Bayesian concept learning

Kevin P. Murphy

Last updated August 27, 2007

1 Introduction

This chapter, which is a summary of Josh Tenenbaum's PhD thesis [**?**], provides an intuitive introduction to the key ideas behind Bayesian learning in relatively simple settings. In subsequent chapters we will study more sophisticated models and more efficient computational techniques. Bayesian techniques are particularly useful when learning from small datasets, as humans often have to do.

Consider the problem of learning to understand the meaning of a word, such as "dog". Presumably, as a child, one'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. This is in contrast to many machine learning approaches to **concept learning (binary classification)**, which require positive and negative data. (We will study such methods later.)

In this chapter, we will explain how it is possible to learn concepts from positive only data, using three examples proposed by Tenenbaum. The first is a simple discrete domain, where must identify an arithmetic rule from a series of numbers (see Section 2). The second is a simple continuous domain, where one must identify the true (rectangular-shaped) boundary that distinguishes positive from negative examples, given positive examples alone (see Section 3). The third example is an attempt to model human word learning based on visual similarity of objects.

2 A discrete domain (the number concept)

Suppose I tell you I am thinking of some simple arithemetical concept C. I give you a series of *randomly chosen positive examples* $X = \{x_1, \ldots, x_N\}$, and ask you whether any other test cases y belong to the **extension** of C. Suppose, for simplicity, that all numbers are between 1 and 100, so the task is to compute whether $y \in C$ given X, for $y \in \{1, \ldots, 100\}$; this is called the the **generalization function**.

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 the predictive distribution is quite vague: see Figure 1(a). 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 with 16, 32 is similar because it is also even and a power of 2, but 99 does not seem similar. Thus some concepts are more likely than others, which induces a non-uniform predictive distribution: see Figure 2(top). Learning from one example is called **one-shot learning**, although arguably we haven't actually learned much yet (because our prior was so vague).

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": see Figure 1(b). This is an example of **induction**. Given this hypothesis, the predictive distribution is quite specific: see Figure 2(bottom). We predict that other numbers belong to the concept of the basis of a **rule** rather than on the basis of similarity.

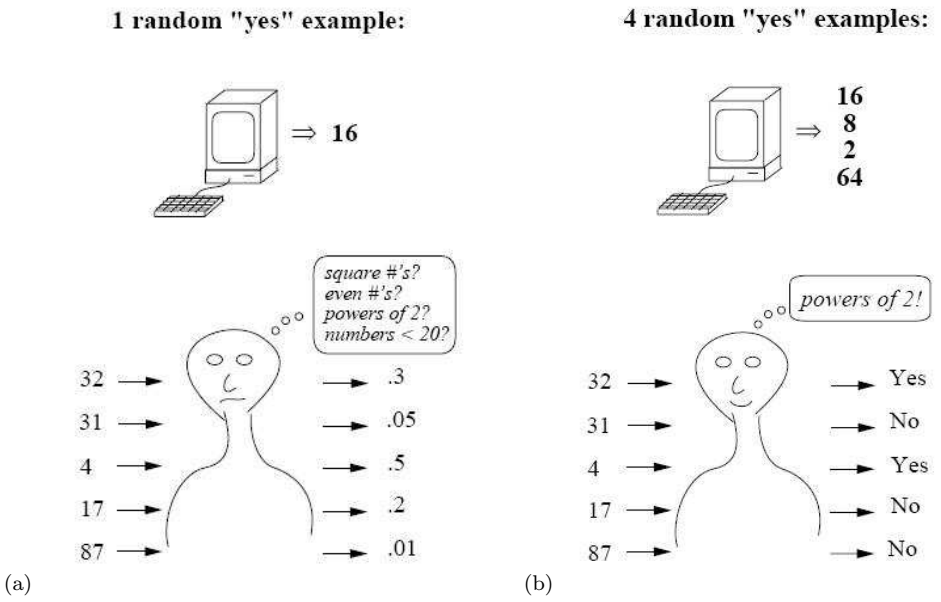


Figure 1: The number game. Belief state after seeing (a) 1 example, (b) 4 examples.

