Informatics 1 Cognitive Science – Tutorial 4 Solutions

Frank Keller, Carina Silberer, Frank Mollica

Week 5

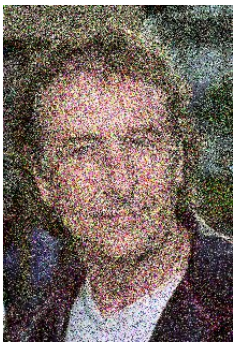
1 Bayesian Modeling

General guidance: The key concept to get across in this exercise is how to infer $P(H|d)$ from $P(d|H)$. You will probably remind students of the intuition behind that inference, and of what conditional probabilities mean. The rest of the exercise is supposed to take students through the whole process of Bayesian modeling: defining a hypothesis space, choosing a prior, defining and computing the likelihood term, etc. There is lots of room for discussion, esp. for subquestions 4 to 6.

Last week, we discussed Bayesian Modeling as a way of capturing human reasoning and decision making. In this exercise, we will look at an example of how we can formalize a cognitive process in Bayesian terms.

Exercise 1 In an experiment on face recognition, participants 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 participants 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 D and what kinds of values can it take on?
3. Write down an equation that expresses the inference problem that the participants 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 participants 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 participants' performance with noisy images compared to clear images?

Solution for Exercise 1

1. The hypothesis space is the set of different people that the participant knows, a finite (though very large) discrete space.

2. The data is the image the participant sees, or more precisely, what the participant 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. $$P(H|d) = \frac{P(d|H)P(H)}{P(d)} \tag{1}$$

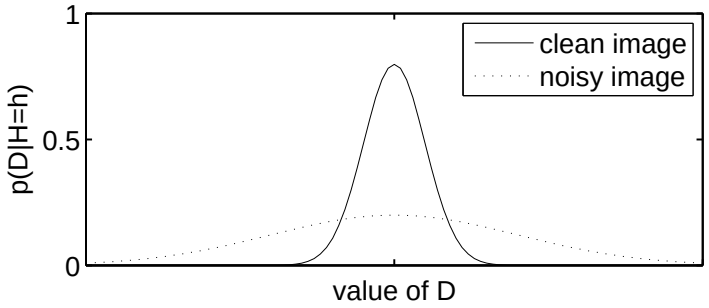
   where the $P(H)$ is the participant's prior belief that any particular person will appear in the photo, $P(H|d)$ is the participant's posterior belief that any particular person is in the image, given what the participant sees in the image, $P(d|H)$ (likelihood) is the probability that a particular set of features will be perceived given that a particular person is shown in the image, and $P(d)$ 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 participant 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 participant to the person, or the frequency with which the participant 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. $P(d)$ will be different since there is a different overall distribution of what the images look like. And since $P(d)$ and $P(d|H)$ 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 $P(d|h)$ 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 d, but we can also get an intuition for what's going on by imagining that the different possible values for d are continuous values along a 1-dimensional space, and then plotting $P(d|h)$ against d:



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 participants will have a harder time discriminating Idle from Cleese in the noisy scenario. This is because the likelihood distribution is more spread out when there is noise (see previous question), which means that the different hypotheses are closer together in terms of their likelihood (see figure from previous question). As a consequence, the posterior probabilities of the different hypotheses will also be more similar (assuming the prior distribution doesn't change when we move from the regular to the noisy scenario). If the posteriors are close together, then the different hypotheses will be harder to tell apart.

2 Word Learning

General guidance: This is a worked example of a Bayesian model which was discussed in the lectures. The main points are: (i) students should understand the mechanics of the model, i.e., how to compute the prior and the likelihood (though note that we're assuming a different prior from in the lectures!); (ii) they should realize that these models (and the modeling results you get) heavily rely on certain assumptions. In particular using the hypothesis length prior we assume here, we are not able to model the data correctly.

In the lectures, we discussed Piantadosi et al.'s program induction model for learning number words. The model considers several possible hypotheses for the definition of number words and uses Bayes Rule to infer the most likely hypothesis as a learner sees data. Recall Bayes rule:

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_i P(d|h_i)P(h_i)}. \tag{2}$$

Exercise 2 and solution We will attempt to replicate Piantadosi et al.'s results. Instead of searching a vast hypothesis space at multiple data amounts, let's focus on the four hypotheses in Figure 1 and the two datasets in Table 1.

Dataset	(one, $\cdot$)	(two, :)	(three, :$\cdot$)	(four, ::)	(five, ::$\cdot$)	(six, :::)	(seven, :::$\cdot$)	(eight, ::::)
($N = 30$)	20	4	2	2	0	1	0	1
($N = 60$)	41	10	5	2	0	1	0	1

Table 1: The datasets we are considering. Each column denotes a type of data. Each row contains the number of times that type of data has been seen.

