Bayesian Learning Example

Jesse Hoey
David R. Cheriton School of Computer Science
University of Waterloo,
Waterloo, Ontario, CANADA, N2L3G1
jhoey@cs.uwaterloo.ca

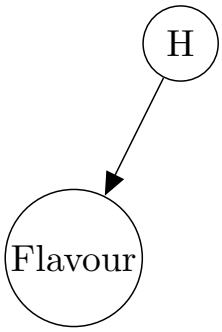
You have a bag of candy with 2 flavors (Lime, Cherry), and you know these candies are sold in bags with different ratios $h_i, i \in \{1 \dots 5\}$ as follows:

- h_1*: 100% cherry*
- h_2*: 75% cherry+25% lime*
- h_3*: 50% cherry + 50% lime*
- h_4*: 25% cherry + 75% lime*
- h_5 *100% lime*

By drawing a random sample of candies from the bag, what ratio is in the bag? What is the probability the next candy is a lime?

The Bayesian network for this example is very simple:

But in this case, we don't know the probability distribution $P(Flavour)$. However, we do know it is one of the five hypotheses, $h_i, i \in \{1 \dots 5\}$. If we represent these hypotheses as a discrete random variable H with 5 values, the BN is actually as follows:



Our goal is to answer the two questions *"What ratio is in the bag?"* and *"Is the next candy going to be a lime?"*, based on our data $\mathbf{d} = \{d_1, d_2, d_3, \ldots, d_N\}$, where $d_j \in \{lime, cherry\}$. The most complete answer is Bayesian estimation, but we can also look at two approximations: Maximum a-posteriori (MAP) estimation, or maximum likelihood (ML) estimation. We assume the bags are infinitely large, or that we put each candy back and shuffle the bag after drawing the candy. Otherwise, the ratios would all change (and hence so would our hypotheses) after each draw. Here is how these three methods compute answers to the two questions:

- **Bayesian**:
  - *What ratio is in the bag?*: The Bayesian answer does not commit to one ratio, but instead is a belief about what ratio is in the bag, as given by the distribution $P(H|\mathbf{d})$:

$$\begin{aligned} P(H|\mathbf{d}) &= \frac{P(\mathbf{d}|H)P(H)}{\sum_H P(\mathbf{d}|H)P(H)} \\ &\propto P(\mathbf{d}|H)P(H) \end{aligned} \tag{1}$$

    The Bayesian solution is to compute $P(H|\mathbf{d})$ exactly, but note that this requires the computation of the normalization factor $\sum_H P(\mathbf{d}|H)P(H)$ which may be intractable if the model space is large.
  - *Is the next candy going to be a lime?*: This is taken as an expectation over the distribution over hypotheses:

$$\begin{aligned} P(d_{N+1} = lime|\mathbf{d}) &= \sum_i P(d_{N+1} = lime, h_i|\mathbf{d}) \\ &= \sum_i P(d_{N+1} = lime|h_i)P(h_i|\mathbf{d}) \end{aligned} \tag{2}$$

    where $P(h_i|\mathbf{d}) = P(H = h_i|\mathbf{d})$ is given by Equation (1).

- **MAP**:
  - *What ratio is in the bag?*: The most likely posterior hypothesis, or

$$h^* = \arg\max_{h_i} P(h_i|\mathbf{d}), \tag{3}$$

    found by maximizing over Equation (1). However, this is easier because the maximization can be done over the numerator alone, and the normalization factor does not need to be computed.
  - *Is the next candy going to be a lime?*: The probability is now based on the maximum a posteriori hypothesis:

$$P(d_{N+1} = lime|\mathbf{d}) = P(d_{N+1} = lime|h^*). \tag{4}$$

- **ML**:
  - *What ratio is in the bag?*: The hypothesis that best explains the data, or

$$h^\dagger = \arg\max_{h_i} P(\mathbf{d}|h_i). \tag{5}$$

– *Is the next candy going to be a lime?*: The probability is now based on which hypothesis makes the data most likely:

