Chapter 9 Bayesian Inference

Bibliography: Jeremy Orloff and, Jonathan Bloom, Lectures MIT

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available.

9.1 Review of Bayes' theorem

Recall that Bayes' theorem allows us to `invert' conditional probabilities. If *H* and *D* are events, then:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \qquad (9.1)$$

When we first learned Bayes' theorem we worked an example about screening tests showing that $P(D|H)$ can be very different from $P(H|D)$.

Example:

A screening test for a disease is sensitive. 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?

We first do the computation using trees. Next we will redo the computation using tables.

Notation: Let H_+ be the hypothesis (event) that the person has the disease and let H_- be the hypothesis they do not. Likewise, let T_+ and T_- represent the data of a positive and negative screening test respectively. We are asked to compute $P(H_+|T_+)$

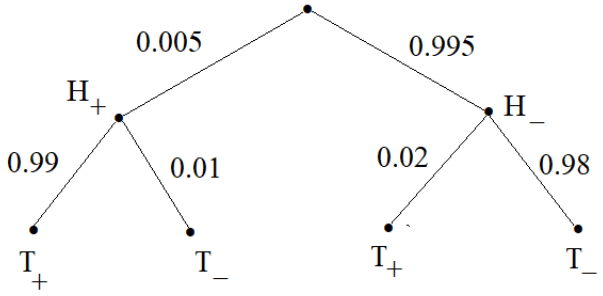
We are given:

$$P(T_+|H_+)=0.99 \qquad P(T_+|H_-)=0.02 \qquad P(H_+)=0.005$$

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

$$P(T_-|H_+)=0.01 \qquad P(T_-|H_-)=0.98$$

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



Bayes' theorem yields

$$P(H_+|T_+)=\frac{P(T_+|H_+)P(H_+)}{P(T_+)}=$$

$$=\frac{0.99\times 0.005}{0.02485}=0.199195\approx 20\%$$

$P(T_+)=P(T_+|H_+)P(H_+)+P(T_+|H_-)P(H_-)$ the total probability law

$$=0.99\times 0.005+0.02\times 0.995=$$

$$=0.00495+0.0199=0.02485$$

Now we redo this calculation using a *Bayesian update table*:

Hypothesis H	Prior $P(H)$	Likelihood $P(T_+\|H)$	Bayes numerator $P(T_+\|H)P(H)$	Posterior $P(H\|T_+)$
H_+	0.005	0.99	0.00495	0.19919
H_-	0.995	0.02	0.01990	0.80080
Total	1	No sum	0.02485	1

The table shows that the posterior probability $P(H_+|T_+)$ 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%).

9.2 Terminology and Bayes' theorem in tabular form

We now use a coin tossing problem to introduce terminology and a tabular format for Bayes' theorem.

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

Type *A* coins are fair, with probability 0.5 of heads

Type *B* coins are bent and have probability 0.6 of heads

Type *C* coins are bent and have probability 0.9 of heads

Suppose I have a drawer containing 5 coins: 2 of type *A*, 2 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*?

Let *A*, *B*, and *C* be the event that the chosen coin was type *A*, type *B*, and type *C*. Let *D* be the event that the toss is heads. The problem asks us to find

$$P(A|D) \qquad P(B|D) \qquad P(C|D)$$

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 *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, 2 of type B and 1of type C we have

$$P(A) = 0.4 \qquad P(B) = 0.4 \qquad P(C) = 0.2$$

Likelihood: (This is the same likelihood we used for the MLE.) The likelihood function is $P(D|H)$, 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,

$P(D|A)$ = probability of heads if the coin is type A.

In our case the likelihoods are

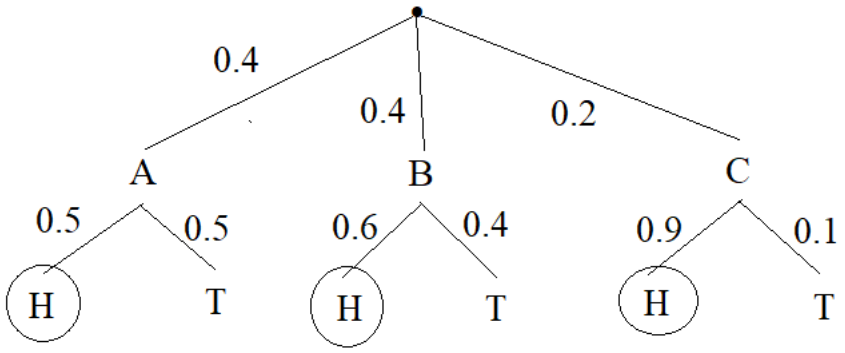
$$P(D|A) = 0.5 \qquad P(D|B) = 0.6 \qquad P(D|C) = 0.9$$

