Computational Cognitive Science (2018–2019)

School of Informatics, University of Edinburgh
Original exercises by Frank Keller, with modifications by Chris Lucas

Solutions for Tutorial 3: Bayesian Estimation; Concept Learning

Please work through this tutorial sheet on your own time as much as possible before arriving in tutorial. We encourage you to work together and discuss your methods and solutions.

1 Exercises in Bayesian Estimation

Question 1: I have three fair dice in a jar: one with 6 sides, one with 8 sides, and one with 12 sides. I select one at random from the jar, roll it once, and tell you that the number is greater than four.

1. If you had to place a bet on which die I picked, which would you bet on? Why?
2. If I roll the **same** die again, what's the probability that the outcome will be greater than four?

Solution 1:

1. This problem is easiest to solve if the different dice are considered as different probability distributions over two outcomes (greater than four, not greater than four) rather than even probability distributions over multiple outcomes. Let $H \in \{d6, d8, d12\}$ be the die, and $O \in \{5+, 4-\}$ be the observation. Then we have $P(O = 5+|H = d6) = \frac{2}{6}$, $P(O = 5+|H = d8) = \frac{4}{8}$, and $P(O = 5+|H = d12) = \frac{8}{12}$. The Maximum Likelihood Estimate is then $d12$, since $\frac{8}{12} > \frac{4}{8} > \frac{2}{6}$.

2. We don't know which die we are dealing with, so our best estimate of the probability of an outcome greater than four should be based on the mean of all three hypotheses (Bayesian average). In addition, we need to realize that the old posterior $P(H = h|O_{old} = 5+)$ functions as the prior for the new posterior, $P(O_{new} = 5+|O_{old} = 5+)$. The new posterior is therefore:

$$P(O_{new} = 5+|O_{old} = 5+) = \sum_{h \in H} P(O_{new} = 5+|H = h)P(H = h|O_{old} = 5+) \tag{1}$$

First, let's compute the $P(H = h|O_{old} = 5+)$ terms for each value of H using Bayes' rule:

$$\begin{aligned} P(H = d6|O_{old} = 5+) &= \frac{P(O_{old} = 5+|H = d6)P(H = d6)}{P(O_{old} = 5+)} \\ &= \frac{P(O_{old} = 5+|H = d6)P(H = d6)}{\sum_{h \in H} P(O_{old} = 5+|H = h)P(H = h)} \\ &= \frac{(1/3)(1/3)}{(1/3)(1/3) + (1/2)(1/3) + (2/3)(1/3)} \\ &= \frac{1/9}{1/2} \\ &= 2/9 \end{aligned} \tag{2}$$

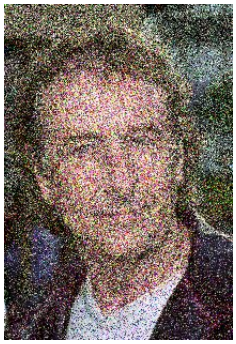
Similarly, $P(H = d8|O_{old} = 5+) = 1/3$ and $P(H = d12|O_{old} = 5+) = 4/9$. (Notice we can reuse the value of $P(O_{old} = 5+)$ in the denominator an do not need to recompute it each time).

We can now plug these values into Equation 1 to compute the Bayesian estimate:

$$
\begin{aligned}
P(O_{new} = 5+|O_{old} = 5+) &= \sum_{h \in H} P(O_{new} = 5+|H = h)P(H = h|O_{old}) \qquad (3)\\
&= P(O_{new} = 5+|H = d6)P(H = d6|O_{old})\\
&\quad + P(O_{new} = 5+|H = d8)P(H = d8|O_{old})\\
&\quad + P(O_{new} = 5+|H = d12)P(H = d12|O_{old})\\
&= (1/3)(2/9) + (1/2)(1/3) + (2/3)(4/9)\\
&= 29/54\\
&\approx 0.537
\end{aligned}
$$

Question 2: In an experiment on face recognition, subjects are presented with images of people they know, and asked to identify them. The images are presented for a very short period of time so that subjects may not have time to see the details of the entire face, but are likely to get a general impression of things like hair color and style, overall shape, skin color, etc. In this question we will consider how to formulate the face recognition problem as a probabilistic inference model.

1. What is the hypothesis space in this problem? Is it continuous or discrete? Finite or infinite?
2. What constitutes the observed data y and what kinds of values can it take on?
3. Write down an equation that expresses the inference problem that the subjects must solve to identify each face. Describe what each term in the equation represents.
4. What factors might influence the prior in this situation?
5. Suppose one group of subjects sees clear images, such as the one on the left below, and another group sees noisy images, such as the one on the right below. Which term(s) in your equation will be different for the noisy group compared to the clear group?