$$P(d_{N+1} = lime|\mathbf{d}) = P(d_{N+1} = lime|h^{\dagger}). \tag{6}$$

Let us suppose we have a prior distribution over H which is $P(H) = \{0.1, 0.2, 0.4, 0.2, 0.1\}$ for $H = \{h_1, h_2, h_3, h_4, h_5\}$, respectively. That is, we think its less likely the distribution is skewed towards more cherries or limes. Then, after we draw $N = 1$ candy and find $\mathbf{d} = \{d_1 = lime\}$, we have:

	$P(\mathbf{d}\|H)$	$P(H)$	$P(\mathbf{d}\|H)P(H) \propto P(H\|\mathbf{d})$	$P(H\|\mathbf{d})$
h_1	0.0	0.1	0.0	0.0
h_2	0.25	0.2	0.05	0.1
h_3	0.5	0.4	0.2	0.4
h_4	0.75	0.2	0.15	0.3
h_5	1.0	0.1	0.1	0.2

Where the last column is obtained from the second-to-last column by dividing out the sum of all elements in the second-to-last column (0.5).
We can answer the questions using Bayesian estimation, Maximum a-posteriori (MAP) estimation, or maximum likelihood (ML) estimation.

- **Bayesian**:
  - *What ratio is in the bag?*: Our belief about what ratio is in the bag is given by the distribution $P(H|\mathbf{d})$ shown in the table above. h_3 has the same probability as our prior belief (0.4), but the probability of $H = h_4$ and $H = h_5$ have gone up, whereas the probability of $H = h_1$ and $H = h_2$ have gone down.
  - *Is the next candy going to be a lime?*: The probability the next candy is a lime is

$$\begin{aligned} P(d_2 = lime|d_1 = lime) &= \sum_i P(d_2 = lime, h_i|d_1 = lime) \\ &= \sum_i P(d_2 = lime|h_i)P(h_i|d_1 = lime) \end{aligned}$$

    which is $0 + 0.25 * 0.1 + 0.5 * 0.4 + 0.75 * 0.3 + 1 * 0.2 = 0.65$.

- **MAP**:
  - *What ratio is in the bag?*: The most likey in the posterior is h_3, which says its 50% lime/50% cherry
  - *Is the next candy going to be a lime?*: The probability is now based on the maximum a posteriori hypothesis h_3: $P(d_2 = lime|d_1 = lime) = P(d_2 = lime|h^*)$ where $h^* = \arg\max_{h_i} P(h_i|d_1 = lime)$, so it is 0.5.

- **ML**:
  - *What ratio is in the bag?*: The most likely in the likelihood is h_5, which says its 100% lime.

– *Is the next candy going to be a lime?*: The probability is now based on which hypothesis makes the data most likely, which is h_5: $P(d_2 = lime|d_1 = lime) = P(d_2 = lime|h^\dagger)$ where $h^\dagger = \arg\max_{h_i} P(d_1 = lime|h_i)$, so it is 1.0. Notice how this is massively *overfit*: the ML estimate is totally committed to the next candy being lime.

Let's draw another 2 limes, so we have $\mathbf{d} = \{lime, lime, lime\}$, and if we start from $P(H)$, we get:

	$P(\mathbf{d}\|H)$	$P(H)$	$\propto P(H\|\mathbf{d})$	$P(H\|\mathbf{d})$
h_1	$0^3 = 0$	0.1	0	0
h_2	$0.25^3 = 0.015625$	0.2	0.003125	0.013
h_3	$0.5^3 = 0.125$	0.4	0.05	0.211
h_4	$0.75^3 = 0.4219$	0.2	0.0844	0.356
h_5	$1^3 = 1$	0.1	0.1	0.421

If, instead we proceed incrementally, and start from $P(H|d_1 = lime)$, which is what we computed in the table on the previous page, and multiply by the likelihood of observing two more limes, we get the same answer:

