Bayesian Updating: Probabilistic Prediction and odds
Class 12, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to use the law of total probability to compute prior and posterior predictive probabilities.
2. Be able to convert between odds and probability.
3. Be able to update prior odds to posterior odds using Bayes factors.
4. Understand how Bayes factors measure the extent to which data provides evidence for or against a hypothesis.

2 Introduction to probabilistic prediction

In the previous class we looked at updating the probability of hypotheses based on data. We can also use the data to update the probability of each possible outcome of a future experiment. In this class we will look at how this is done.

2.1 Probabilistic prediciton; words of estimative probability (WEP)

There are many ways to word predictions:

- Prediction: "It will rain tomorrow."
- Prediction using words of estimative probability (WEP): "It is likely to rain tomorrow."
- Probabilistic prediction: "Tomorrow it will rain with probability 60% (and not rain with probability 40%)."

Each type of wording is appropriate at different times.

In this class we are going to focus on probabilistic prediction and precise quantitative statements. You can see `http://en.wikipedia.org/wiki/Words_of_Estimative_Probability` for an interesting discussion about the appropriate use of words of estimative probability. The article also contains a list of *weasel words* such as 'might', 'cannot rule out', 'it's conceivable' that should be avoided as almost certain to cause confusion.

There are many places where we want to make a probabilistic prediction. Examples are

- Medical treatment outcomes
- Weather forecasting

- Climate change
- Sports betting
- Elections
- ...

These are all situations where there is uncertainty about the outcome and we would like as precise a description of what could happen as possible.

3 Predictive Probabilities

Probabilistic prediction simply means assigning a probability to each possible outcomes of an experiment.

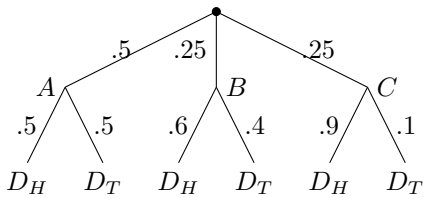
Recall the coin example from the previous class notes: there are three types of coins which are indistinguishable apart from their probability of landing heads when tossed.

- Type A coins are fair, with probability 0.5 of heads
- Type B coins have probability 0.6 of heads
- Type C coins have probability 0.9 of heads

You have a drawer containing 4 coins: 2 of type A, 1 of type B, and 1 of type C. You reach into the drawer and pick a coin at random. We let A stand for the event 'the chosen coin is of type A'. Likewise for B and C.

3.1 Prior predictive probabilities

Before taking data we can compute the probability that our chosen coin will land heads (or tails) if flipped. Let D_H be the event it lands heads and let D_T the event it lands tails. We can use the law of total probability to determine the probabilities of these events. Either by drawing a tree or directly proceeding to the algebra, we get:



$$\begin{aligned} P(D_H) &= P(D_H|A)P(A) + P(D_H|B)P(B) + P(D_H|C)P(C) \\ &= 0.5 \cdot 0.5 + 0.6 \cdot 0.25 + 0.9 \cdot 0.25 = 0.625 \\ P(D_T) &= P(D_T|A)P(A) + P(D_T|B)P(B) + P(D_T|C)P(C) \\ &= 0.5 \cdot 0.5 + 0.4 \cdot 0.25 + 0.1 \cdot 0.25 = 0.375 \end{aligned}$$

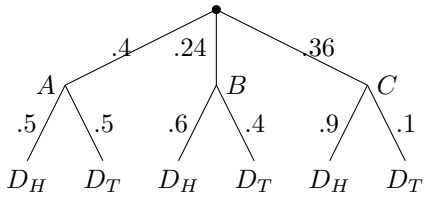
Definition: These probabilities give a (probabilistic) prediction of what will happen if the coin is tossed. Because they are computed before we collect any data they are called prior predictive probabilities.