6. What does the model predict about subjects' performance with noisy images compared to clear images? Rather than working with the full scenario above, you can simplify by supposing the experiment has images of only two people, $h_1 =$ Eric Idle and $h_2 =$ John Cleese. How does image noisiness affect the model's probability of inferring that the image is of Eric Idle, rather than John Cleese? (Hint: since you are considering only two hypotheses, think about the *posterior odds*. How will this quantity differ between the case where the image is noisy and the case where it is clear?) Do you need to make any further assumptions in order to make predictions about subject behavior?

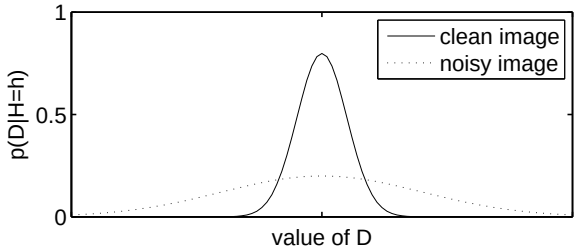
Solution 2:

1. The hypothesis space is the set of different people that the subject knows, a finite (though very large) discrete space.
2. The data is the image the subject sees, or more precisely, what the subject actually perceives. It's difficult to say precisely what that might be without knowing more about the low-level features used by the visual system, but it could be things like the color and intensity in different regions of the image. In this case the values of the observed data would be continuous. However, if we were actually going to model this problem, we might want to simplify by assuming that higher-level discrete features are directly observed, e.g. face shape, hair style, skin color. But note that not all of these higher-level features would necessarily be observed for each trial. (We could also make an intermediate assumption, using a discretized space of color/intensity features, or maybe intermediate-level features like edges).
3. $$P(H|y) = \frac{P(y|H)P(H)}{P(y)} \tag{4}$$

   where the $P(H)$ is the subject's prior belief that any particular person will appear in the photo, $P(H|y)$ is the subject's posterior belief that any particular person is in the image, given what the subject sees in the image, $P(y|H)$ (likelihood) is the probability that a particular set of features will be perceived given that a particular person is shown in the image, and $P(y)$ is the overall probability of perceiving a particular set of features.
4. The prior could be influenced by factors such as the frequency or recency with which the subject has seen each of the people outside of the experimental situation, and the frequency of the particular person's face within the experimental situation (if images are reused). One could imagine other possible factors such as the emotional closeness of the subject to the person, or the frequency with which the subject has seen photographs of the person (as opposed to the live person).
5. The prior will be the same. The likelihood will be different, since the manipulation changes how the images look – there will be a different distribution over observed features given the same person being shown. $P(y)$ will be different since there is a different overall distribution of what the images look like. And since $P(y)$ and $P(y|H)$ are different, the posterior will also be different.

   In preparation for the next question it helps if we are more specific about the changes we expect in the likelihood. Consider the distribution $P(y|h)$ for a specific hypothesis h (say, Eric Idle). If the images are clear, then we would expect relatively little variation in the features we perceive when shown his image. That is, a relatively small number of possible values for y will have high probability, and other possible values will have low probability. However, if the images are noisy, this effectively spreads out the probability mass over a larger number of possible values for y: we are more likely to see features that are further from those in the original image, but less likely to see features that are exactly those in the original image.

   The above description assumes discrete values for y, but we can also get an intuition for what's going on by imagining that the different possible values for y are continuous values along a 1-dimensional space, and then plotting $P(y|h)$ against y:



The mean of these curves represents the "average" data that would be seen when Eric's picture is shown. The curves for the noisy and clean cases have the same mean, but the distribution of observations in the noisy case is broader than that in the clean case. [I'm assuming that the noise itself is unbiased, otherwise the mean could change also, but this would needlessly complicate the analysis.]

6. The model predicts that subject will have a harder time discriminating Idle from Cleese in the noisy scenario. To see why, consider the posterior odds in each scenario. For a particular observation d the posterior odds can be written as

$$\frac{P(H=h_1|y)}{P(H=h_2|y)} = \frac{P(y|H=h_1)P(H=h_1)}{P(y|H=h_2)P(H=h_2)}$$

and represents how much more favored h_1 is than h_2 given y.