	$P(d_2 = lime, d_3 = lime\|H)$	$P(H\|d_1 = lime)$	$\propto P(H\|\mathbf{d})$	$P(H\|\mathbf{d})$
h_1	$0^2 = 0$	0.0	0	0
h_2	$0.25^2 = 0.0625$	0.1	0.00625	0.013
h_3	$0.5^2 = 0.25$	0.4	0.1	0.211
h_4	$0.75^2 = 0.5625$	0.3	0.1688	0.356
h_5	$1^2 = 1$	0.2	0.2	0.421

We can answer the questions again with our three methods.

- **Bayesian**:
  - *What ratio is in the bag?*: Our belief about what ratio is in the bag is given by the distribution $P(H|\mathbf{d})$ shown in the table above. h_3 has a probability of 0.21, and h_5 has a probability of 0.42.
  - *Is the next candy going to be a lime?*: The probability the next candy is a lime is $0 + 0.25 * 0.01 + 0.5 * 0.21 + 0.75 * 0.35 + 1.0 * 0.42 = 0.79$.
- **MAP**:
  - *What ratio is in the bag?*: The most likey in the posterior is h_5, which says its 100% lime
  - *Is the next candy going to be a lime?*: based on h_5 its 1.0
- **ML**:
  - *What ratio is in the bag?*: The most likely in the likelihood is h_5, which says its 100% lime.
  - *Is the next candy going to be a lime?*: again, it is 1.0.

We now draw a fourth candy and it is a cherry, so we have $\mathbf{d} = \{lime, lime, lime, cherry\}$. Starting again from the original prior, we get:

	$P(\mathbf{d}\|H)$	$P(H)$	$\propto P(H\|\mathbf{d})$	$P(H\|\mathbf{d})$
h_1	$0^3 * 1 = 0$	0.1	0	0
h_2	$0.25^3 * 0.75 = 0.0117$	0.2	0.0023	0.05
h_3	$0.5^3 * 0.5 = 0.0625$	0.4	0.025	0.51
h_4	$0.75^3 * 0.25 = 0.105$	0.2	0.021	0.44
h_5	$1^3 * 0 = 0$	0.1	0	0

Or, doing it incrementally, we get the same:

	$P(d_4 = cherry\|H)$	$P(H\|d_1 = d_2 = d_3 = lime)$	$\propto P(H\|\mathbf{d})$	$P(H\|\mathbf{d})$
h_1	1	0	0	0
h_2	0.75	0.013	0.00975	0.05
h_3	0.5	0.211	0.1055	0.51
h_4	0.25	0.356	0.089	0.44
h_5	0	0.421	0	0

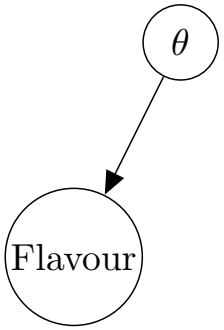
Question answers:

- **Bayesian**:
  - *What ratio is in the bag?*: $P(H|\mathbf{d})$ shown in the table above. h_3 has a probability of 0.51, and h_5 has a probability of 0.
  - *Is the next candy going to be a lime?*: The probability the next candy is a lime is $0 + 0.25 * 0.05 + 0.5 * 0.51 + 0.75 * 0.44 + 1.0 * 0.0 = 0.5975$.
- **MAP**:
  - *What ratio is in the bag?*: The most likey in the posterior is h_3, which says its 50% lime
  - *Is the next candy going to be a lime?*: based on h_3 its 0.5
- **ML**:
  - *What ratio is in the bag?*: The most likely in the likelihood is h_4, which says its 75% lime.
  - *Is the next candy going to be a lime?*: again, it is 0.75.