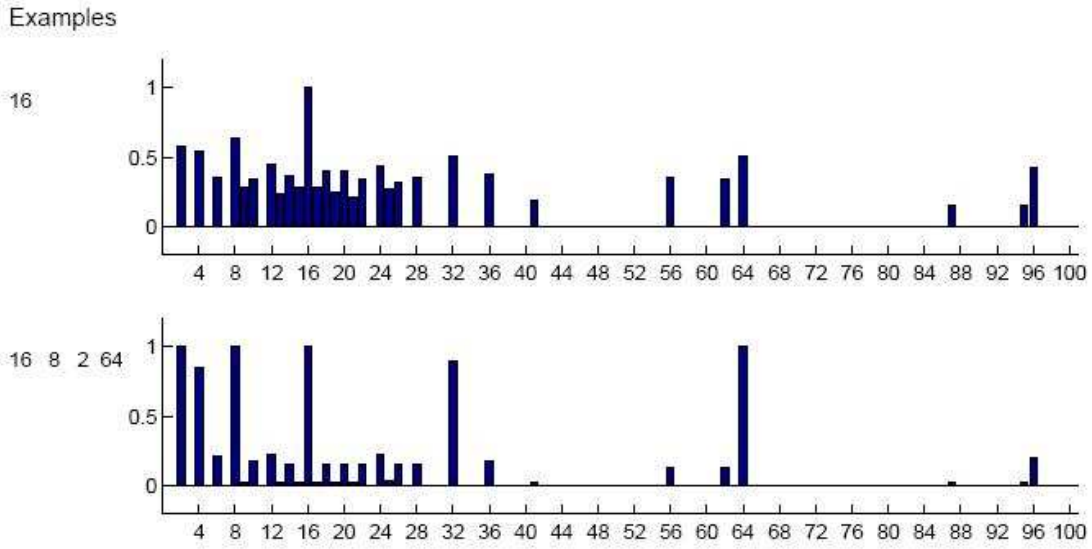


Figure 2: Empirical predictive distribution of humans in the number game. Top: after seeing one example. Bottom: after seeing 4 examples.

The classic approach to rule based 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. (We discuss where the hypothesis space comes from in Section 2.6). The subset of $\mathcal{H}$ that is consistent with the data X is called the **version space** and is written $\mathcal{H}_X$. As we see more examples, the version space shrinks and we become increasingly certain about the extension of the concept.

However, the version space is not the whole story. After seeing $X = 16$, there are many consistent rules; how do you combine them to predict if $y \in C$? Also, after seeing $X = \{16, 8, 2, 64\}$, 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? This is what we seek to explain.

2.1 Generalization

In the Bayesian approach, we maintain a *probability distribution over hypotheses*, $p(h|X)$, which is like the version space but has much more information. We call it our **posterior belief state**.

Assuming we have such a distribution, we can predict the future even when we are uncertain about the exact concept. Specifically, we can compute the **posterior predictive distribution** by **marginalizing out** the **nuisance variable** h:

$$p(y \in C|X) = \sum_{h \in \mathcal{H}} p(y \in C|h)p(h|X) \tag{1}$$

where $\sum_{h \in \mathcal{H}} p(h|X) = 1$. This is called **Bayesian model averaging**.

In this simple (noise-free) example, $p(y \in C|h) = 1$ if y is consistent with h, and is 0 otherwise. (For example, $p(32 \in C|h = \text{even numbers}) = 1.0$, but $p(33 \in C|h = \text{even numbers}) = 0.0$.) Hence we can rewrite the above as

$$p(y \in C|X) = \sum_{h \in \mathcal{H}_y} p(h|X) \tag{2}$$

where $\mathcal{H}_y$ are all hypothesis that are consistent with y. Thus the predictive distribution is just a weighted sum of consistent hypotheses; we discuss how to compute the weights $p(h|X)$ below.

When we have a small dataset, $p(h|X)$ is vague (has high **entropy**) which induces a broad predictive distribution: see Figure 3. In this case, generalization is similarity-like. But when the dataset increases, the posterior (usually) becomes sharper (has lower entropy), and so does the predictive distribution: see Figure 4. In this case, generalization is rule-like.

2.2 Bayesian inference

By Bayes rule, we can compute the posterior distribution as follows

$$p(h|X) = \frac{p(X|h)p(h)}{\sum_{h'} p(X|h')p(h')} \tag{3}$$

We therefore need to specify the prior $p(h)$ and the likelihood function $p(X|h)$. For more realistic problems, we will also need to discuss how to compute this summation in the denominator tractably, but in this simple case, we can use exhaustive enumeration.

2.3 Likelihood

We must explain why we chose $h =$"powers of two", and not, say, $h' =$"even numbers" after seeing $X = \{16, 8, 2, 64\}$, given that both hypotheses are consistent with the evidence. The key idea is that we want to **avoid suspicious coincidences**. If the true concept were even numbers, how come we didn't see any numbers that weren't powers of two? (See Figure 5.)