3.2 Posterior predictive probabilities

Suppose we flip the coin once and it lands heads. We now have data D, which we can use to update the prior probabilities of our hypotheses to posterior probabilities. Last class we learned to use a Bayes table to facilitate this computation:

hypothesis	prior	likelihood	Bayes numerator	posterior
H	$P(H)$	$P(D\|H)$	$P(D\|H)P(H)$	$P(H\|D)$
A	0.5	0.5	0.25	0.4
B	0.25	0.6	0.15	0.24
C	0.25	0.9	0.225	0.36
total	1		0.625	1

Having flipped the coin once and gotten heads, we can compute the probability that our chosen coin will land heads (or tails) if flipped a second time. We proceed just as before, but using the posterior probabilities $P(A|D)$, $P(B|D)$, $P(C|D)$ in place of the prior probabilities $P(A)$, $P(B)$, $P(C)$.



$$\begin{aligned} P(D_H|D) &= P(D_H|A)P(A|D) + P(D_H|B)P(B|D) + P(D_H|C)P(C|D) \\ &= 0.5 \cdot 0.4 + 0.6 \cdot 0.24 + 0.9 \cdot 0.36 = 0.668 \\ P(D_T|D) &= P(D_T|A)P(A|D) + P(D_T|B)P(B|D) + P(D_T|C)P(C|D) \\ &= 0.5 \cdot 0.4 + 0.4 \cdot 0.24 + 0.1 \cdot 0.36 = 0.332 \end{aligned}$$

Definition: These probabilities give a (probabilistic) prediction of what will happen if the coin is tossed again. Because they are computed after collecting data and updating the prior to the posterior, they are called posterior predictive probabilities.

Note that heads on the first toss increases the probability of heads on the second toss.

3.3 Review

Here's a succinct description of the preceding sections that may be helpful:

Each hypothesis gives a different probability of heads, so the total probability of heads is a weighted average. For the prior predictive probability of heads, the weights are given by the prior probabilities of the hypotheses. For the posterior predictive probability of heads, the weights are given by the posterior probabilities of the hypotheses.

Remember: Prior and posterior probabilities are for hypotheses. Prior predictive and posterior predictive probabilities are for data. To keep this straight, remember that the latter predict future data.

4 Introduction to odds

When comparing two events, it common to phrase probability statements in terms of odds.

Definition The odds of event E versus event E' are the ratio of their probabilities $P(E)/P(E')$. If unspecified, the second event is assumed to be the complement E^c. So the odds of E are:

$$O(E) = \frac{P(E)}{P(E^c)}.$$

For example, $O(\text{rain}) = 2$ means that the probability of rain is twice the probability of no rain (2/3 versus 1/3). We might say 'the odds of rain are 2 to 1.'

Example. For a fair coin, $O(\text{heads}) = \dfrac{1/2}{1/2} = 1$. We might say the odds of heads are 1 to 1 or fifty-fifty.

Example. For a standard die, the odds of rolling a 4 are $\dfrac{1/6}{5/6} = \dfrac{1}{5}$. We might say the odds are '1 to 5 for' or '5 to 1 against' rolling a 4.

Example. The probability of a pair in a five card poker hand is 0.42257. So the odds of a pair are 0.42257/(1-0.42257) = 0.73181.

We can go back and forth between probability and odds as follows.

Conversion formulas: if $P(E) = p$ then $O(E) = \dfrac{p}{1-p}$. If $O(E) = q$ then $P(E) = \dfrac{q}{1+q}$.

Notes:

1. The second formula simply solves $q = p/(1-p)$ for p.
2. Probabilities are between 0 and 1, while odds are between 0 to ∞.
3. The property $P(E^c) = 1 - P(E)$ becomes $O(E^c) = 1/O(E)$.

Example. Let F be the event that a five card poker hand is a full house. Then $P(F) = 0.00145214$ so $O(F) = 0.0014521/(1 - 0.0014521) = 0.0014542$.

The odds not having a full house are $O(F^c) = (1 - 0.0014521)/0.0014521 = 687 = 1/O(F)$.

4. If $P(E)$ or $O(E)$ is small then $O(E) \approx P(E)$. This follows from the conversion formulas.

Example. In the poker example where $F =$ 'full house' we saw that $P(F)$ and $O(F)$ differ only in the fourth significant digit.

5 Updating odds

5.1 Introduction

In Bayesian updating, we used the likelihood of data to update prior probabilities of hypotheses to posterior probabilities. In the language of odds, we will update prior odds to posterior odds. One of our key points will be that the data can provide evidence supporting or negating a hypothesis depending on whether its posterior odds are greater or less than its prior odds.

5.2 Example: Marfan syndrome

Marfan syndrome is a genetic disease of connective tissue that occurs in 1 of every 15000 people. The main ocular features of Marfan syndrome include bilateral ectopia lentis (lens dislocation), myopia and retinal detachment. About 70% of people with Marfan syndrome have a least one of these ocular features; only 7% of people without Marfan syndrome do. (We don't guarantee the accuracy of these numbers, but they will work perfectly well for our example.)