1. In the original model, they used a simplicity prior. Let's try using a prior based on the length of the hypotheses. For each hypothesis, write down its length (include parentheses and punctuation).

One-knower

```
λ S . (if (singleton? S)
          "one"
          undef)
```

Two-knower

```
λ S . (if (singleton? S)
          "one"
          (if (doubleton? S)
              "two"
              undef))
```

Three-knower

```
λ S . (if (singleton? S)
          "one"
          (if (doubleton? S)
              "two"
              (if (tripleton? S)
                  "three"
                  undef)))
```

CP-knower

```
λ S . (if (singleton? S)
          "one"
          (next (L (set-difference S
                                   (select S)))))
```

Figure 1: The four hypotheses for number word meanings that we are considering.

One-knower: 30 characters;
Two-knower: 52 chars;
Three-knower: 75 chars;
Cardinal Principle: 60 chars.

2. Let's say the prior probability of a hypothesis h is inversely proportional to its length L_h. So longer lengths are less probable a priori.

$$P(h) = \frac{{L_h}^{-1}}{\sum_i {L_{h_i}}^{-1}} \tag{3}$$

Fill in the prior in Tables 2 and 3.

3. In the original model, the authors used a noisy size principle likelihood, which considers three possible ways the data might have been generated: (i) undefined, i.e., we get no response (ii) correct according to the hypothesis and (iii) correct by random guessing. In the lecture, we formulated the noisy size principle as:

$$P(w|s,h) = \begin{cases} \frac{1}{N} & \text{if } h(s) = \text{undef} \\ \alpha + (1-\alpha)\frac{1}{N} & \text{if } h(s) = w \\ (1-\alpha)\frac{1}{N} & \text{else} \end{cases} \tag{4}$$

Here, let's simplify things and leave out the first case (undefined answers) and assume that the number of words is constant with $N = 10$. We get:

$$P(w|s,h) = \begin{cases} \alpha + \frac{1-\alpha}{10} & \text{if } h(s) = w \\ \frac{1-\alpha}{10} & \text{else} \end{cases} \tag{5}$$

The parameter α reflects how reliably the data comes from the hypothesis. Let's consider $\alpha = 0.9$ for this exercise. Fill in the rest of the likelihoods in Table 3.

4. Now use Bayes Rule to calculate the posterior beliefs over knower levels. In the real world, we often deal with probabilities of events that are really small. To make the calculations easier we can work in logarithms. Here is the log of Bayes rule:

$$\log P(h|d) = \log P(d|h) + \log P(h) - \log P(d), \tag{6}$$

where $P(d) = \sum_h \exp(\log P(d|h) + \log P(h))$.

Most programming languages have a function that computes this `LogSumExp` operation. To make your life easier, the attached python script walks through this computation. You should be able to plug in the above information and it will return the posterior. You can run it locally or copy it on notable.

5. Take a look at the results from the original model (in the slides). Did we replicate their results? Was hypothesis length an appropriate prior? Take a guess on how we might have to change the prior.

   Solution We did not replicate. While length is consistent with simplicity, this implementation does not replicate the original model. What might be done to fix this? Well we need to penalize the CP-knower hypothesis. This is exactly what the original model did, noting that recursive reasoning is a particularly difficult operation. For fun you could implement this bias and add -45 to the CP-knower hypothesis prior.

Knower Level	**Prior**	**Likelihood**$_{N=30}$	**Posterior**
1-knower	0.4037	$(0.91)^{20}(0.01)^{10}$	0.0000
2-knower	0.2329	$(0.91)^{24}(0.01)^{6}$	0.0000
3-knower	0.1615	$(0.91)^{26}(0.01)^{4}$	0.0000
CP-knower	0.2019	$(0.91)^{30}(0.01)^{0}$	0.9999

Table 2: Use this table to write down the components of Bayes Rule for the first dataset $N = 30$.

Knower Level	**Prior**	**Likelihood**$_{N=60}$	**Posterior**
1-knower	0.4037	$(0.91)^{41}(0.01)^{19}$	0.0000
2-knower	0.2329	$(0.91)^{51}(0.01)^{9}$	0.0000
3-knower	0.1615	$(0.91)^{56}(0.01)^{4}$	0.0000
CP-knower	0.2019	$(0.91)^{60}(0.01)^{0}$	0.9999

Table 3: Use this table to write down the components of Bayes Rule for the second dataset $N = 60$.