Note that the fact that $X = \{16, 8, 2, 64\}$ is considered "suspicious" is because we are implicitly making the **strong sampling assumption**, namely that the examples were chosenly randomly from the concept's

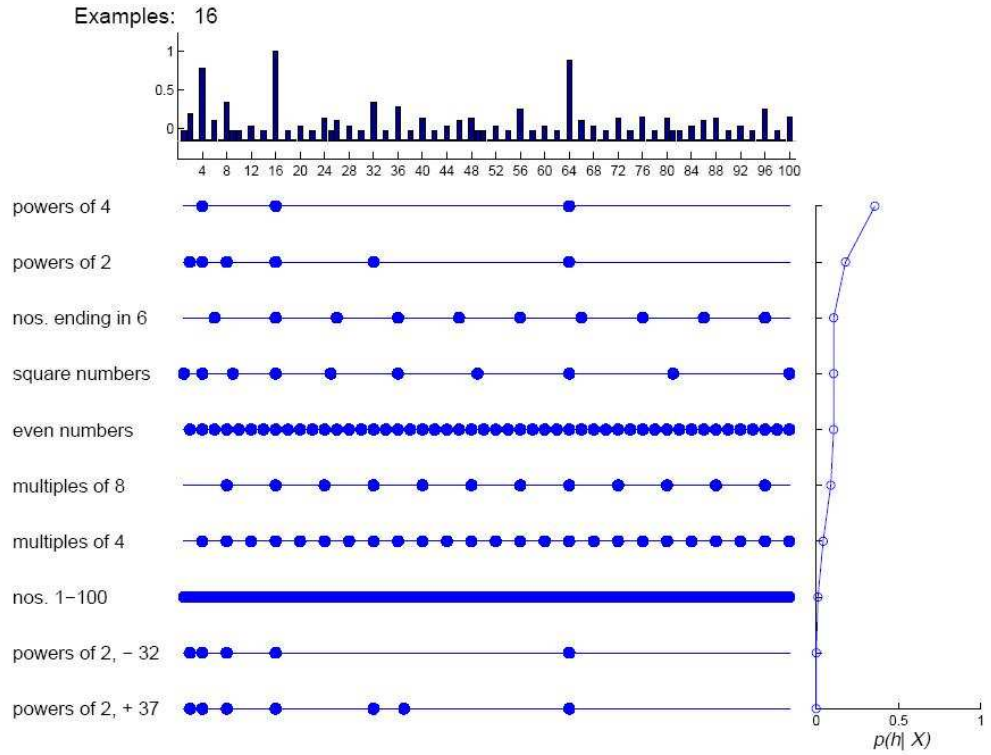


Figure 3: Posterior over hypotheses and the corresponding predictive distribution after seeing one example. A dot means this number is consistent with this hypothesis. The graph $p(h|X)$ on the right is the weight given to hypothesis h. By taking a weighed sum of dots, we get $p(y \in C|X)$ (top).

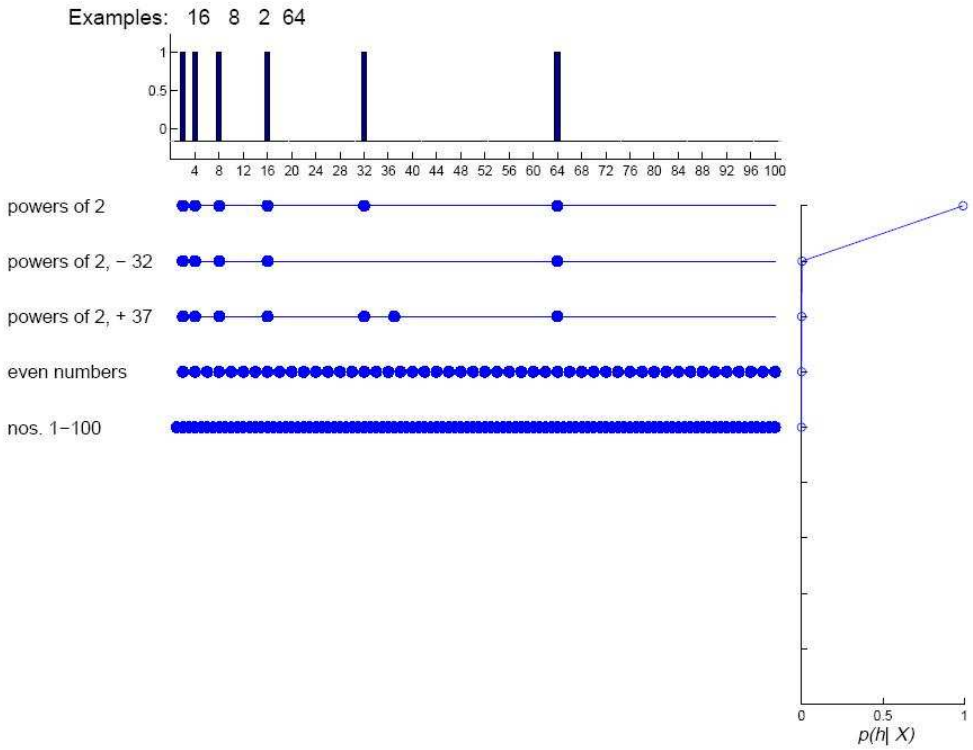


Figure 4: Posterior over hypotheses and the corresponding predictive distribution after seeing four examples.