Now let us consider the more general problem:

> *You have a bag of candy with 2 flavors (Lime, Cherry). You don't know what proportion of limes are in the bag. By drawing a random sample of candies from the bag, what ratio of limes to cherries is in the bag? What is the probability the next candy is a lime?*

The Bayesian network for this example is the same, however, we now replace the hypothesis space with a continuous parameter θ giving the probability that the flavour of a candy in the bag is *cherry*. Thus, our hypothesis space is H_θ which says that $P(Flavour = cherry) = \theta$ and $P(Flavour = lime) = (1 - \theta)$.



Now suppose we have drawn $\mathbf{d} = \{d_1, \ldots, d_c, d_{c+1}, \ldots, d_N\}$ such that $d_i = cherry$ if $i \leq c$ and $d_i = lime$ if $i > c$, such that c is the number of cherries in $\mathbf{d}$ and we let $l = N - c$ be the number of limes in $\mathbf{d}$. As before the data are idependently and identically distributed ("i.i.d" so the bag is infinitely large or we replace the candies after each draw). The likelihood of drawing c cherry and l lime candies is

$$P(\mathbf{d}|H_\theta) = \theta^c(1-\theta)^l$$

We will assume a β distribution as a prior, such that

$$P(H_\theta) = \beta(\theta, a, b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

which you can think of as "imagining" you'd already seen a cherries and b limes. Therefore, the posterior distribution is

$$P(H_\theta|\mathbf{d}) \propto P(\mathbf{d}|H_\theta)P(H_\theta) = \theta^{c+a-1}(1-\theta)^{l+b-1}$$

Now we can proceed to answer our two questions using our three methods:

- **Bayesian**:
  - *What ratio is in the bag?*: The Bayesian solution is to compute $P(H_\theta|\mathbf{d})$ exactly, but this requires the computation of the normalization factor:

$$\begin{aligned} P(H_\theta|\mathbf{d}) &= \frac{P(\mathbf{d}|H_\theta)P(H_\theta)}{\int_\theta P(\mathbf{d}|\theta)P(\theta)} \\ &= \frac{\theta^{c+a-1}(1-\theta)^{l+b-1}}{\int_\theta \theta^{c+a-1}(1-\theta)^{l+b-1}d\theta} \end{aligned} \tag{7}$$

relabing $s = c + a - 1$ and $t = l + b - 1$, we can compute (using integration by parts, see Appendix)

$$\int_\theta \theta^s (1-\theta)^t d\theta = \frac{s!t!}{(s+t+1)!},$$

(which is the normalization constant from the β distribution that we ignored), so that

$$P(H_\theta|\mathbf{d}) = \frac{(c+a+l+b-1)!}{(c+a-1)!(l+b-1)!}\theta^{c+a-1}(1-\theta)^{l+b-1} \tag{8}$$

– *Is the next candy going to be a lime?*: This is taken as an expectation over the distribution over hypotheses:

$$\begin{aligned} P(d_{N+1} = lime|\mathbf{d}) &= \int_\theta P(d_{N+1} = lime, \theta|\mathbf{d}) \\ &= \int_\theta P(d_{N+1} = lime|\theta) P(\theta|\mathbf{d}) \end{aligned} \tag{9}$$

where $P(d_{N+1} = lime|\theta) = (1-\theta)$ and $P(\theta|\mathbf{d}) = P(H_\theta = \theta|\mathbf{d})$ is given by Equation (8), giving

$$\begin{aligned} &= \frac{(c+a+l+b-1)!}{(c+a-1)!(l+b-1)!}\int_\theta \theta^{c+a-1}(1-\theta)^{l+b} \\ &= \frac{(c+a+l+b-1)!}{(c+a-1)!(l+b-1)!}\frac{(c+a-1)!(l+b)!}{(c+a+l+b)!} \\ &= \frac{l+b}{c+a+l+b} \end{aligned} \tag{10}$$

And we see that the Bayesian prediction is the fraction of limes seen so far, but imagining that we'd seen a cherries and b limes (the prior) before observing the actual data of c cherries and l limes. The reason this works out so nicely, of course, is that the β function is the conjugate prior of the binomial distribution.