If a person has at least one of these ocular features, what are the odds that they have Marfan syndrome?

answer: This is a standard Bayesian updating problem. Our hypotheses are:
M = 'the person has Marfan syndrome'
M^c = 'the person does not have Marfan syndrome'
The data is:
F = 'the person has at least one ocular feature'.

We are given the prior probability of M and the likelihoods of F given M or M^c:

$$P(M) = 1/15000, \qquad P(F|M) = 0.7, \qquad P(F|M^c) = 0.07.$$

As before, we can compute the posterior probabilities using a table:

hypothesis	prior	likelihood	Bayes numerator	posterior
H	$P(H)$	$P(F\|H)$	$P(F\|H)P(H)$	$P(H\|F)$
M	0.000067	0.7	0.0000467	0.00066
M^c	0.999933	0.07	0.069995	0.99933
total	1		0.07004	1

First we find the prior odds:

$$O(M) = \frac{P(M)}{P(M^c)} = \frac{1/15000}{14999/15000} = \frac{1}{14999} \approx 0.000067.$$

The posterior odds are given by the ratio of the posterior probabilities or the Bayes numerators, since the normalizing factor will be the same in both numerator and denominator.

$$O(M|F) = \frac{P(M|F)}{P(M^c|F)} = \frac{P(F|M)P(M)}{P(F|M^c)P(M^c)} = 0.000667.$$

The posterior odds are a factor of 10 larger than the prior odds. In that sense, having an ocular feature is strong evidence in favor of the hypothesis M. However, because the prior odds are so small, it is still highly unlikely the person has Marfan syndrome.

6 Bayes factors and strength of evidence

The factor of 10 in the previous example is called a Bayes factor. The exact definition is the following.

Definition: For a hypothesis H and data D, the Bayes factor is the ratio of the likelihoods:

$$\text{Bayes factor} = \frac{P(D|H)}{P(D|H^c)}.$$

Let's see exactly where the Bayes factor arises in updating odds. We have

$$\begin{aligned} O(H|D) &= \frac{P(H|D)}{P(H^c|D)} \\ &= \frac{P(D|H)\,P(H)}{P(D|H^c)P(H^c)} \\ &= \frac{P(D|H)}{P(D|H^c)} \cdot \frac{P(H)}{P(H^c)} \\ &= \frac{P(D|H)}{P(D|H^c)} \cdot O(H) \end{aligned}$$

$$\text{posterior odds} = \textbf{Bayes factor} \times \text{prior odds}$$

From this formula, we see that the Bayes' factor (BF) tells us whether the data provides evidence for or against the hypothesis.

- If $BF > 1$ then the posterior odds are greater than the prior odds. So the data provides evidence for the hypothesis.
- If $BF < 1$ then the posterior odds are less than the prior odds. So the data provides evidence against the hypothesis.
- If $BF = 1$ then the prior and posterior odds are equal. So the data provides no evidence either way.

The following example is taken from the textbook *Information Theory, Inference, and Learning Algorithms* by David J. C. Mackay, who has this to say regarding trial evidence.

> In my view, a jury's task should generally be to multiply together carefully evaluated likelihood ratios from each independent piece of admissible evidence with an equally carefully reasoned prior probability. This view is shared by many statisticians but learned British appeal judges recently disagreed and actually overturned the verdict of a trial because the jurors *had* been taught to use Bayes' theorem to handle complicated DNA evidence.

Example 1. Two people have left traces of their own blood at the scene of a crime. A suspect , Oliver, is tested and found to have type 'O' blood. The blood groups of the two traces are found to be of type 'O' (a common type in the local population, having frequency 60%) and type 'AB' (a rare type, with frequency 1%). Does this data (type 'O' and 'AB' blood were found at the scene) give evidence in favor of the proposition that Oliver was one of the two people present at the scene of the crime?"

answer: There are two hypotheses:
S = 'Oliver and another unknown person were at the scene of the crime
S^c = 'two unknown people were at the scene of the crime'

The data is:
D = 'type 'O' and 'AB' blood were found'

The Bayes factor for Oliver's presence is $BF_{\text{Oliver}} = \dfrac{P(D|S)}{P(D|S^c)}$. We compute the numerator and denominator of this separately.

The data says that both type O and type AB blood were found. If Oliver was at the scene then 'type O' blood would be there. So $P(D|S)$ is the probability that the other person had type AB blood. We are told this is .01, so $P(D|S) = 0.01$.