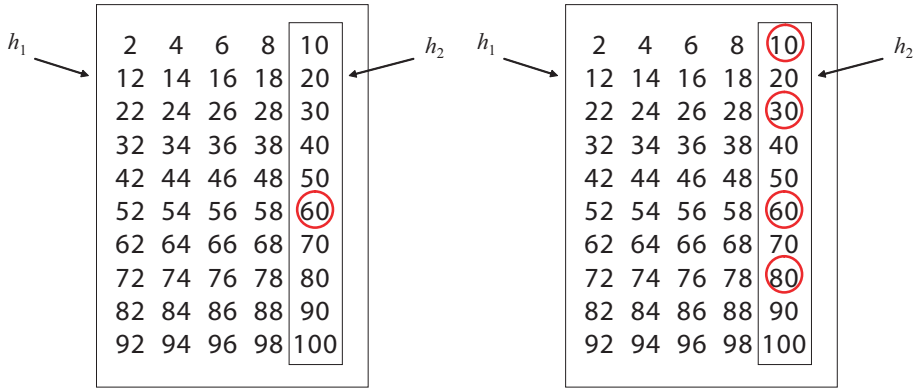


Figure 5: Illustration of the size principle. Consider h_1 = even numbers and h_2 = multiples of 10. If $X = \{60\}$, it is slightly more of a coincidence under h_1; but if $X = \{10, 30, 60, 80\}$, it is much more of a coincidence under h_1, i.e., $p(X|h_1) \ll p(X|h_2)$. Thus the more data we get, the more likely the simpler hypothesis becomes. This is an example of the Bayesian Occam's razor.

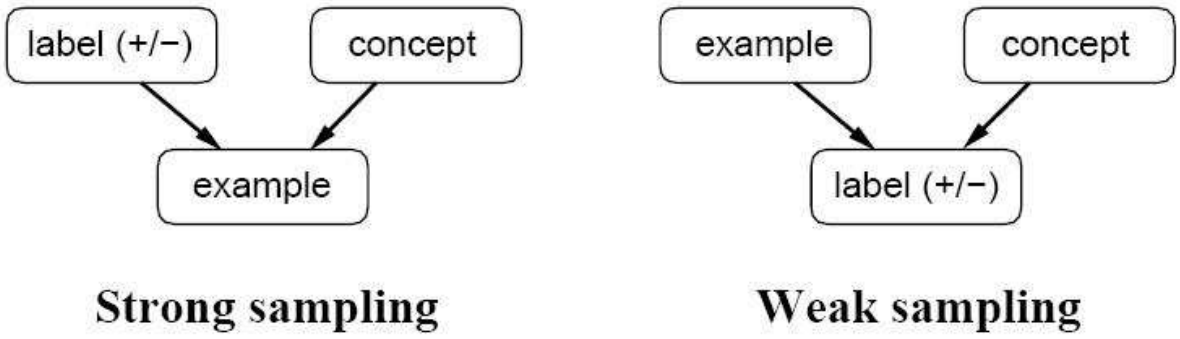


Figure 6: Strong vs weak sampling.

extension (see Figure 6). Under a **weak sampling assumption**, whereby numbers are chosen at random and then are merely labeled as positive or negative, the surprise would focus on the random number generator, that happened to generate four powers of two in a row, rather than on the program, which merely labeled them as positive.

Under the strong sampling model, the probability of independently sampling n items (with replacement) from h is given by

$$p(X|h) = \left[\frac{1}{\text{size}(h)}\right]^n = \left[\frac{1}{|h|}\right]^n \tag{4}$$

