Bayesian Updating with Discrete Priors
Class 11, 18.05, Spring 2014
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to apply Bayes theorem to compute probabilities.
2. Be able to identify the definition and roles of prior probability, likelihood (Bayes term), posterior probability, data and hypothesis in the application of Bayes Theorem.
3. Be able to use a Bayesian update table to compute posterior probabilities.

2 Review of Bayes theorem

Recall that Bayes theorem allows us to 'invert' conditional probabilities. If $\mathcal{H}$ and $\mathcal{D}$ are events, then:

$$P(\mathcal{H}\,|\,\mathcal{D}) = \frac{P(\mathcal{D}\,|\,\mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}$$

Our view is that Bayes theorem forms the foundation for inferential statistics. We will begin to justify this view today.

2.1 The base rate fallacy

When we first learned Bayes theorem we worked an example about screening tests showing that $P(\mathcal{D}|\mathcal{H})$ can be very different from $P(\mathcal{H}|\mathcal{D})$. In the appendix we work a similar example. If you are not comfortable with Bayes theorem you should read the example in the appendix now.

3 Terminology and Bayes theorem in tabular form

We now use a coin tossing problem to introduce terminology and a tabular format for Bayes theorem. This will provide a simple, uncluttered example that shows our main points.

Example 1. There are three types of coins which have different probabilities of landing heads when tossed.

- Type A coins are fair, with probability .5 of heads
- Type B coins are bent and have probability .6 of heads
- Type C coins are bent and have probability .9 of heads

Suppose I have a drawer containing 4 coins: 2 of type A, 1 of type B, and 1 of type C. I reach into the drawer and pick a coin at random. Without showing you the coin I flip it once and get heads. What is the probability it is type A? Type B? Type C?

answer: Let A, B, and C be the event that the chosen coin was type A, type B, and type C. Let $\mathcal{D}$ be the event that the toss is heads. The problem asks us to find

$$P(A|\mathcal{D}), \quad P(B|\mathcal{D}), \quad P(C|\mathcal{D}).$$

Before applying Bayes theorem, let's introduce some terminology.

- *Experiment*: pick a coin from the drawer at random, flip it, and record the result.
- *Data*: the result of our experiment. In this case the event $\mathcal{D}$ = 'heads'. We think of D as data that provides evidence for or against each hypothesis.
- *Hypotheses*: we are testing three hypotheses: the coin is type A, B or C.
- *Prior probability*: the probability of each hypothesis prior to tossing the coin (collecting data). Since the drawer has 2 coins of type A and 1 each of type B and C we have
$$P(A) = .5, \qquad P(B) = .25, \qquad P(C) = .25.$$
- *Likelihood*: (This is the same likelihood we used for the MLE.) The likelihood function is $P(\mathcal{D}|\mathcal{H})$, i.e., the probability of the data assuming that the hypothesis is true. Most often we will consider the data as fixed and let the hypothesis vary. For example, $P(\mathcal{D}|A)$ = probability of heads if the coin is type A. In our case the likelihoods are
$$P(\mathcal{D}|A) = .5, \qquad P(\mathcal{D}|B) = .6, \qquad P(\mathcal{D}|C) = .9.$$

  The name likelihood is so well established in the literature that we have to teach it to you. However in colloquial language likelihood and probability are synonyms. This leads to the likelihood function often being confused with the probabity of a hypothesis. Because of this we'd prefer to use the name Bayes term. However since we are stuck with 'likelihood' we will try to use it very carefully and in a way that minimizes any confusion.
- *Posterior probability*: After (posterior to) tossing the coin we use Bayes theorem to compute the probability of each hypothesis given the data. These posterior probabilities are what the problem asks us to find. It will turn out that
$$P(A|\mathcal{D}) = .4, \qquad P(B|\mathcal{D}) = .24, \qquad P(C|\mathcal{D}) = .36.$$

We can organize all of this very neatly in a table:

hypothesis	prior	likelihood	unnormalized posterior	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}\|\mathcal{D})$
A	.5	.5	.25	.4
B	.25	.6	.15	.24
C	.25	.9	.225	.36
total	1		.625	1

