than that one innocent Person should suffer," equivalent to convicting when the probability of guilt exceeds $100/(100+1)$. Using Franklin's Ratio, would you still convict Mrs. Hubbard?

4.5 Exercise 4: Vampires, ghosts, and lies

In a small village in Transylvania 15% of the population are vampires, 20% are ghosts, and 65% are ordinary people.

Vampires never tell the truth, ghosts tell the truth 37% of the time, and ordinary people tell the truth 95% of the time.

It is impossible to tell apart vampires, ghosts and ordinary people by the way they look (between 6:00am and 11:59:59pm).

You are introduced to a gentleman from the village. What is the probability that he did not give you his real name?

Just at 11:59:59pm you realize that the gentleman lied about his name. What is the probability that your companion is a vampire or ghost?

5 Prediction

We are almost always interested in the probability of some outcome Y conditional on observing some features $X_1, X_2, \ldots, X_d$, written as the conditional probability $P(Y|\mathbf{x})$.

- $Y \in 0, 1$, **binary classification**
- Y is a discrete variable with more than two possible values, **multiclass classification**
- Y is a continuous variable, **regression**

We will focus on the binary classification problem for now.

What if you need to make a classification decision based on $P(Y|\mathbf{x})$? Perhaps predict that $Y = 1$ whenever $P(Y = 1|\mathbf{x}) > 0.5$. But what if, the cost of a false negative (failing to identify a case where $Y = 1$) is different from the cost of a false positive (incorrectly predicting $Y = 1$).

Let's say a false positive costs c units while a false negative costs 1 units.

What is our expected cost if we label the case as a 0?

$$0 \times P(Y = 0|\mathbf{x}) + 1 \times P(Y = 1|\mathbf{x}) = P(Y = 1|\mathbf{x})$$

What is our expected cost if we label the case as a 1?

$$c \times P(Y = 0|\mathbf{x}) + 0 \times P(Y = 1|\mathbf{x}) = cP(Y = 0|\mathbf{x})$$

We should classify as a 1 if

$$\begin{aligned} cP(Y = 0|\mathbf{x}) &< P(Y = 1|\mathbf{x}) \\ c(1 - P(Y = 1|\mathbf{x})) &< P(Y = 1|\mathbf{x}) \\ c - cP(Y = 1|\mathbf{x}) &< P(Y = 1|\mathbf{x}) \\ c &< P(Y = 1|\mathbf{x}) + cP(Y = 1|\mathbf{x}) \\ c &< (c + 1)P(Y = 1|\mathbf{x}) \\ P(Y = 1|\mathbf{x}) &> \frac{c}{c + 1} \end{aligned}$$

If false positives cost nothing, then $c = 0$ and we would classify everyone as a 1. If false positives cost 1 (equal to false negatives), then classify as a 1 if $P(Y = 1|\mathbf{x}) > \frac{1}{2}$. Whenever you see examples in which the decision boundary is $\frac{1}{2}$, the analysts are assuming equal false positive and false negative costs. If false positives are infinitely expensive, then $c \to \infty$ and we classify as a 1 only when $P(Y = 1|\mathbf{x}) = 1$ and all other classifications should be $Y = 0$.

6 Random variables, expected value, variance

A random variable is a numerical summary of the outcome of a random event.

- $X = 1$ if the student graduates, $X = 0$ if the student does not graduate
- T is the time until a parolee reoffends
- Y is total earnings within 5 years of college graduation

We often summarize a random variable by its average or expected value. A random variable is *discrete* if it takes on a finite or countable number of values, like 0/1 outcomes, the number of times something happens, or the count of a number of items.

Expected value of discrete random variables

$$\mathbb{E}(Y) = \sum_{y=0}^{\infty} yP(Y = y) \tag{3}$$

Equation (3) says that the expected value is the sum over all the possible values Y can have times the probability that that value of Y occurs.

6.1 Examples:

- Starting with the classic coin flip, if Y counts the number of tails that occurs in a single coin flip, then we have $\mathbb{E}(Y) = 0 \times 0.5 + 1 \times 0.5 = 0.5$. Note that the expected value does not necessarily have to equal one of the possible values of Y
- In the US, of mothers over age 40, assume that the distribution of the number of children is

# children (y)	0	1	2	3	4
$P(Y = y)$	0.18	0.19	0.32	0.20	0.11

The expected value of Y is

$$\mathbb{E}(Y) = 0 \times 0.18 + 1 \times 0.19 + 2 \times 0.32 + 3 \times 0.20 + 4 \times 0.11 = 1.87$$

For a physical understanding of the expected value, imagine a (weightless) rod of length 4 cm. Place an 18g weight at the left end, a 19g weight at the 1cm mark, a 32g weight at the 2cm mark, a 20g weight at the 3cm mark, and an 11g weight at the 4cm mark (the right end of the rod). The expected value is the point at which this rod would balance.