Posterior probability: the probability (posterior to) of each hypothesis given the data from tossing the coin.

$$P(A|D) \qquad P(B|D) \qquad P(C|D)$$

These posterior probabilities are what the problem asks us to find.

We now use Bayes' theorem to compute each of the posterior probabilities. First we organize the probabilities into a tree:



Probability tree for choosing and tossing a coin.

Bayes' theorem says, e.g.

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

The denominator $P(D)$ is computed using the law of total probability:

$$P(D) = P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C) =$$

$$= 0.5 \times 0.4 + 0.6 \times 0.4 + 0.9 \times 0.2 =$$

$$= 0.2 + 0.24 + 0.18 = 0.62$$

Now each of the three posterior probabilities can be computed:

$$P(A|D) = \frac{P(D|A)P(A)}{P(D)} = \frac{0.5 \times 0.4}{0.62} = \frac{0.2}{0.62}$$

$$P(B|D) = \frac{P(D|B)P(B)}{P(D)} = \frac{0.6 \times 0.4}{0.62} = \frac{0.24}{0.62}$$

$$P(C|D) = \frac{P(D|C)P(C)}{P(D)} = \frac{0.9 \times 0.2}{0.62} = \frac{0.18}{0.62}$$

Notice that the total probability $P(D)$ is the same in each of the denominators and that it is the sum of the three numerators.

We can organize all of this in a *Bayesian update table*:

Hypothesis H	Prior $P(H)$	Likelihood $P(D\|H)$	Bayes numerator $P(D\|H)P(H)$	Posterior $P(H\|D)$
A	0.4	0.5	0.2	0.32258
B	0.4	0.6	0.24	0.38709
C	0.2	0.9	0.18	0.29032
total	1		0.62	1

The Bayes numerator is the product of the prior and the likelihood. We see in each of the Bayes' formula computations above that the posterior probability is obtained by dividing the Bayes numerator by $P(D) = 0.62$. We also see that the law of total probability says that $P(D)$ is the sum of the entries in the Bayes numerator column.

Bayesian updating: The process of going from the prior probability $P(H)$ to the posterior $P(H|D)$ is called *Bayesian updating*. Bayesian updating uses the data to alter our understanding of the probability of each of the possible hypotheses.

Important things to notice:

1.There are two types of probabilities: Type one is the standard probability of data, e.g. the probability of heads is $p = 0.9$. Type two is the probability of the hypotheses, e.g. the probability the chosen coin is type A, B or C. This second type has prior (before the data) and posterior (after the data) values.

2.The posterior (after the data) probabilities for each hypothesis are in the last column. We see that coin B is now the most probable, though its probability has decreased from a prior probability of 0.4 to a posterior probability of 0.39. Meanwhile, the probability of type C has increased from 0.2 to 0.29.

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

4. If all we care about is finding the most likely hypothesis, the Bayes numerator works as well as the normalized posterior.

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

6.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.

7.The maximum likelihood estimate (MLE) for Exercise 1 is hypothesis C, with a likelihood $P(D|C) = 0.9$. The MLE is useful, but you can see in this example that it is not the entire story, since type B has the greatest posterior probability. Terminology in hand, we can express Bayes' theorem in various ways:

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

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

With the data fixed, the denominator $P(D)$ 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 *H* (i.e, of the last two columns of the table).

$$P(\text{hypothesis}|\text{data}) \sim P(\text{data}|\text{hypothesis})\,P(\text{hypothesis})$$