The unnormalized posterior (in red) is the product of the prior and the likelihood. It is "unnormalized" because it does not sum to 1. Bayes formula for, say, $P(A|\mathcal{D})$ is

$$P(A|\mathcal{D}) = \frac{P(\mathcal{D}|A)P(A)}{P(\mathcal{D})}$$

So Bayes theorem says that the posterior probability is obtained by dividing the unnormalized posterior by $P(\mathcal{D})$. We can find $P(\mathcal{D})$ using the law of total probability:

$$P(\mathcal{D}) = P(\mathcal{D}|A)P(A) + P(\mathcal{D}|B)P(B) + P(\mathcal{D}|C)P(C).$$

This is just the sum of the unnormalized (red) column. So to go from the unnormalized column to the final column, we simply divide by $P(\mathcal{D}) = .625$.

Bayesian updating: The process of going from the prior probability $P(\mathcal{H})$ to the posterior $P(\mathcal{H}|\mathcal{D})$ is called *Bayesian updating.*

3.1 Important things to notice

1. The posterior (after the data) probabilities for each hypothesis are in the last column. We see that coin A is still the most likely, though its probability has decreased from a prior probability of .5 to a posterior probability of .4. Meanwhile, the probability of type C has increased from .25 to .36.

2. The unnormalized posterior column determines the posterior probability column. To compute the latter, we simply rescaled the unnormalized posterior so that it sums to 1.

3. If all we care about is finding the most likely hypothesis, the unnormalized posterior works as well as the normalized posterior.

4. The likelihood column does not sum to 1. The likelihood function is *not* a probability function.

5. The posterior probability represents the outcome of a 'tug-of-war' between the likelihood and the prior. When calculating the posterior, a large prior may be deflated by a small likelihood, and a small prior may be inflated by a large likelihood.

6. The maximum likelihood estimate (MLE) for Example 1 is hypothesis C, with a likelihood $P(\mathcal{D}|C) = .9$. The MLE is useful, but you can see in this example that it is not the entire story.

Terminology in hand, we can express Bayes theorem in various ways:

$$P(\mathcal{H}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}$$

$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{data}|\text{hypothesis})P(\text{hypothesis})}{P(\text{data})}$$

With the data fixed, the denominator $P(\mathcal{D})$ just serves to normalize the total posterior probability to 1. So we can also express Bayes theorem as a statement about the proportionality of two functions of $\mathcal{H}$ (i.e, of the last two columns of the table).

$$P(\text{hypothesis}|\text{data}) \;\propto\; P(\text{data}|\text{hypothesis})P(\text{hypothesis})$$

This leads to the most elegant form of Bayes theorem in the context of Bayesian updating:

$$\boxed{\text{posterior} \;\propto\; \text{likelihood} \times \text{prior}}$$

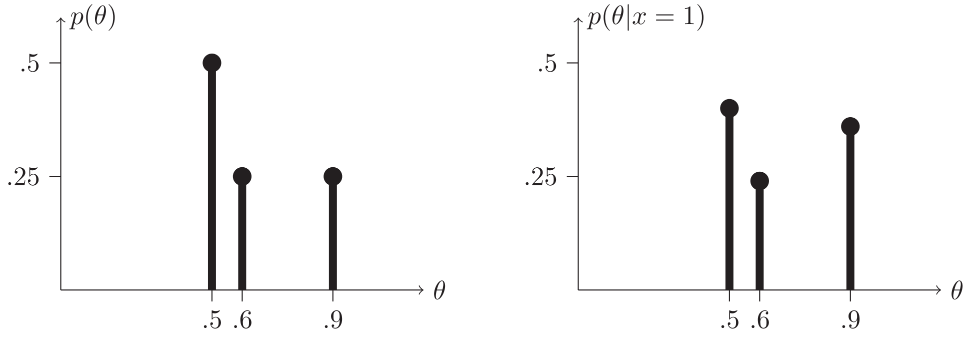
3.2 Prior and posterior probability mass functions

We saw earlier in the course that it was convenient to use random variables and probability mass functions. To do this we had to assign values to events (head is 1 and tails is 0). We will do the same thing in the context of Bayesian updating.