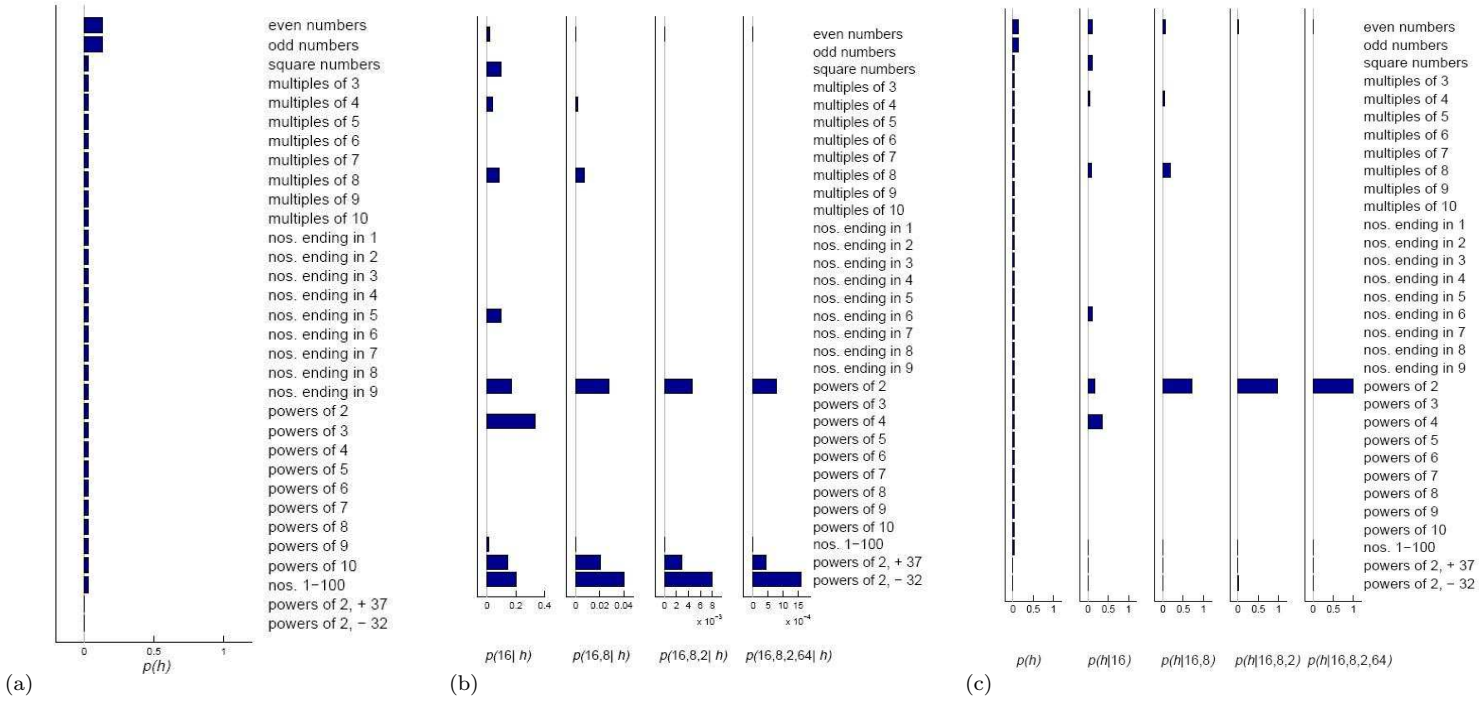
(We will consider more realistic likelihood models, than can handle noise, outliers, etc. later.) This crucial equation is called the **size principle**, and is a form of **Ockham's razor**, which says one should pick the simplest explanation that is consistent with the data. To see how it works, let $X = 16$. Then $p(X|h = \text{powers of two}) = 1/6$, since there are only 6 powers of two less than 100. This is more likely than the following, more general concept: $p(X|h = \text{even numbers}) = 1/50$. Of course, both of these are more likely than inconsistent concepts: $p(X|h = \text{odd numbers}) = 0$. Figure 7(b) shows how the likelihood function becomes exponentially more peaked on the smallest consistent hypotheses. After 4 examples, the likelihood of "powers of two" is $1/6^4 = 7.7 \times 10^{-4}$, whereas the likelihood of "even numbers" is $1/50^4 = 1.6 \times 10^{-7}$. This is a **likelihood ratio** of almost 5000:1 in favor of "power of two". This quantifies our earlier intuition that $X = \{16, 8, 2, 64\}$ would be a very suspicious coincidence if generated by "even numbers".

However, note that the most likely hypothesis is not "powers of two", but rather the rather unnatural hypothesis, "powers of two except 32". This has higher likelihood because it does not need to explain the (small) coincidence that we did not see 32. To rule out such "unnatural" concepts, we need a prior, as we discuss in Section 2.4.

2.4 Priors

We must explain why we chose $h =$"powers of two", and not, say, $h' =$"powers of two except 32", after seeing $X = \{16, 8, 2, 64\}$. After all, h' has higher likelihood, since it does not need to explain the coincidence that 32 is missing from the set of examples. However, h' is much less likely than h a priori, because it is "conceptually unnatural". It is the combination of the likelihood *and* the prior that determines the posterior.

