Chapter 1

Bayes' rule and its applications

1.1 The chain rule

This chapter is going to focus on how to re-write joint and conditional probabilities. When we turn to statistics later on, it will turn out that it is often hard to define a joint distribution over many variables. Likewise, it can be hard to calculate the probability distribution of a RV X conditioned on a RV Y but it may be much easier to find the distribution of Y conditioned on X. In this chapter we are essentially trying to find simpler expressions for distributions that may be hard to compute.

The first general method for simplifying a joint distribution is know as the **chain rule**. We could actually have introduced the chain rule much earlier, when talking about the probability of events. However, we chose not to do so as it would have cluttered the presentation. Moreover, we only need the chain rule now. For completeness' sake we are first going to formulate the chain rule for events and then for random variables.

Theorem 1 ***(Chain rule)*** *The joint probability of events $E_1, \ldots, E_n$ can be factorised as*

$$\mathbb{P}(E_1, \ldots, E_n) = \mathbb{P}(E_1) \times \mathbb{P}(E_2|E_1) \times \ldots \times \mathbb{P}(E_n|E_1, \ldots, E_{n-1})$$

There are a couple of things to note here. First of all, the numbering of the events is arbitrary. That means that it does not matter in which order we decompose the joint probability. We could just as well start with any E_i for $1 \leq i \leq n$. Second we used the word *factorise*. This simply means that we decompose any expression (in this case a joint probability) into a product. Products are nice in that we can arrange them in any order that we like (i.e. they commute). Moreover, products make a lot of calculations easier, as we will see later.

But now let us go ahead and actually prove the chain rule. We are going to do so inductively and chose $\mathbb{P}(E_1, E_2)$ as our base case. Then we simply employ the definition of conditional probability to get

$$\mathbb{P}(E_1, E_2) = \mathbb{P}(E_1) \times \frac{\mathbb{P}(E_1, E_2)}{\mathbb{P}(E_1)} = \mathbb{P}(E_1) \times \mathbb{P}(E_2|E_1) \tag{1.1}$$

Now let us assume that the chain rule holds for events $E_1, \ldots, E_{n-1}$. We will abbreviate them as E_1^{n-1}. Then we get

$$\mathbb{P}(E_1^{n-1}, E_n) = \mathbb{P}(E_1^{n-1}) \times \frac{\mathbb{P}(E_1^{n-1}, E_n)}{\mathbb{P}(E_1^{n-1})} = \mathbb{P}(E_1^{n-1}) \times \mathbb{P}(E_n|E_1^{n-1}) \tag{1.2}$$

Since $\mathbb{P}(E_1^{n-1})$ factorises according to the chain rule by our induction hypothesis, we have completed the proof.

The chain rule can make our lives even simpler if we have independent events. Assume we have to compute the joint probability of 3 events E_1, E_2, E_3 and we also know that $E_1 \perp E_2$. In this case our factorisation becomes 1.3 where the first equality follows from the chain rule and the second equality follows from independence between E_1 and E_2.

$$\begin{aligned} \mathbb{P}(E_1, E_2, E_3) &= \mathbb{P}(E_1) \times \mathbb{P}(E_2|E_1) \times \mathbb{P}(E_3|E_1, E_2) \\ &= \mathbb{P}(E_1) \times \mathbb{P}(E_2) \times \mathbb{P}(E_3|E_1, E_2) \end{aligned} \tag{1.3}$$

We are now in a position to state the chain rule for random variables. There are two ways you can go about proving it. Either you calculate the probability of a specific setting of the variables or you just do the proof based on the distributions of the RVs. So in the first case you'd have to prove that

$$\begin{aligned} &P_{X_1^{n-1}}(X_1 = x_1, \ldots, X_n = x_n) \\ &= P_{X_1}(X_1 = x_1) \times \ldots \times P_{X_n|X_1^{n-1}}(X_n = x_n|X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) \end{aligned}$$

whereas in the second case you'd simply prove that

$$P_{X_1^n} = \sum_{i=1}^{n} P_{X_i|X_1^{i-1}}$$

Incidentally, we also introduce a very short notation for the chain rule above. Note that it is not quite correct, since if $i = 1$ we'd be conditioning on X_0. That's not to bad however, since we can always define ourselves a dummy variable X_0 that does not affect the distribution. Moreover, this notation is really just meant to be convenient, so you should just accept it as is when you encounter it in papers.