In Example 1 we can represent the three hypotheses A, B, and C by $\theta = .5, .6, .9$. For the data we'll let $x = 1$ mean heads and $x = 0$ mean tails. Then the prior and posterior probabilities in the table define prior and posterior probability mass functions.

prior pmf	$p(\theta)$:	$p(.5) = P(A)$,	$p(.6) = P(B)$,	$p(.9) = P(C)$
post. pmf	$p(\theta \,\vert\, x = 1)$:	$p(.5 \,\vert\, x = 1) = P(A\vert D)$,	$p(.6 \,\vert\, x = 1) = P(B\vert D)$,	$p(9 \,\vert\, x = 1) = P(C\vert D)$

Here are plots of the prior and posterior pmf's from the example.



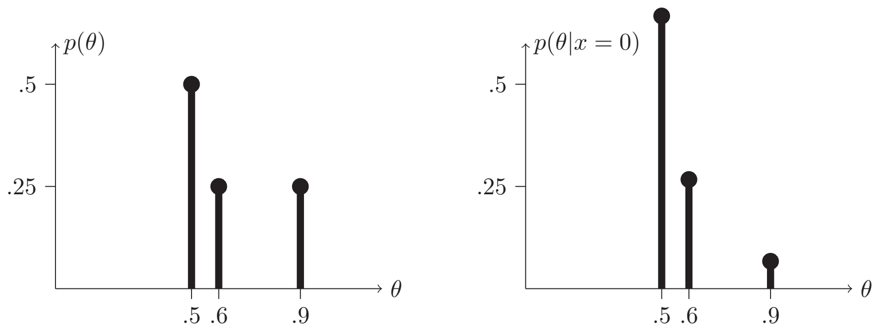
Prior pmf $p(\theta)$ and posterior pmf $p(\theta|x = 1)$ for Example 1

Example 2. Using the notation $p(\theta)$, etc., redo Example 1 assuming the flip was tails.

answer: We redo the table with data of tails represented by $x = 0$.

			unnormalized	
hypothesis	prior	likelihood	posterior	posterior
θ	$p(\theta)$	$p(x = 0 \,\vert\, \theta)$	$p(x = 0 \,\vert\, \theta)p(\theta)$	$p(\theta \,\vert\, x = 0)$
.5	.5	.5	.25	.6667
.6	.25	.4	.1	.2667
.9	.25	.1	.025	.0667
total	1		.375	1

Now the probability of type A has increased from .5 to .6667, while the probability of type C has decreased from .25 to only 0.0667. Here are the corresponding plots:



Prior pmf $p(\theta)$ and posterior pmf $p(\theta|x=0)$

3.3 Food for thought.

Suppose that in Example 1 you didn't know how many coins of each type were in the drawer. You picked one at random and got heads. How would you go about deciding which hypothesis (coin type) if any was most supported by the data?

4 Updating again and again

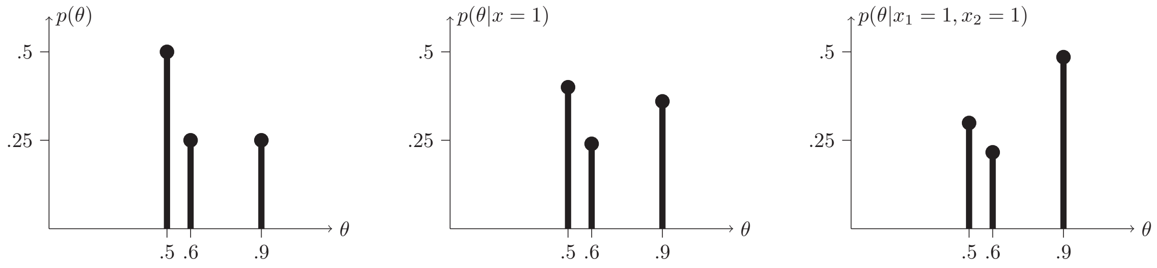
In life we are continually updating our beliefs with each new experience of the world. In Bayesian inference, after updating the prior to the posterior, we can take more data and update again! For the second update, the posterior from the first data becomes the prior for the second data.

