3 Generative models for discrete data

3.1 Introduction

In Section 2.2.3.2, we discussed how to classify a feature vector $\mathbf{x}$ by applying Bayes rule to a generative classifier of the form

$$p(y = c|\mathbf{x}, \boldsymbol{\theta}) \propto p(\mathbf{x}|y = c, \boldsymbol{\theta})p(y = c|\boldsymbol{\theta}) \tag{3.1}$$

The key to using such models is specifying a suitable form for the class-conditional density $p(\mathbf{x}|y = c, \boldsymbol{\theta})$, which defines what kind of data we expect to see in each class. In this chapter, we focus on the case where the observed data are discrete symbols. We also discuss how to infer the unknown parameters $\boldsymbol{\theta}$ of such models.

3.2 Bayesian concept learning

Consider how a child learns to understand the meaning of a word, such as "dog". Presumably the child's parents point out positive examples of this concept, saying such things as, "look at the cute dog!", or "mind the doggy", etc. However, it is very unlikely that they provide negative examples, by saying "look at that non-dog". Certainly, negative examples may be obtained during an active learning process — the child says "look at the dog" and the parent says "that's a cat, dear, not a dog" — but psychological research has shown that people can learn concepts from positive examples alone (Xu and Tenenbaum 2007).

We can think of learning the meaning of a word as equivalent to **concept learning**, which in turn is equivalent to binary classification. To see this, define $f(x) = 1$ if x is an example of the concept C, and $f(x) = 0$ otherwise. Then the goal is to learn the indicator function f, which just defines which elements are in the set C. By allowing for uncertainty about the definition of f, or equivalently the elements of C, we can emulate **fuzzy set theory**, but using standard probability calculus. Note that standard binary classification techniques require positive and negative examples. By contrast, we will devise a way to learn from positive examples alone.

For pedagogical purposes, we will consider a very simple example of concept learning called the **number game**, based on part of Josh Tenenbaum's PhD thesis (Tenenbaum 1999). The game proceeds as follows. I choose some simple arithmetical concept C, such as "prime number" or "a number between 1 and 10". I then give you a series of randomly chosen positive examples $\mathcal{D} = \{x_1, \ldots, x_N\}$ drawn from C, and ask you whether some new test case $\tilde{x}$ belongs to C, i.e., I ask you to classify $\tilde{x}$.

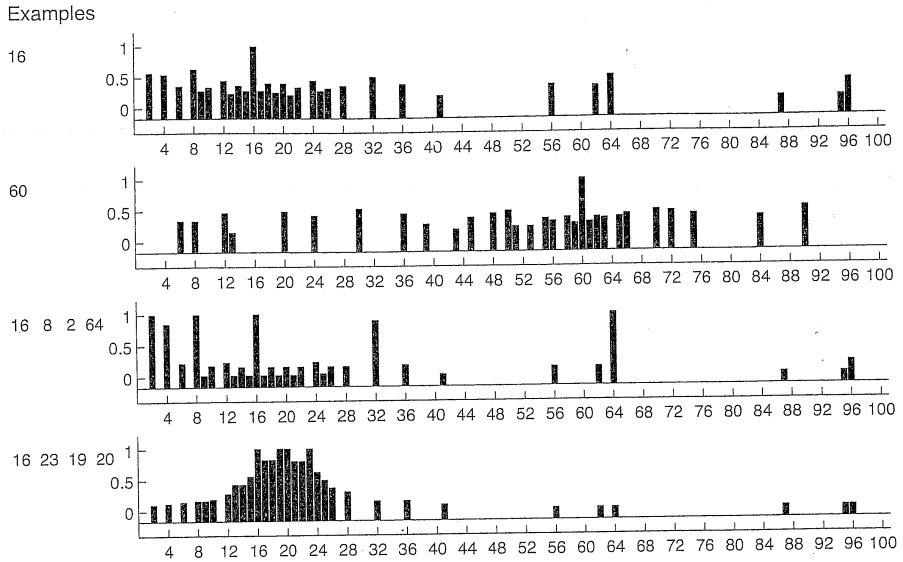

Figure 3.1 Empirical predictive distribution averaged over 8 humans in the number game. First two rows: after seeing $\mathcal{D} = \{16\}$ and $\mathcal{D} = \{60\}$. This illustrates diffuse similarity. Third row: after seeing $\mathcal{D} = \{16, 8, 2, 64\}$. This illustrates rule-like behavior (powers of 2). Bottom row: after seeing $\mathcal{D} = \{16, 23, 19, 20\}$. This illustrates focussed similarity (numbers near 20). Source: Figure 5.5 of (Tenenbaum 1999). Used with kind permission of Josh Tenenbaum.