One possible prior on hypotheses is shown in Figure 7(a). This puts less weight on "unnatural" concepts such as "powers of two except 32", and more weight on very simple concepts like "even numbers". Of course, your prior might be different. This **subjective** aspect of Bayesian reasoning is a source of controversy, since it means, for example, that a child and a math professor (who presumably not only have different priors, but different hypothesis spaces) will reach different answers. (Note that we can define the hypothesis space of the child and the math professor to be the same, and simply set 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.)



(a) (b) (c)

Figure 7: 7(a) One possible prior. 7(b) Likelihood as a function of sample size. 7(c) Posterior as a function of sample size.

On the other hand, this **context dependence** of the prior is actually quite useful. If you are told the numbers are from some arithmetic rule, 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 cholestrol levels, you would probably think 400 is unlikely and 1183 is likely. So the prior is the mechanism by which **background knowledge** can be brought to bear.

2.5 Posterior

The posterior is simply the likelihood times the prior, normalized:

$$p(h|X) = \frac{p(X|h)p(h)}{\sum_{h' \in \mathcal{H}} p(X, h')} \tag{5}$$

$$= \frac{p(h)/|h|^n}{\sum_{h' \in \mathcal{H}_X} p(h')/|h'|^n} \tag{6}$$

(In more complex models, this normalization procedure can be computationally difficult, but we ignore this for now.) The result is shown in Figure 7(c). Note that the single sharp peak obtained after 4 examples is not present in either the prior (Figure 7(a)) or the likelihood (Figure 7(b)).

2.6 More accurate model

The hypothesis space used above contains just 30 hypotheses for simplicity. To more accurately model the human data in Figure 2, Tenenbaum used the 5090 hypotheses in Figure 8, with results shown in Figure 9. This hypothesis space, which contains 40 mathematical concepts and 5050 interval/ magnitude hypotheses, was derived by analysing some experimental data of how people measure similarity between numbers (see [**?**, p208] for details).

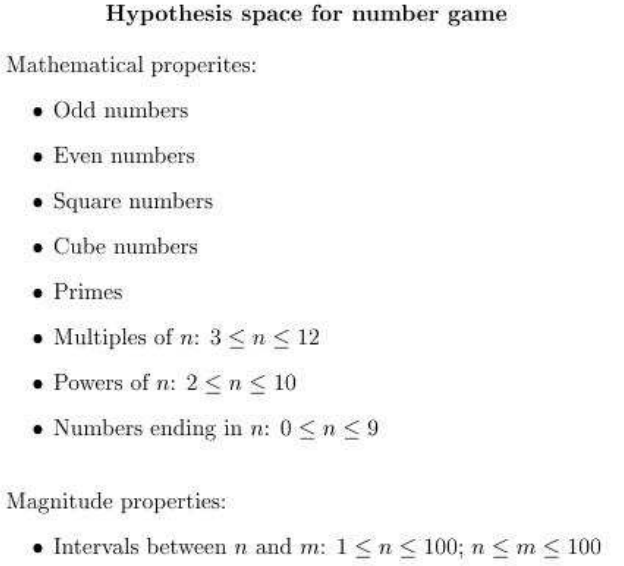


Figure 8: Complete hypothesis space for the number game. There are 40 mathematical hypotheses, and 5050 magnitude/ interval hypotheses.

To specify a prior on this hypothesis space, let us put weight $0 < \lambda < 1$ on the mathematical concepts, and weight $1 - \lambda$ on the interval concepts. (This is an example of a **mixture model**; λ and $1 - \lambda$ are called the **mixing weights**.) Within the mathematical concepts, we will use a uniform prior, so each one has prior $\lambda/40$. λ is called a **hyper-parameter**, since it is a parameter of the prior; Tenenbaum used $\lambda = 2/3$ (chosen by hand). Within the interval concepts, we can also use a uniform prior[1], in which case each hypothesis gets weight $(1 - \lambda)/5050$. Hence any individual interval hypothesis has lower prior, reflecting an a priori preference to explain data using compact rules. (This is orthogonal to the likelihood-induced bias towards small hypotheses.) This two-stage definition is an example of a **hierarchical prior**.

The overall model is called a **generative model**, since it specifies a procedure for generating data (positive examples) as follows: first decide if the concept is mathematical or interval (by tossing a coin with probability of heads λ); second, pick a specific rule or interval from within the set (by choosing a number uniformly between 1 and 40, or 1 and 5050); finally, pick a specific number (uniformly at random) consistent with the rule or interval. In more realistic models, we may also add **noise** to the observation as a final step. See Section **??**.

2.7 Special cases of the Bayesian framework

A summary of the Bayesian approach is given in Figure 10. The key "ingredients" are:

1. A constrained hypothesis space. Without this, it is impossible to generalize from a finite data set, because any hypothesis consistent with the evidence is possible.