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

$$\text{posterior} \sim \text{likelihood} \times \text{prior}$$

9.3 Prior and posterior probability functions

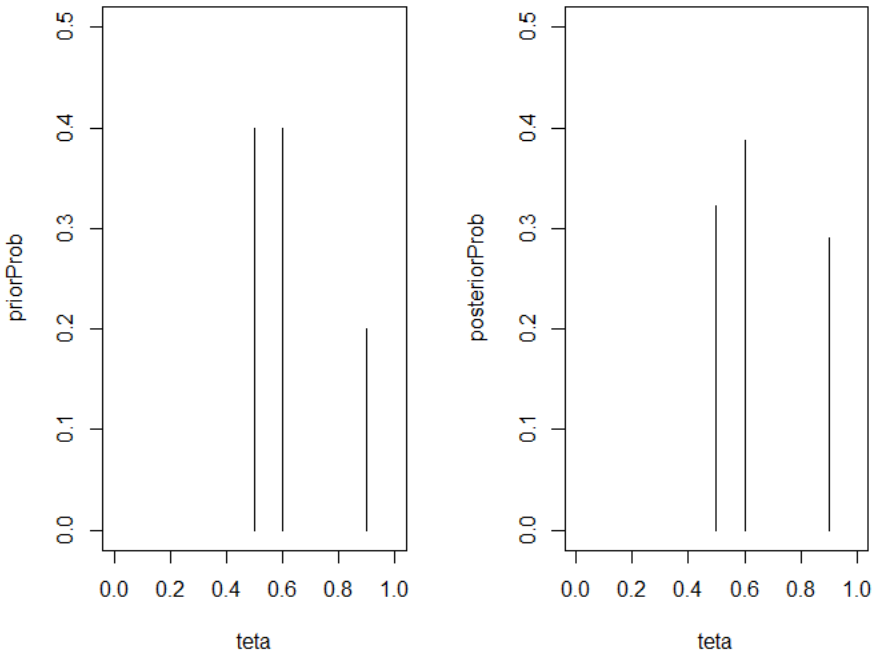
Earlier in the course we saw that it is convenient to use random variables and probability functions. To do this we had to assign values to events (head is 1 and tails is 0). We will do the same thing in the context of Bayesian updating. Our standard notations will be:

- θ is the value of the hypothesis.
- $p(\theta)$ is the prior probability function of the hypothesis.
- $p(\theta|D)$ is the posterior probability function of the hypothesis given the data.
- $p(D|\theta)$ is the likelihood function. (This is not a probability function!)

In Exercise 1 we can represent the three hypotheses A, B, and C by $\theta = 0.5,\ 0.6,\ 0.9$. 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 the prior and posterior probability functions.

Hypothesis	θ	Prior pf $p(\theta)$	Posterior pf $p(\theta\|x=1)$
A	0.5	$P(A) = P(0.5) = 0.4$	$P(A\|D) = p(0.5\|x=1) = 0.3226$
B	0.6	$P(B) = P(0.6) = 0.4$	$P(B\|D) = p(0.6\|x=1) = 0.3871$
C	0.9	$P(C) = P(0.9) = 0.2$	$P(C\|D) = p(0.9\|x=1) = 0.2903$

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



If the data was different then the likelihood column in the Bayesian update table would be different. We can plan for different data by building the entire likelihood table ahead of time. In the coin example there are two possibilities for the data: the toss is heads or the toss is tails. So the full *likelihood table* has two likelihood columns:

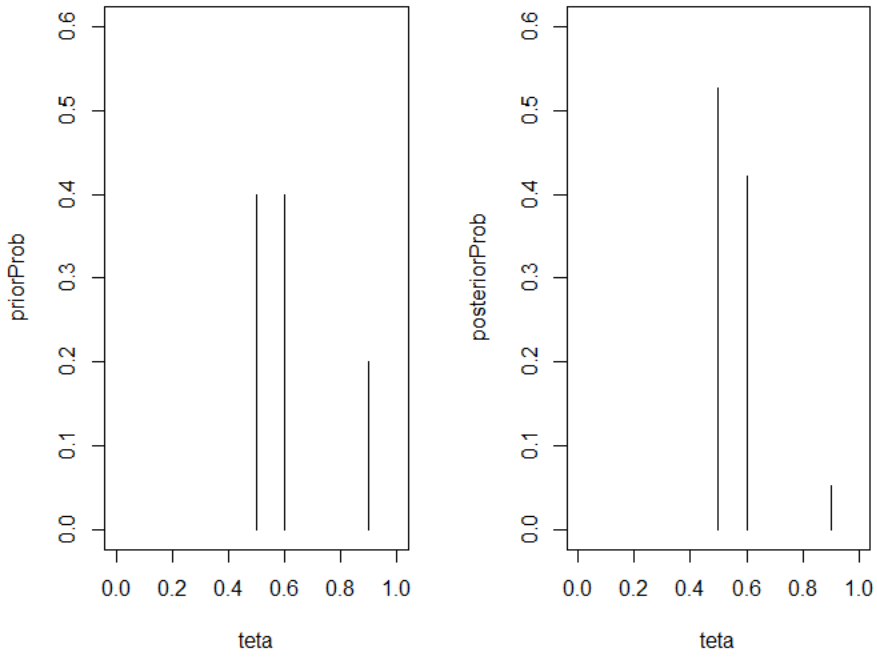
Hypothesis	likelihood	$p(x\|\theta)$
θ	$p(x=0\|\theta)$	$p(x=1\|\theta)$
0.5	0.5	0.5
0.6	0.4	0.6
0.9	0.1	0.9