Suppose, for simplicity, that all numbers are integers between 1 and 100. Now suppose I tell you "16" is a positive example of the concept. What other numbers do you think are positive? 17? 6? 32? 99? It's hard to tell with only one example, so your predictions will be quite vague. Presumably numbers that are similar in some sense to 16 are more likely. But similar in what way? 17 is similar, because it is "close by", 6 is similar because it has a digit in common, 32 is similar because it is also even and a power of 2, but 99 does not seem similar. Thus some numbers are more likely than others. We can represent this as a probability distribution, $p(\tilde{x}|\mathcal{D})$, which is the probability that $\tilde{x} \in C$ given the data $\mathcal{D}$ for any $\tilde{x} \in \{1, \ldots, 100\}$. This is called the **posterior predictive distribution**. Figure 3.1(top) shows the predictive distribution of people derived from a lab experiment. We see that people predict numbers that are similar to 16, under a variety of kinds of similarity.

Now suppose I tell you that 8, 2 and 64 are *also* positive examples. Now you may guess that the hidden concept is "powers of two". This is an example of **induction**. Given this hypothesis, the predictive distribution is quite specific, and puts most of its mass on powers of 2, as shown in Figure 3.1(third row). If instead I tell you the data is $\mathcal{D} = \{16, 23, 19, 20\}$, you will get a different kind of **generalization gradient**, as shown in Figure 3.1(bottom).

How can we explain this behavior and emulate it in a machine? The classic approach to induction is to suppose we have a **hypothesis space** of concepts, $\mathcal{H}$, such as: odd numbers, even numbers, all numbers between 1 and 100, powers of two, all numbers ending in j (for

$0 \leq j \leq 9$), etc. The subset of $\mathcal{H}$ that is consistent with the data D is called the **version space**. As we see more examples, the version space shrinks and we become increasingly certain about the concept (Mitchell 1997).

However, the version space is not the whole story. After seeing $\mathcal{D} = \{16\}$, there are many consistent rules; how do you combine them to predict if $\tilde{x} \in C$? Also, after seeing $\mathcal{D} = \{16, 8, 2, 64\}$, why did you choose the rule "powers of two" and not, say, "all even numbers", or "powers of two except for 32", both of which are equally consistent with the evidence? We will now provide a Bayesian explanation for this.

3.2.1 Likelihood

We must explain why we chose $h_{\text{two}} \triangleq$"powers of two", and not, say, $h_{\text{even}} \triangleq$ "even numbers" after seeing $\mathcal{D} = \{16, 8, 2, 64\}$, given that both hypotheses are consistent with the evidence. The key intuition is that we want to avoid **suspicious coincidences**. If the true concept was even numbers, how come we only saw numbers that happened to be powers of two?

To formalize this, let us assume that examples are sampled uniformly at random from the **extension** of a concept. (The extension of a concept is just the set of numbers that belong to it, e.g., the extension of h_{even} is $\{2, 4, 6, \ldots, 98, 100\}$; the extension of "numbers ending in 9" is $\{9, 19, \ldots, 99\}$.) Tenenbaum calls this the **strong sampling assumption**. Given this assumption, the probability of independently sampling N items (with replacement) from h is given by

$$p(\mathcal{D}|h) = \left[\frac{1}{\text{size}(h)}\right]^N = \left[\frac{1}{|h|}\right]^N \tag{3.2}$$

This crucial equation embodies what Tenenbaum calls the **size principle**, which means the model favors the simplest (smallest) hypothesis consistent with the data. This is more commonly known as **Occam's razor**.[1]