Exercise 1 *Prove the chain rule for random variables. The proof is totally analogous to the one give for events.*

1.2 Bayes' rule

In this section we are going to prove **Bayes' rule**. In order to prove it we will actually employ the chain rule. However, the proof is stupidly simple and thus of no great interest in and by itself. The consequences of Bayes' rule are huge however. It will basically allow us to invert a conditional probability distribution. Now you may rightfully aks: what's the deal? Well, as we said in the beginning, it may be hard to compute a conditional distribution in one direction but much easier to compute it in the other direction.

On top of that Bayes' rule opens up a whole range of new possibilities. We will discuss those as we proceed in this chapter. First of all let us state and prove Bayes's rule.

Theorem 2 ***(Bayes' rule)*** *The probability distribution of a random variable X given a random variable Y can be computed as*

$$P_{X|Y} = \frac{P_{Y|X}P_X}{P_Y}$$

And here comes the proof:

$$P_{X|Y} = \frac{P_{XY}}{P_Y} = \frac{P_{Y|X}P_X}{P_Y} \tag{1.4}$$

That was the proof! Considering how simple it was, it will be surprising to see what kind of benefits we can get out of Bayes' rule. To get us started let us introduce some terminology first. In particular, each of the terms in Bayes' rule has a specific name. You should really learn these names by heart as they crop up all over the place.

$$posterior = \frac{likelihood \times prior}{marginal\ likelihood}$$

The posterior is what we get after we have completed the computation. However, its name is related to the prior. The prior is just the probability that we would place on $P_X(X = x)$ *a priori*. Therefore P_X is also known as the prior distribution. Then when we multiply the prior by the likelihood we get a new distribution over X that is conditioned on Y. This is the distribution that we place on X *a posteriori*, i.e. after having taken into account information about X that we may get from knowing the value of Y. The marginal likelihood of Y is simply needed to normalize the expression to a probability distribution (i.e. to make sure that it sums to one). Why is it called marginal likelihood? The reason for this is how you can compute

it. Recall that when we are given a joint distribution $P_{X,Y}$ and X can take on n states, we can obtain the distribution P_Y by simply marginalizing over X.

$$P_Y(Y = y) = \sum_{i=1}^{n} P_{XY}(X = x_i, Y = y) \tag{1.5}$$

In addition to that, the chain rule allows us to factorise the joint probability. Thus we get

$$P_Y(Y = y) = \sum_{i=1}^{n} P_{Y|X}(Y = y|X = x_i) \times P_X(X = x_i) \tag{1.6}$$

Now if you think that this looks an awful lot like the enumerator of Bayes' rule then you are exactly on the right track. Essentially, we are just summing over all possible denominators (with respect to X). Let us make this more concrete with an example. Assume that we are given two coins. One of them is fair, meaning that it is equally probable to come up heads or tails. The other coin is biased towards tails and we happen to know that its probability to come up heads is only 0.3. Which coin is flipped is captured by a random variable X that takes on the value 0 if the fair coin is used and the value 1 if the biased coin is used. We have no idea which coin is going to be tossed, it could be either one. Therefore we set our prior as $P_X(X = 0) = P_X(X = 1) = 0.5$.

Now we flip the chosen coin 10 times and obtain 8 heads. The number of heads obtained during the 10 tosses is going to be encoded by Y. Since all tosses are independent each other, Y will follow a binomial distribution. For each of the two coins we also know the parameter of the binomial distribution. For the fair coin it is $\theta = 0.5$ and for the biased coin it is $\theta = 0.3$. Let us compute each of the enumerators separately.

$$P_{Y|X}(Y = 8|X = 0) \times P_X(X = 0) = 0.0439 \times 0.5 = 0.02195 \tag{1.7}$$

$$P_{Y|X}(Y = 8|X = 1) \times P_X(X = 1) = 0.0014 \times 0.5 = 0.0007 \tag{1.8}$$

We calculated the likelihood values with the help of a computer. You can do that yourself if you remember that Y $binom(10, \theta)$ and that $\theta = 0.5$ if $X = 0$ and $\theta = 0.3$ if $X = 1$.

All that is left do now is to compute the marginal likelihood of Y. Luckily for us, X only assumes two values, so we only need to add up 1.7 and 1.8.

$$\begin{aligned} P_Y(Y = 8) =& P_{Y|X}(Y = 8|X = 0) \times P_X(X = 0) + \\ & P_{Y|X}(Y = 8|X = 1) \times P_X(X = 1) = 0.02265 \end{aligned} \tag{1.9}$$

And finally we can apply Bayes' rule to compute the posterior probabilities of X.

$$P_{X|Y}(X=0|Y=8) = \frac{P_{Y|X}(Y=8|X=0) \times P_X(X=0)}{P_Y(Y=8)} \tag{1.10}$$

$$= \frac{0.2195}{0.02265} = 0.969$$

$$P_{X|Y}(X=1|Y=8) = \frac{P_{Y|X}(Y=8|X=1) \times P_X(X=1)}{P_Y(Y=8)} \tag{1.11}$$

$$= \frac{0.0005}{0.02265} = 0.031$$

(1.12)