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

Since the data has changed, the likelihood column in the Bayesian update table is now for $x=0$. That is, we must take the $p(x=0|\theta)$ column from the likelihood table.

Hypothesis	prior	likelihood	Bayes numerator	posterior
θ	$p(\theta)$	$p(x=0\|\theta)$	$p(x=0\|\theta)\,p(\theta)$	$p(\theta\|x=0)$
0.5	0.4	0.5	0.2	0.5263
0.6	0.4	0.4	0.16	0.4211
0.9	0.2	0.1	0.02	0.0526
Total	1		0.38	1

Now the probability of type A has increased from 0.4 to 0.5263, while the probability of type C has decreased from 0.2 to only 0.0526. Here are the corresponding plots:

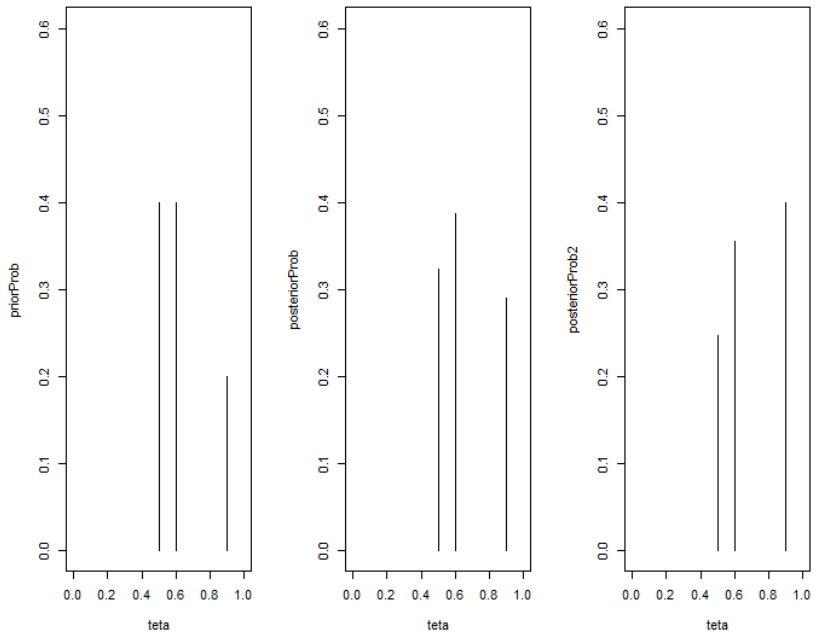


9.4 Updating again and again

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.

Exercise 3. Suppose you have picked a coin as in Exercise 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*?

Hypothesis θ	prior $p(\theta)$	likelihood 1 $p(x_1=1\|\theta)$	Bayes numerator 1 $p(x_1=1\|\theta)\,p(\theta)$	likelihood 2 $p(x_2=1\|\theta)$	Bayes numerator 2 $p(x_2=1\|\theta)\,p(x_1=1\|\theta)\,p(\theta)$	posterior 2 $p(\theta\|x_1=1,x_2=1)$
0.5	0.4	0.5	0.2	0.5	0.1	0.2463
0.6	0.4	0.6	0.24	0.6	0.144	0.3546
0.9	0.2	0.9	0.18	0.9	0.162	0.3990
Total	1				0.406	1



Note that the *second Bayes numerator* is computed by multiplying the first Bayes numerator and the second 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!