To see how it works, let $\mathcal{D} = \{16\}$. Then $p(\mathcal{D}|h_{\text{two}}) = 1/6$, since there are only 6 powers of two less than 100, but $p(\mathcal{D}|h_{\text{even}}) = 1/50$, since there are 50 even numbers. So the likelihood that $h = h_{\text{two}}$ is higher than if $h = h_{\text{even}}$. After 4 examples, the likelihood of h_{two} is $(1/6)^4 = 7.7 \times 10^{-4}$, whereas the likelihood of h_{even} is $(1/50)^4 = 1.6 \times 10^{-7}$. This is a **likelihood ratio** of almost 5000:1 in favor of h_{two}. This quantifies our earlier intuition that $D = \{16, 8, 2, 64\}$ would be a very suspicious coincidence if generated by h_{even}.

3.2.2 Prior

Suppose $D = \{16, 8, 2, 64\}$. Given this data, the concept h' ="powers of two except 32" is more likely than h ="powers of two", since h' does not need to explain the coincidence that 32 is missing from the set of examples.

However, the hypothesis h' ="powers of two except 32" seems "conceptually unnatural". We can capture such intution by assigning low prior probability to unnatural concepts. Of course, your prior might be different than mine. This **subjective** aspect of Bayesian reasoning is a source of much controversy, since it means, for example, that a child and a math professor

1. William of Occam (also spelt Ockham) was an English monk and philosopher, 1288–1348.

will reach different answers. In fact, they presumably not only have different priors, but also different hypothesis spaces. However, we can finesse that by defining the hypothesis space of the child and the math professor to be the same, and then setting the child's prior weight to be zero on certain "advanced" concepts. Thus there is no sharp distinction between the prior and the hypothesis space.

Although the subjectivity of the prior is controversial, it is actually quite useful. If you are told the numbers are from some arithmetic rule, then given 1200, 1500, 900 and 1400, you may think 400 is likely but 1183 is unlikely. But if you are told that the numbers are examples of healthy cholesterol levels, you would probably think 400 is unlikely and 1183 is likely. Thus we see that the prior is the mechanism by which background knowledge can be brought to bear on a problem. Without this, rapid learning (i.e., from small samples sizes) is impossible.

So, what prior should we use? For illustration purposes, let us use a simple prior which puts uniform probability on 30 simple arithmetical concepts, such as "even numbers", "odd numbers", "prime numbers", "numbers ending in 9", etc. To make things more interesting, we make the concepts even and odd more likely apriori. We also include two "unnatural" concepts, namely "powers of 2, plus 37" and "powers of 2, except 32", but give them low prior weight. See Figure 3.2(a) for a plot of this prior. We will consider a slightly more sophisticated prior later on.

3.2.3 Posterior

The posterior is simply the likelihood times the prior, normalized. In this context we have

$$p(h|\mathcal{D}) = \frac{p(\mathcal{D}|h)p(h)}{\sum_{h' \in \mathcal{H}} p(\mathcal{D}, h')} = \frac{p(h)\mathbb{I}(\mathcal{D} \in h)/|h|^N}{\sum_{h' \in \mathcal{H}} p(h')\mathbb{I}(\mathcal{D} \in h')/|h'|^N} \tag{3.3}$$

where $\mathbb{I}(\mathcal{D} \in h)$ is 1 **iff** (if and only if) all the data are in the extension of the hypothesis h. Figure 3.2 plots the prior, likelihood and posterior after seeing $\mathcal{D} = \{16\}$. We see that the posterior is a combination of prior and likelihood. In the case of most of the concepts, the prior is uniform, so the posterior is proportional to the likelihood. However, the "unnatural" concepts of "powers of 2, plus 37" and "powers of 2, except 32" have low posterior support, despite having high likelihood, due to the low prior. Conversely, the concept of odd numbers has low posterior support, despite having a high prior, due to the low likelihood.

Figure 3.3 plots the prior, likelihood and posterior after seeing $\mathcal{D} = \{16, 8, 2, 64\}$. Now the likelihood is much more peaked on the powers of two concept, so this dominates the posterior. Essentially the learner has an **aha** moment, and figures out the true concept. (Here we see the need for the low prior on the unnatural concepts, otherwise we would have overfit the data and picked "powers of 2, except for 32".)