2. An informative prior, that ranks members of the hypothesis space. The alternative is to have a uniform prior, $p(h) = 1/|\mathcal{H}|$.

3. The size principle, which is the likelihood function of a strong sampling model. The alternative is simply to enforce consistency, $p(X|h) = 1$ if $h \in \mathcal{H}_X$ and 0 otherwise.

[1] In fact Tenenbaum used an **Erlang prior** for the intervals, with hyperparameter $\sigma = 10$: see Section **??** for details.)

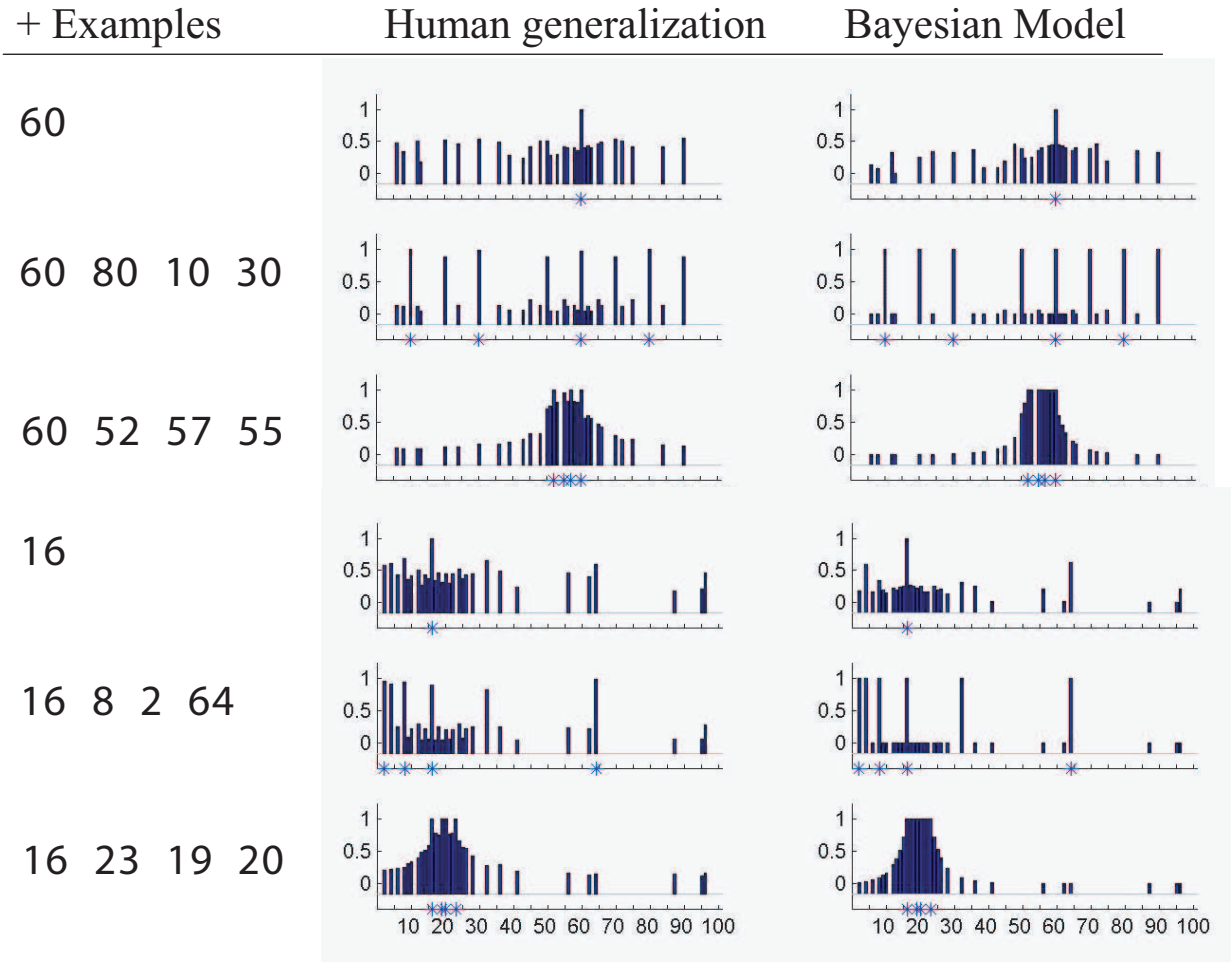


Figure 9: Predictive distributions for people and model using the full hypothesis space. We either get rule-like generalization or similarity-like generalization, depending on which hypotheses have higher posterior probability.