9.5 Prior and Posterior Predictive Probabilities

So far, 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.

There are many ways to word predictions:

1. Prediction: ”It will rain tomorrow."

2. Prediction using words of estimative probability: ”It is likely to rain tomor-row."

3. Probabilistic prediction: ”Tomorrow it will rain with probability 60%.”

We are going to focus on *probabilistic prediction* and precise quantitative statements. There are many places where we want to make a probabilistic prediction: Medical treatment outcomes, Weather forecasting, Climate change, Elections.

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

Recall the coin example: 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

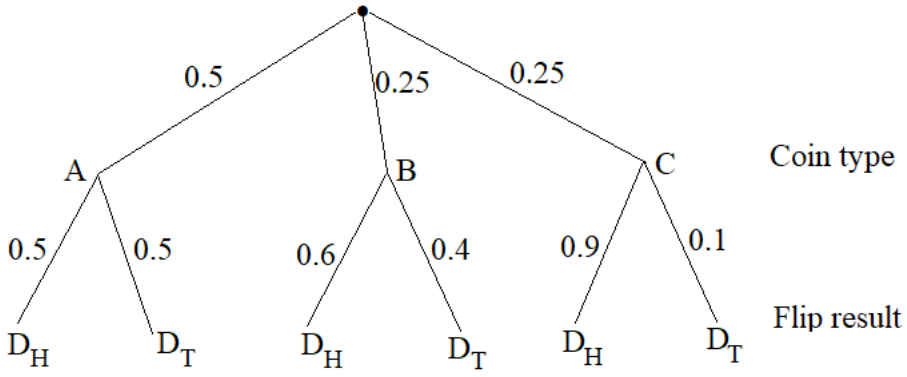
Type *B* coins have probability 0.6 of heads

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*.

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. By either drawing a tree or directly proceeding to the algebra, we get:



$$P(D_H) = P(D_H|A)P(A) + P(D_H|B)P(B) + P(D_H|C)P(C)$$

$$= 0.5 \times 0.5 + 0.6 \times 0.25 + 0.9 \times 0.25 = 0.625$$

$$P(D_T) = P(D_T|A)P(A) + P(D_T|B)P(B) + P(D_T|C)P(C)$$

$$= 0.5 \times 0.5 + 0.4 \times 0.25 + 0.1 \times 0.25 = 0.375$$

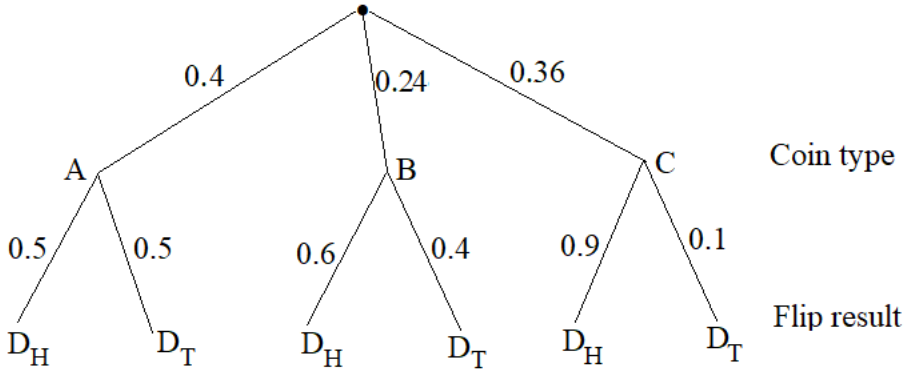
Definition: These probabilities give a (probabilistic) prediction of what will happen if the coin is tossed. Because they are computed before we collect any data they are called *prior predictive probabilities*.

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. 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 $P(A)$, $P(B)$, $P(C)$



$$P(D_H|D) = P(D_H|A)P(A|D) + P(D_H|B)P(B|D) + P(D_H|C)P(C|D)$$

$$= 0.5\times 0.4 + 0.6\times 0.24 + 0.9\times 0.36 = 0.668$$

$$P(D_T|D) = P(D_T|A)P(A|D) + P(D_T|B)P(B|D) + P(D_T|C)P(C|D)$$

$$= 0.5\times 0.4 + 0.4\times 0.24 + 0.1\times 0.36 = 0.332$$

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.

Summary:

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.