In general, when we have enough data, the posterior $p(h|\mathcal{D})$ becomes peaked on a single concept, namely the MAP estimate, i.e.,

$$p(h|\mathcal{D}) \to \delta_{\hat{h}^{MAP}}(h) \tag{3.4}$$

where $\hat{h}^{MAP} = \text{argmax}_h \, p(h|\mathcal{D})$ is the posterior mode, and where δ is the **Dirac measure** defined by

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} \tag{3.5}$$

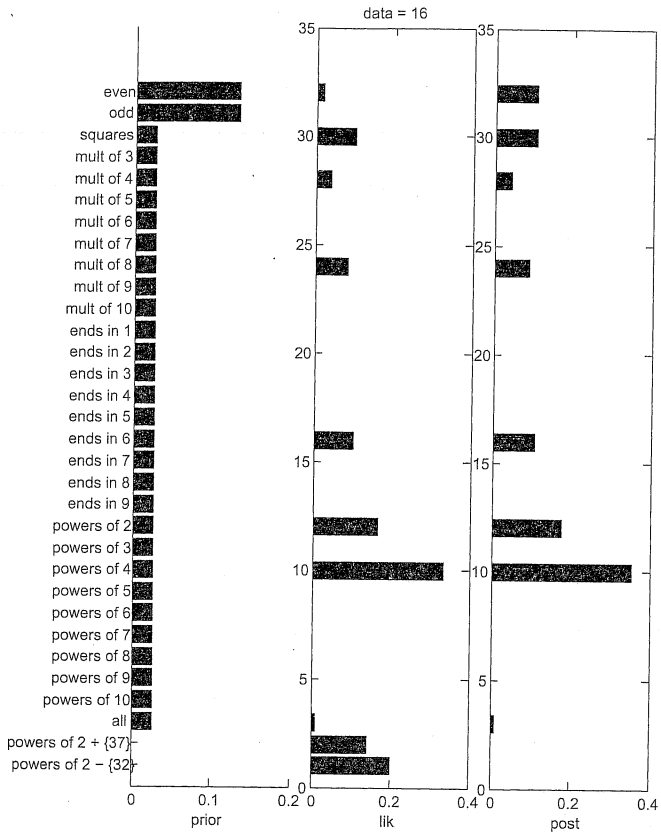

Figure 3.2 Prior, likelihood and posterior for $\mathcal{D} = \{16\}$. Based on (Tenenbaum 1999). Figure generated by `numbersGame`.

Note that the MAP estimate can be written as

$$\hat{h}^{MAP} = \underset{h}{\operatorname{argmax}}\, p(\mathcal{D}|h)p(h) = \underset{h}{\operatorname{argmax}}\, [\log p(\mathcal{D}|h) + \log p(h)] \tag{3.6}$$

Since the likelihood term depends exponentially on N, and the prior stays constant, as we get more and more data, the MAP estimate converges towards the **maximum likelihood estimate** or **MLE**:

$$\hat{h}^{mle} \triangleq \underset{h}{\operatorname{argmax}}\, p(\mathcal{D}|h) = \underset{h}{\operatorname{argmax}} \log p(\mathcal{D}|h) \tag{3.7}$$

In other words, if we have enough data, we see that the **data overwhelms the prior**. In this