- **MAP**:

  – *What ratio is in the bag?*: The most likely posterior hypothesis, or

$$\theta^* = \arg\max_\theta P(\theta|\mathbf{d}), \tag{11}$$

found by maximizing over Equation (7), which we do by taking a derivative of the log and setting to zero (we can ignore the denominator in Equation 7 as it is independent of θ.

$$\begin{aligned} 0 = \frac{d\log P(\theta|\mathbf{d})}{d\theta} &= \frac{d}{d\theta}\left[(c+a-1)\log\theta + (l+b-1)\log(1-\theta)\right] \\ 0 &= \frac{(c+a-1)}{\theta^*} - \frac{(l+b-1)}{1-\theta^*} \\ \theta^*(l+b-1) &= (1-\theta^*)(c+a-1) \\ \theta^* &= \frac{c+a-1}{c+a+l+b-2} \end{aligned}$$

– *Is the next candy going to be a lime?*: The probability is now based on the maximum a posteriori hypothesis, θ^*, and is therefore $1 - \frac{c+a-1}{c+a+l+b-2} = \frac{l+b-1}{c+a+l+b-2}$.

- **ML**:

  – *What ratio is in the bag?*: The hypothesis that best explains the data, or

$$\theta^\dagger = \arg\max_\theta P(\mathbf{d}|\theta), \tag{12}$$

  found by maximizing the liklihood function only

$$\begin{aligned} 0 = \frac{d \log P(\mathbf{d}|\theta)}{d\theta} &= \frac{d}{d\theta}\left[c \log\theta + l \log(1-\theta)\right] \\ 0 &= \frac{c}{\theta^\dagger} - \frac{l}{1-\theta^\dagger} \\ \theta^\dagger &= \frac{c}{c+l} \end{aligned}$$

  – *Is the next candy going to be a lime?*: The probability is now based on the hypothesis that makes the data most likely, $\theta^\dagger$, and is therefore $1 - \frac{c}{c+l} = \frac{l}{c+l}$, or the fraction of limes drawn so far.

Appendix: derivation of

$$\int_\theta \theta^s (1-\theta)^t d\theta = \frac{s!t!}{(s+t+1)!},$$

$$\begin{aligned} \int_\theta \theta^s (1-\theta)^t d\theta &= \theta^s \int_\theta (1-\theta)^t d\theta - \int_\theta s\theta^{s-1} \left(\int_\theta (1-\theta)^t d\theta\right) d\theta \\ &= \left[\theta^s \frac{-(1-\theta)^{t-1}}{(t+1)}\right]_0^1 + \int_\theta s\theta^{s-1} \frac{(1-\theta)^{t-1}}{(t+1)} d\theta \\ &= \left[\frac{s}{t+1}\right] \int_\theta \theta^{s-1}(1-\theta)^{t+1} d\theta \\ &= \left[\frac{s}{t+1}\right]\left[\frac{s-1}{t+2}\right] \int_\theta \theta^{s-2}(1-\theta)^{t+2} d\theta \\ &= \cdots \\ &= \left[\frac{s(s-1)(s-2)\ldots(s-(s-1))}{(t+1)(t+2)\ldots(t+s)}\right] \int_\theta (1-\theta)^{t+s} d\theta \\ &= \frac{s(s-1)(s-2)\ldots(s-(s-1))}{(t+1)(t+2)\ldots(t+s)} \left[\frac{-(1-\theta)^{t+s+1}}{(t+s+1)}\right]_0^1 \\ &= \frac{s(s-1)(s-2)\ldots 1}{(t+1)(t+2)\ldots(t+s)(t+s+1)} \\ &= \frac{s!}{\frac{(t+s+1)!}{t!}} \\ &= \frac{s!t!}{(s+t+1)!} \end{aligned}$$