What can we say about the posterior odds in the noisy and clean situations? Well, first of all, any differences between the two situations will come from the *likelihood ratio* $P(y|H=h_1)/P(y|H=h_2)$ because the priors (and prior odds $P(H=h_1)/P(H=h_2)$) are the same in both cases. So let's focus on the likelihood ratio, using our analysis of the likelihood from the previous question part. According to our graphs, $P(y|H=h_1)$ and $P(y|H=h_2)$ will have similar shapes, except that the means will be different. In the clean case, when we see an image of Idle, the actual observation d is very likely to be close to the mean of $P(y|H=h_1)$, and not close to the mean of $P(y|H=h_2)$. Therefore because $P(y|H=h_1)$ and $P(y|H=h_2)$ are highly peaked, $P(y|H=h_1)$ will be very high while $P(y|H=h_2)$ will be very low and the likelihood ratio will strongly favor h_1. On the other hand in the noisy case, the observation d' is less likely to be near the mean of $P(y|H=h_1)$, and in addition observations that are far from the mean of $P(y|H=h_2)$ have higher probability than in the clean case. So overall the likelihood ratio will not favor h_1 as strongly.

Since the likelihood ratio favors h_1 more strongly in the clean case, so do the posterior odds. If we assume that subjects are more likely to give the name of the person corresponding to the higher probability hypothesis, then they would be more accurate in the clean scenario. If we wanted to make any more quantitative predictions, we would need to be more specific about the relationship between probability and subject's responses (e.g. do their responses exactly match the posterior distribution, is there some non-linear relationship between the two, etc.)

2 A Bayesian Model of Concept Learning

Question 3: In this question we will consider a simplified version of the Tenenbaum, 2000 model of generalization where there are only four hypotheses under consideration, each of which has equal prior probability:

$$h_1 = \{\text{odd numbers}\},\ h_2 = \{\text{even numbers}\},\ h_3 = \{\text{multiples of 5}\},\ h_4 = \{\text{multiples of 10}\}$$

Given the set of examples $X = \{10, 40\}$ from the target concept, first compute the posterior probability of each hypothesis under the model. Then, determine the model's predicted probability that each of the following new data points is also part of the same concept: 2, 3, 5, and 20. Now explain, with reference to the terms in the model, why the model's predictions tend to become more 'rule-like' as more examples are seen (e.g., if 50 is shown as a third example from the concept).

Solution 3: For i =1-4, we have

$$\begin{aligned} P(h_i|X) &= \frac{P(X|h_i)P(h_i)}{P(X)} \\ &= \frac{P(X|h_i)P(h_i)}{\sum_j P(X|h_j)P(h_j)} \end{aligned}$$

Note that the denominator is the same for any i, and is the sum of the numerators. So we can first compute each numerator:

$$\begin{aligned} P(h_1|X) &\propto P(X|h_1)P(h_1) \\ &= 0 \cdot \frac{1}{4} \\ &= 0 \end{aligned}$$

$$\begin{aligned} P(h_2|X) &\propto P(X|h_2)P(h_2) \\ &= (\frac{1}{50})^2(\frac{1}{4}) \\ &= .0001 \end{aligned}$$

And similarly, $P(h_3|X) \propto (\frac{1}{20})^2(\frac{1}{4}) = .000625$ and $P(h_4|X) \propto (\frac{1}{10})^2(\frac{1}{4}) = .0025$. Thus, after normalizing, we have

$$\begin{aligned} P(h_1|X) &= 0 \\ P(h_2|X) &\approx .031 \\ P(h_3|X) &\approx .194 \\ P(h_4|X) &\approx .775 \end{aligned}$$

To compute the probability of a new point y being in the concept, we need to find

$$P(y \in C|X) = \sum_{h_i} P(y \in C|C = h_i)P(C = h_i|X). \quad (5)$$

For $y = 3$, $P(y \in C|C = h_i) = 0$ for i =2-4, and $P(C = h_1|X) = 0$, so $P(y \in C|X) = 0$. For $y = 2$, $P(y \in C|C = h_i) = 0$ except for $i = 2$, so only one term in the sum in Equation (5) counts:

$$\begin{aligned} P(y \in C|X) &= P(y \in C|C = h_2)P(C = h_2|X) \\ &= 1 \cdot .031 \\ &= .031 \end{aligned}$$

Similarly, only h_3 is relevant for $y = 5$, and we get $P(y \in C|X) = .194$.
For $y = 20$, we get $P(y \in C|X) = 1$.

The model's behavior becomes more 'rule-like' as more examples are seen because of the form of the likelihood term, $P(X|h_i) = \frac{1}{|h_i|^n}$, where n is the number of examples seen. In particular, the model will favor more and more the smallest hypothesis consistent with all examples seen—the one that follows the same 'rule'. This is because the likelihood of larger hypotheses decreases a lot more than the likelihood of smaller hypothesis with every additional example (because of $|h_i|$ in the denominator). Eventually the smallest hypothesis dominates all others.