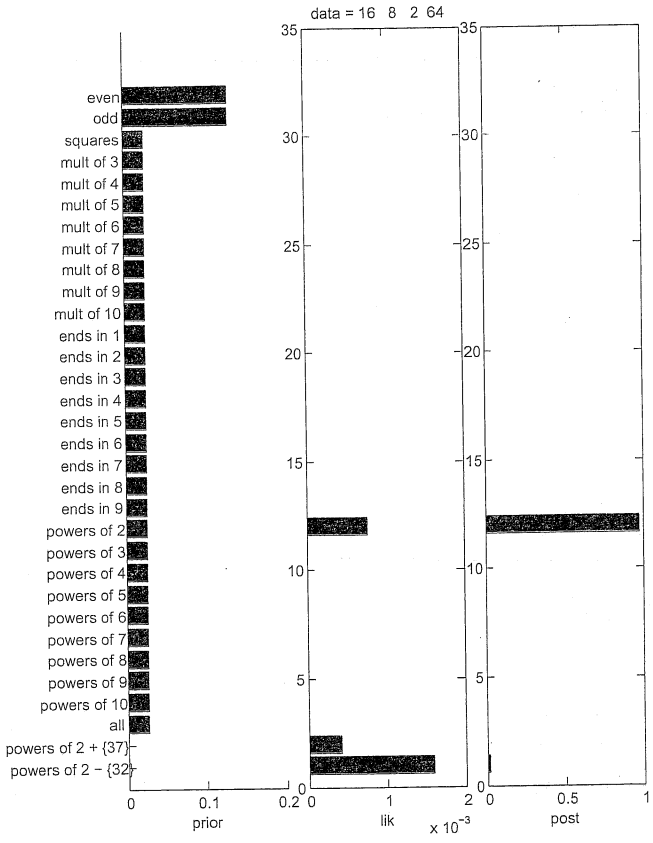

Figure 3.3 Prior, likelihood and posterior for $\mathcal{D} = \{16, 8, 2, 64\}$. Based on (Tenenbaum 1999). Figure generated by `numbersGame`.

case, the MAP estimate converges towards the MLE.

If the true hypothesis is in the hypothesis space, then the MAP/ ML estimate will converge upon this hypothesis. Thus we say that Bayesian inference (and ML estimation) are consistent estimators (see Section 6.4.1 for details). We also say that the hypothesis space is **identifiable in the limit**, meaning we can recover the truth in the limit of infinite data. If our hypothesis class is not rich enough to represent the "truth" (which will usually be the case), we will converge on the hypothesis that is as close as possible to the truth. However, formalizing this notion of "closeness" is beyond the scope of this chapter.

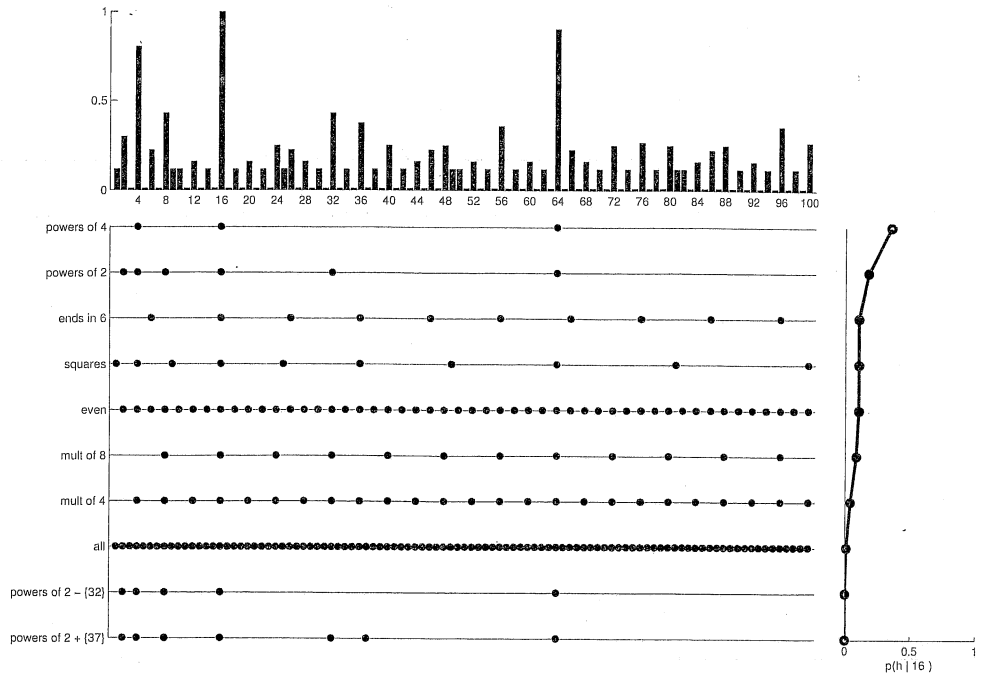

Figure 3.4 Posterior over hypotheses and the corresponding predictive distribution after seeing one example, $\mathcal{D} = \{16\}$. A dot means this number is consistent with this hypothesis. The graph $p(h|\mathcal{D})$ on the right is the weight given to hypothesis h. By taking a weighed sum of dots, we get $p(\tilde{x} \in C|\mathcal{D})$ (top). Based on Figure 2.9 of (Tenenbaum 1999). Figure generated by `numbersGame`.