If Oliver was not at the scene then there were two random people one with type O and one with type AB blood. The probability of this is $2 \cdot 0.6 \cdot 0.01$. The factor of 2 is because there are two ways this can happen –the first person is type O and the second is type AB or vice versa.*

Thus the Bayes factor for Oliver's presence is

$$BF_{\text{Oliver}} = \frac{P(D|S)}{P(D|S^c)} = \frac{0.01}{2 \cdot 0.6 \cdot 0.01} = 0.83.$$

Since $BF_{\text{Oliver}} < 1$, the data provides (weak) evidence against Oliver being at the scene.

*We have assumed the blood types of the two people are independent. This is not precisely true, but for a large population it is close enough. The exact probability is $\dfrac{2 \cdot N_O \cdot N_{AB}}{N \cdot (N-1)}$ where N_O is the number of people with type O blood, N_{AB} the number with type AB blood and N the size of the population. We have $\dfrac{N_O}{N} = 0.6$. For large N we have $N \approx N - 1$, so $\dfrac{N_{AB}}{N-1} \approx 0.01$. This shows the probability is approximately $2 \cdot 0.6 \cdot 0.01$ as claimed.

Example 2. Another suspect Alberto is found to have type 'AB' blood. Do the same data give evidence in favor of the proposition that Alberto was one of the two people present at the crime?

answer: Reusing the above notation with Alberto in place of Oliver we have:

$$BF_{\text{Alberto}} = \frac{P(D|S)}{P(D|S^c)} = \frac{0.6}{2 \cdot 0.6 \cdot 0.01} = 50.$$

Since $BF_{\text{Alberto}} \gg 1$, the data provides strong evidence in favor of Alberto being at the scene.

Notes:
1. In both examples, we have only computed the Bayes factor, not the posterior odds. To compute the latter, we would need to know the prior odds that Oliver (or Alberto) was at the scene based on other evidence.

2. Note that if 50% of the population had type O blood instead of 60%, then the Oliver's Bayes factor would be 1 (neither for nor against). More generally, the break-even point for blood type evidence is when the proportion of the suspect's blood type in the general population equals the proportion of the suspect's blood type among those who left blood at the scene.

6.1 Updating again and again

Suppose we collect data in two stages, first D_1, then D_2. We have seen in our dice and coin examples that the final posterior can be computed all at once or in two stages where we first update the prior using the likelihoods for D_1 and then update the resulting posterior using the likelihoods for D_2. The latter approach works whenever likelihoods multiply:

$$P(D_1, D_2|H) = P(D_1|H)P(D_2|H).$$

Since likelihoods are conditioned on hypotheses, we say that D_1 and D_2 are conditionally independent if the above equation holds for every hypothesis H.

Example. There are five dice in a drawer, with 4, 6, 8, 12, and 20 sides (these are the hypotheses). I pick a die at random and roll it twice. The first roll gives 7. The second roll gives 11. Are these results conditionally independent? Are they independent?

answer: These results are conditionally independent. For example, for the hypothesis of the 8-sided die we have:

$$\begin{aligned} P(\text{7 on roll 1} \,|\, \text{8-sided die}) &= 1/8 \\ P(\text{11 on roll 2} \,|\, \text{8-sided die}) &= 0 \\ P(\text{7 on roll 1, 11 on roll 2} \,|\, \text{8-sided die}) &= 0 \end{aligned}$$

For the hypothesis of the 20-sided die we have:

$$\begin{aligned} P(\text{7 on roll 1} \,|\, \text{20-sided die}) &= 1/20 \\ P(\text{11 on roll 2} \,|\, \text{20-sided die}) &= 1/20 \\ P(\text{7 on roll 1, 11 on roll 2} \,|\, \text{20-sided die}) &= (1/20)^2 \end{aligned}$$

However, the results of the rolls are *not* independent. That is:

$$P(\text{7 on roll 1, 11 on roll 2}) \neq P(\text{7 on roll 1})P(\text{11 on roll 2}).$$

Intuitively, this is because a 7 on the roll 1 allows us to rule out the 4- and 6-sided dice, making an 11 on roll 2 more likely. Let's check this intuition by computing both sides precisely. On the righthand side we have:

$$\begin{aligned} P(\text{7 on roll 1}) &= \frac{1}{5}\cdot\frac{1}{8} + \frac{1}{5}\cdot\frac{1}{12} + \frac{1}{5}\cdot\frac{1}{20} = \frac{31}{600} \\ P(\text{11 on roll 2}) &= \frac{1}{5}\cdot\frac{1}{12} + \frac{1}{5}\cdot\frac{1}{20} = \frac{2}{75} \end{aligned}$$