What this means is there is a probability of 0.969 that the fair coin has been tossed when a sequence with eight heads is generated and only a probability of 0.031 that the biased coin was tossed. Obviously, the probability of the fair coin is much higher. But how much higher? To find this out we take the ratio of the two probabilities. This gives us $^{0.969}/_{0.031} \approx 31$. We can conclude that the fair coin is 31 times more likely to have generated the sequence with 8 heads than the biased coin. But wait a second, can we maybe find this ration somewhere else? It turns out that the ration of the likelihoods is the same! That is $^{0.0439}/_{0.0014} \approx 31$.

We started out by assuming that both coins were equally likely to be used. However, we then observed a sequence of 10 tosses, 8 of which were heads and that made it 31 times more likely that the fair coin was used. Now what if the priors had not been equal? There actually is a more general story to this. While calculating the actual probabilities involves a lot of number juggling, just telling whether or not an observation will make one or the other event more likely is not too hard.

$$\frac{P_{X|Y}(X=x_1|Y=y)}{P_{X|Y}(X=x_2|Y=y)} \tag{1.13}$$

$$= \frac{\frac{P_{Y|X}(Y=y|X=x_1)P_X(X=x_1)}{P_Y(Y=y)}}{\frac{P_{Y|X}(Y=y|X=x_2)P_X(X=x_2)}{P_Y(Y=y)}} \tag{1.14}$$

$$= \frac{P_{Y|X}(Y=y|X=x_1)P_X(X=x_1)}{P_{Y|X}(Y=y|X=x_2)P_X(X=x_2)} \tag{1.15}$$

In the above we clearly see that the ratio of the posterior probabilities is determined by the ration of the likelihood times the prior. In our coin example the priors were the same so it was only the likelihood that mattered.

If the ratio of any of the above terms is greater than 1, the posterior will change in favour of $X = x_1$. If the ratio is smaller than 1 the posterior changes in favour of $X = x_2$. If the ratio is exactly 1, nothing happens.

Notice that in general, although our observations may shift the posterior in favour of $X = x_2$, say, this does not necessarily imply that $P_{X|Y}(X = x_2|Y = y)$ will be greater than $P_{X|Y}(X = x_1|Y = y)$. In order for this to happen the following conditions must be fulfilled.

$$P_{X|Y}(X = x_1|Y = y) < P_{X|Y}(X = x_2|Y = y) \quad \Leftrightarrow \tag{1.16}$$

$$\frac{P_{Y|X}(Y = y|X = x_1)P_X(X = x_1)}{P_Y(Y = y)} < \frac{P_{Y|X}(Y = y|X = x_2)P_X(X = x_2)}{P_Y(Y = y)} \quad \Leftrightarrow \tag{1.17}$$

$$P_{Y|X}(Y = y|X = x_1)P_X(X = x_1) < P_{Y|X}(Y = y|X = x_2)P_X(X = x_2) \quad \Leftrightarrow \tag{1.18}$$

$$\frac{P_{Y|X}(Y = y|X = x_1)}{P_{Y|X}(Y = y|X = x_2)} < \frac{P_X(X = x_2)}{P_X(X = x_1)} \tag{1.19}$$

The last line is of particular interest as it elucidates the relationship between the prior and the likelihood. Only if the likelihood ratio for $X = x_1$ over $X = x_2$ is smaller than the reversed prior ratio will the posterior probability of $X = x_2$ be greater than that of $X = x_1$. This means that if we have strong very asymmetric priors (like $P_X(X = x_1) = 0.9$ and $P_X(X = x_2) = 0.1$), the likelihood needs to discriminate very well between the two cases in order to tip the scale in favour of $X = x_2$. In that sense the prior and the likelihood can be seen as battling forces whose equilibrium gives us the posterior.

But enough theory about Bayes' rule. It is about time you apply it. To that end we present you an exercise that is, in some variation, contained in virtually every textbook on probability theory, statistics or machine learning. Have fun with it!