Example 3. Suppose you have picked a coin as in Example 1. You flip it once and get heads. Then you flip the same coin and get heads again. What is the probability that the coin was type A? Type B? Type C?

answer: Since both flips were heads, the likelihood function is the same for the first and second rolls. So we only record the likelihood function once in our table.

hypothesis	prior	likelihood	unnormalized posterior 1	unnormalized posterior 2	posterior 2
θ	$p(\theta)$	$p(x_1=1\|\theta)$	$p(x_1=1\|\theta)p(\theta)$	$p(x_2=1\|\theta)p(x_1=1\|\theta)p(\theta)$	$p(\theta\|x_1=1,x_2=1)$
.5	.5	.5	.25	.125	.299
.6	.25	.6	.15	.09	.216
.9	.25	.9	.225	.2025	.485
total	1			.41750	1

Note that the second unnormalized posterior is computed by multiplying the first unnormalized posterior and the likelihood; since we are only interested in the final posterior, there is no need to normalize until the last step. As shown in the last column and plot, after two heads the type C hypothesis has finally taken the lead!



The prior $p(\theta)$, first posterior $p(\theta|x_1 = 1)$, and second posterior $p(\theta|x_1 = 1, x_2 = 1)$

5 Appendix: the base rate fallacy

Example 4. A screening test for a disease is both sensitive and specific. By that we mean it is usually positive when testing a person with the disease and usually negative when testing someone without the disease. Let's assume the true positive rate is 99% and the false positive rate is 2%. Suppose the prevalence of the disease in the general population is 0.5%. If a random person tests positive, what is the probability that they have the disease?

answer: As a review we first do the computation using trees. Next we will redo the computation using tables.

Let's use notation established above for hypotheses and data: let $\mathcal{H}_+$ be the hypothesis (event) that the person has the disease and let $\mathcal{H}_-$ be the hypothesis they do not. Likewise, let $\mathcal{D}_+$ and $\mathcal{D}_-$ represent the data of a positive and negative screening test respectively. We are asked to compute $P(\mathcal{H}_+|\mathcal{D}_+)$.

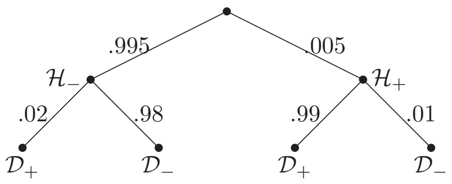
We are given

$$P(\mathcal{D}_+|\mathcal{H}_+) = .99, \quad P(\mathcal{D}_+|\mathcal{H}_-) = .02, \quad P(\mathcal{H}_+) = .005.$$

From these we can compute the false negative and true negative rates:

$$P(\mathcal{D}_-|\mathcal{H}_+) = .01, \quad P(\mathcal{D}_-|\mathcal{H}_-) = .98$$

All of these probabilities can be displayed quite nicely in a tree.



Bayes theorem yields

$$P(\mathcal{H}_+|\mathcal{D}_+) = \frac{P(\mathcal{D}_+|\mathcal{H}_+)P(\mathcal{H}_+)}{P(\mathcal{D}_+)} = \frac{.99 \cdot .005}{.99 \cdot .005 + .02 \cdot .995} = 0.19920 \approx 20\%$$

Now we redo this calculation using a Bayesian update table:

hypothesis	prior	likelihood	unnormalized posterior	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{D}_+\|\mathcal{H})$	$P(\mathcal{D}_+\|\mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}\|\mathcal{D}_+)$
$\mathcal{H}_+$	.005	.99	.00495	.19920
$\mathcal{H}_-$	.995	.02	.01990	.80080
total	1		.02485	1

The table shows that the posterior probability $P(\mathcal{H}_+|\mathcal{D}_+)$ that a person with a positive test has the disease is about 20%. This is far less than the sensitivity of the test (99%) but much higher than the prevalence of the disease in the general population (0.5%).