On the lefthand side we have:

$$\begin{aligned} P(\text{7 on roll 1, 11 on roll 2}) &= P(\text{11 on roll 2} \,|\, \text{7 on roll 1})P(\text{7 on roll 1}) \\ &= \left(\frac{30}{93}\cdot\frac{1}{12} + \frac{6}{31}\cdot\frac{1}{20}\right)\cdot\frac{31}{600} \\ &= \frac{17}{465}\cdot\frac{31}{600} = \frac{17}{9000} \end{aligned}$$

Here $\frac{30}{93}$ and $\frac{6}{31}$ are the posterior probabilities of the 12- and 20-sided dice given a 7 on roll 1. We conclude that, without conditioning on hypotheses, the rolls are not independent.

Returning the to general setup, if D_1 and D_2 are conditionally independent for H and H^c then it makes sense to consider each Bayes factor independently:

$$BF_i = \frac{P(D_i|H)}{P(D_i|H^c)}.$$

The prior odds of H are $O(H)$. The posterior odds after D_1 are

$$O(H|D_1) = BF_1 \cdot O(H).$$

And the posterior odds after D_1 and D_2 are

$$\begin{aligned} O(H|D_1, D_2) &= BF_2 \cdot O(H|D_1) \\ &= BF_2 \cdot BF_1 \cdot O(H) \end{aligned}$$

We have the beautifully simple notion that updating with new data just amounts to multiplying the current posterior odds by the Bayes factor of the new data.

Example 3. Other symptoms of Marfan Syndrome

Recall from the earlier example that the Bayes factor for a least one ocular feature (F) is

$$BF_F = \frac{P(F|M)}{P(F|M^c)} = \frac{0.7}{0.07} = 10.$$

The wrist sign (W) is the ability to wrap one hand around your other wrist to cover your pinky nail with your thumb. Assume 10% of the population have the wrist sign, while 90% of people with Marfan's have it. Therefore the Bayes factor for the wrist sign is

$$BF_W = \frac{P(W|M)}{P(W|M^c)} = \frac{0.9}{0.1} = 9.$$

We will assume that F and W are conditionally independent symptoms. That is, among people with Marfan syndrome, ocular features and the wrist sign are independent, and among people without Marfan syndrome, ocular features and the wrist sign are independent. Given this assumption, the posterior odds of Marfan syndrome for someone with both an ocular feature and the wrist sign are

$$O(M|F, W) = BF_W \cdot BF_F \cdot O(M) = 9 \cdot 10 \cdot \frac{1}{14999} \approx \frac{6}{1000}.$$

We can convert the posterior odds back to probability, but since the odds are so small the result is nearly the same:

$$P(M|F, W) \approx \frac{6}{1000 + 6} \approx 0.596\%.$$

So ocular features and the wrist sign are both strong evidence in favor of the hypothesis M, and taken together they are very strong evidence. Again, because the prior odds are so small, it is still unlikely that the person has Marfan syndrome, but at this point it might be worth undergoing further testing given potentially fatal consequences of the disease (such as aortic aneurysm or dissection).

Interesting

Note also that if a person has exactly one of the two symptoms, then the product of the Bayes factors is near 1 (either 9/10 or 10/9). So the two pieces of data essentially cancel each other out with regard to the evidence they provide for Marfan's syndrome.

7 Log odds

In practice, people often find it convenient to work with the natural log of the odds in place of odds. Naturally enough these are called the log odds. The Bayesian update formula

$$O(H|D_1, D_2) = BF_2 \cdot BF_1 \cdot O(H)$$

becomes

$$\ln(O(H|D_1, D_2)) = \ln(BF_2) + \ln(BF_1) + \ln(O(H)).$$

We can interpret the above formula for the posterior log odds as the sum of the prior log odds and all the evidence $\ln(BF_i)$ provided by the data. Note that by taking logs, evidence in favor ($BF_i > 1$) is positive and evidence against ($BF_i < 1$) is negative.

To avoid lengthier computations, we will work with odds rather than log odds in this course. Log odds are nice because sums are often more intuitive then products. Log odds also play a central role in logistic regression, an important statistical model related to linear regression.