3.2.4 Posterior predictive distribution

The posterior is our internal **belief state** about the world. The way to test if our beliefs are justified is to use them to predict objectively observable quantities (this is the basis of the scientific method). Specifically, the posterior predictive distribution in this context is given by

$$p(\tilde{x} \in C|\mathcal{D}) = \sum_h p(y = 1|\tilde{x}, h)p(h|\mathcal{D}) \tag{3.8}$$

This is just a weighted average of the predictions of each individual hypothesis and is called **Bayes model averaging** (Hoeting et al. 1999). This is illustrated in Figure 3.4. The dots at the bottom show the predictions from each hypothesis; the vertical curve on the right shows the weight associated with each hypothesis. If we multiply each row by its weight and add up, we get the distribution at the top.

When we have a small and/or ambiguous dataset, the posterior $p(h|\mathcal{D})$ is vague, which induces a broad predictive distribution. However, once we have "figured things out", the posterior becomes a delta function centered at the MAP estimate. In this case, the predictive distribution

becomes

$$p(\tilde{x} \in C|\mathcal{D}) = \sum_h p(\tilde{x}|h)\delta_{\hat{h}}(h) = p(\tilde{x}|\hat{h}) \tag{3.9}$$

This is called a **plug-in approximation** to the predictive density and is very widely used, due to its simplicity. However, in general, this under-represents our uncertainty, and our predictions will not be as "smooth" as when using BMA. We will see more examples of this later in the book.

Although MAP learning is simple, it cannot explain the gradual shift from similarity-based reasoning (with uncertain posteriors) to rule-based reasoning (with certain posteriors). For example, suppose we observe $\mathcal{D} = \{16\}$. If we use the simple prior above, the minimal consistent hypothesis is "all powers of 4", so only 4 and 16 get a non-zero probability of being predicted. This is of course an example of overfitting. Given $\mathcal{D} = \{16, 8, 2, 64\}$, the MAP hypothesis is "all powers of two". Thus the plug-in predictive distribution gets broader (or stays the same) as we see more data: it starts narrow, but is forced to broaden as it sees more data. In contrast, in the Bayesian approach, we start broad and then narrow down as we learn more, which makes more intuitive sense. In particular, given $\mathcal{D} = \{16\}$, there are many hypotheses with non-negligible posterior support, so the predictive distribution is broad. However, when we see $\mathcal{D} = \{16, 8, 2, 64\}$, the posterior concentrates its mass on one hypothesis, so the predictive distribution becomes narrower. So the predictions made by a plug-in approach and a Bayesian approach are quite different in the small sample regime, although they converge to the same answer as we see more data.

3.2.5 A more complex prior

To model human behavior, Tenenbaum used a slightly more sophisticated prior which was derived by analysing some experimental data of how people measure similarity between numbers; see (Tenenbaum 1999, p208) for details. The result is a set of arithmetical concepts similar to those mentioned above, plus all intervals between n and m for $1 \leq n, m \leq 100$. (Note that these hypotheses are not mutually exclusive.) Thus the prior is a **mixture** of two priors, one over arithmetical rules, and one over intervals:

$$p(h) = \pi_0 p_{\text{rules}}(h) + (1 - \pi_0) p_{\text{interval}}(h) \tag{3.10}$$

The only free parameter in the model is the relative weight, π_0, given to these two parts of the prior. The results are not very sensitive to this value, so long as $\pi_0 > 0.5$, reflecting the fact that people are more likely to think of concepts defined by rules. The predictive distribution of the model, using this larger hypothesis space, is shown in Figure 3.5. It is strikingly similar to the human predictive distribution, shown in Figure 3.1, even though it was not fit to human data (modulo the choice of hypothesis space).

3.3 The beta-binomial model

The number game involved inferring a distribution over a discrete variable drawn from a finite hypothesis space, $h \in \mathcal{H}$, given a series of discrete observations. This made the computations particularly simple: we just needed to sum, multiply and divide. However, in many applications, the unknown parameters are continuous, so the hypothesis space is (some subset) of $\mathbb{R}^K$, where