Exercise 2 *A random person walks into the doctor's office to be tested for a particular disease. The disease can be fatal if not treated. However, successful treatment is possible if the disease is discovered early enough. It is commonly known that the disease occurs in 1 out of 1000 people of the country's population. The doctor will administer a test that with a probability of 99% returns a positive results if the patient does indeed have the disease. At the same time, the test also returns a positive result in 5% of the cases were the patient does not have the disease. After the test has been administered to the patient in question, it returns a positive result. What is the probability that the patient is infected with the disease?*
Proceed as follows:

1. *Write down a guess for what you think the probability might be (do not consider any math at this point).*

2. *Calculate that probability.*

3. *Check whether there is a considerable difference between your initial guess and the calculated probability. Go on to examine how the different factors have influenced the probability of the patient having the disease.*

Let us finish up this section with some more notation. In many applications of Bayes' rule we only want to know which outcome is the most likely, without worrying to much about the actual probabilities. Likewise, there is a range of situation where we just want to assign a score to outcomes and do not demand that this score be necessarily a probability. Throughout this section we have seen that in order to rank the values of an RV according to their probabilities we do not necessarily need to invoke the marginal likelihood since it cancels in all these comparisons anyway. Therefore you will often see authors stating that

$$P_X(X = x|Y = y) \propto P_{Y|X}(Y = y|X = x)P_X(X = x) \tag{1.20}$$

This reads as "the posterior is proportional to the product of the likelihood and the prior". In general if we have to quantities a and b, then by $a \propto b$ we mean that there is some constant $C \in \mathbb{R}$ such that $a = cB$. Notice that the probability distribution is a function and hence we require C to be the same across the domain of that function (that is C should be the same for all values of X).

Exercise 3 *What is the value of C in eq. 1.20?*

1.3 Naïve Bayes

In this section we will introduce a rather crude application of Bayes's rule which is surprisingly successful nonetheless. Assume that instead of one random variable we are observing a sequence of random variables. Thus our problem is the following:

$$P_{Y|X_1^n}(Y = y|X_1^n = x_1^n) \propto P_{X_1^n|Y}(X_1^n = x_1^n|Y = y) \times P_Y(Y = y) \tag{1.21}$$

By the chain rule we can decompose this into

$$\begin{aligned} P_{Y|X_1^n}(Y = y|X_1^n = x_1^n) \propto & \\ & P_{X_1|Y}(X_1 = x_1|Y = y) \times \dots \\ & \times P_{X_n|Y,X_1^{n-1}}(X_n = x_n|Y = y, X_1^{n-1}) \times P_Y(Y = y) \end{aligned} \tag{1.22}$$

So far everything is ok and mathematically correct. Now we are going to introduce the aforementioned crudeness into the model. We are just going to assume that all $X_i, 1 \leq i \leq n$ are conditionally independent given Y. Notice that this is just an assumption that we are making without justification. In fact, it is very likely wrong. However, it makes our live much easier because we will only have to deal with very simple terms of the form $P(X_i = x_i|Y = y)$. Because of the crudeness of our assumptions, this probabilistic model has come to be known as **naïve Bayes** (sometimes also stupid Bayes).

Definition 1 *A naïve Bayes model is a probabilistic model that assumes*

$$p(y|x_1^n) \propto p(y)p(x_1|y)p(x_2|y)\dots p(x_1|y)$$

Notice that in the definition we have employed pmfs in order to lighten up the notation a bit. Once we know all the component distributions, calculating the result is pretty straightforward.

In order to illustrate how naïve Bayes works we are going to employ one of its showcase applications where it indeed had a lot of success in real life. The application we are talking about is text classification. The task is the following: you are given some documents and for each of the documents you have to assign a label signifying its class. What you consider a class depends on you actual application setting, but usually classes are broad categories, such as legal texts, medical texts etc. If you manage to succeed at this task you can accomplish a lot of things automatically that needed to be done by humans before. For example, you could tag online news with their relevant categories and people who are interested in a particular category will then have an easier time finding the news related to that category. Crucially, since you will write a computer program that does the classification for you, you will not need to read any of the texts yourself. This will obviously allow you to classify huge quantities of text in a very short amount of time.

Exercise 4 *A collection of text (or any other kind of data for that matter) is often called a* ***corpus****. Here we are going to use a toy corpus. The corpus just consists of two sentences and we will assume that each sentence is its own document. Your task is to classify these two documents correctly and report the posterior probability for the correct label. The categories that you can label the documents with are finance (0), medicine (1) or law (2). Please find the corpus the pmfs of the needed distributions below. For simplicity's sake we are not going to distinguish between lower and upper case words (this is actually common practice). For better readability we are also going to use the actual words instead*

of their numerical encodings as values for the random variables. Just remind yourselves that those words are real numbers in reality.

The corpus:

- evidence ruled irrelevant or inadmissible shall not be considered by the chamber.
- that is probably what investors who poured $90 billion in european equity funds this year are hoping for.

Further Reading

Here, we have only scratched the surface of what Bayes' rule allows us to do. To get a wider outlook on what else is possible, you can consult Kevin Murphy's webpage.