For continuous random variables, we replace the sum with an integral.

Expected value of continuous random variables

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} yp(y)\,dy \tag{4}$$

(4) replaces the sum with an integral and the probability function $P(Y = y)$ with a probability density function, $p(y)$. Physically, $\mathbb{E}(Y)$ is the place where the shape $p(y)$ would balance.

Example:

- Let T be the time to reoffense with an exponential distribution of the form

  $$p(t) = \lambda e^{-\lambda t},\ t \geq 0$$

  The expected value is

  $$\mathbb{E}(T) = \int_{0}^{\infty} t\lambda e^{-\lambda t}\,dt = \frac{1}{\lambda}$$

For this class, feel free to use Wolfram Alpha or Integral Calculator or similar sites to help you solve any problems in this class.

6.2 Properties of expected value

Since the expected value is a sum or an integral, it has all the same properties of sums and integrals.

- $\mathbb{E}(aY) = a\mathbb{E}(Y)$
- $\mathbb{E}(Y_1 + Y_2) = \mathbb{E}(Y_1) + \mathbb{E}(Y_2)$

Example:

- You have two randomly selected 40-year-old mothers. What is the expected total number of children that they have?

  $$\mathbb{E}(Y_1 + Y_2) = \mathbb{E}(Y_1) + \mathbb{E}(Y_2) = 1.87 + 1.87 = 3.74$$

6.3 Monte Carlo estimation of integrals

Some (most?) integrals we will encounter in machine learning are too complicated to compute in a convenient closed-form solution. Fortunately, the Law of Large Numbers tells us that we can use sample averages to approximate integrals.

Law of Large Numbers

If $y_1, \ldots, y_n$ is a random sample from $p(y)$, then as $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n} g(y_i) \to \int_{-\infty}^{\infty} g(y)p(y)\,dy = \mathbb{E}[g(Y)]. \tag{5}$$

Example:

- Returning to the example with the number of children, we can write an R script to estimate the expected value.

```
nChildren <- sample(0:4, 10000,
                    prob=c(0.18,0.19,0.32,0.20,0.11),
                    replace=TRUE)
mean(nChildren)
```

```
[1] 1.8577
```

6.4 Variance

The variance of a random variable describes its variability, specifically how far away from its mean it is on average as measured by squared distance.

Variance

$$\text{Var}(Y) = \mathbb{E}[(Y - \mathbb{E}(Y))^2] \tag{6}$$

In the coming weeks we will be looking at terms like the following where D represents an entire dataset, Y is an outcome, and $\mathbf{x}_0$ is a set of specific case features. $f(\mathbf{x})$ is the best possible predictive model equal to $\mathbb{E}_D(Y|\mathbf{x}_0)$ and $\hat{f}(\mathbf{x}_0)$ is our machine learning method attempting to best predict y.

$$\mathbb{E}_{D,Y|\mathbf{X}=\mathbf{x}_0}(Y - f(\mathbf{x}_0))^2 \tag{7}$$

Equation (7) represents *noise*. It is the variance of Y and is the part of the outcome that is simply random noise, which no possible machine learning method could hope to capture.

$$\mathbb{E}_{D,Y|\mathbf{X}=\mathbf{x}_0}\left[\left(f(\mathbf{x}_0) - \mathbb{E}_D\hat{f}(\mathbf{x}_0|D)\right)^2\right] \tag{8}$$

Equation (8) represents *bias*, on average how far away our model, $\hat{f}(\mathbf{x}|D)$, is from the best model, $f(\mathbf{x})$.

$$\mathbb{E}_{D,Y|\mathbf{X}=\mathbf{x}_0}(\hat{f}(\mathbf{x}_0|D) - \mathbb{E}_D\hat{f}(\mathbf{x}_0|D))^2 \tag{9}$$

Lastly, (9) represents *variance*, how much our predictions change from dataset to dataset. It is the variability of our machine learning approach and how sensitive it is to changes in the dataset that we use to train the machine learning method.

7 Final notes

Twenty-four cities start with the letter M and have more than 1 million people: Moscow (Russia), Mumbai (India), Mexico City (Mexico), Madrid (Spain), Manila (Philippines), Melbourne (Australia), Mashhad (Iran), Medellin (Columbia), Mogadishu (Somalia), Managua (Nicaragua), Manaus (Brazil), Multan (Pakistan), Medan (Indonesia), Maracaibo (Venezuela), Minsk (Belarus), Maiduguri (Nigeria), Montreal (Canada), Mecca (Saudi Arabia), Makassar (Indonesia), Munich (Germany), Milan (Italy), Montevideo (Uruguay), Maputo (Mozambique), and Monterrey (Mexico)