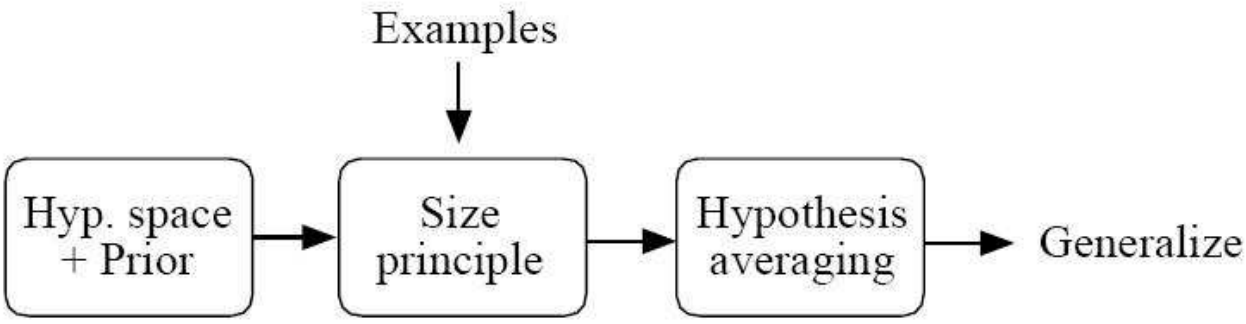


Figure 10: Summary of the Bayesian approach to concept learning.

4. Hypothesis averaging, i.e., integrating out h when making predictions

$$p(y \in C|X) = \sum_h p(y \in C|h)p(h|X) \tag{7}$$

The alternative is simply to pick the most probable **MAP (maximum a posterior)** hypothesis

$$\hat{h} = \arg\max_h p(h|X) \tag{8}$$

and then use this for prediction as a **plug-in estimate**:

$$p(y \in C|X) \approx p(y \in C|\hat{h}) \tag{9}$$

If the posterior is peaked, so $p(h|X) \approx \delta(h - \hat{h})$[2] then the plug-in predictor is a good approximation, since

$$p(y \in C|X) = \sum_h p(y \in C|h)p(h|X) \approx \sum_h p(y \in C|h)\delta(h - \hat{h}) = p(y \in C|\hat{h}) \tag{10}$$

Various other models have been proposed that lack one or more of these ingredients. It is interesting to consider their weaknesses.

Maximum likelihood (ML) learning is ingredients 1 and 3 (no prior, no averaging). This is also called the MIN method, since it picks the smallest (minimal) consistent hypothesis. Since there is no hypothesis averaging, its generalization behavior is all-or-none. For example, given $X = 16$, the minimal consistent hypothesis is $16 : 16$, so only 16 gets a non-zero probability. (If we only use mathematical concept,s the minimal consistent hypothesis is "all powers of 4", so only 4 and 16 get a non-zero probability.) Given $X = \{16, 8, 2, 64\}$, the minimal consistent hypothesis is "all powers of two", which is the same as the Bayesian model. Thus the ML predictive distribution gets broader (or stays the same) as we see more data, contrary to the Bayesian approach, which gets narrower as we see more data. The Bayesian approach seems more natural, since more data should reduce our uncertainty and hence narrow the predictive distribution. But this implies that Bayes was initially broad; in contrast, ML is very **conservative** and is initially narrow, to avoid the risk of over-generalizing. As the amount of data goes to infinity, the Bayesian and the ML approach reach the same answer, because the prior has constant magnitude, whereas the likelihood term depends exponentially on n. If **truth is in the hypothesis space**, then both methods will converge upon the correct hypothesis ; thus both techniques are **consistent**. We say that the hypothesis space is **identifiable in the limit**.

MAP learning is ingredients 1, 2 and 3 (no averaging). This cannot explain the shift from similarity-based reasoning (with uncertain posteriors) to rule-based reasoning (with certain posteriors). But in the large sample limit, it does as well as Bayes, since the likelihood overwhelms the prior.

One can imagine using ingredients 1 and 4 only — no prior and using weak sampling, $p(X|h) = 1$ is X is consistent with h, and 0 otherwise. With this model, the predictive function is just

$$p(y \in C|X) = \frac{|\mathcal{H}_{X,y}|}{|\mathcal{H}_X|} \tag{11}$$

This is similar to the way similarity based approaches work: the probability y belongs to the same set as X is the number of features it shares with the examples X, divided by the number of features common to all examples in X. Unfortunately, this does not work very well. If $X = \{16, 8, 2, 64\}$, there are 3 consistent hypotheses: all powers of two, all even numbers, and all numbers less than 100. Each of these gets equal weight, so a number such as 88, which is consistent with two of the hypotheses, gets probability 2/3 of being positive, and numbers such as 87, which is consistent with one hypothesis, gets a non-negligible 1/3 probability. For this reason, the "weak Bayes" model is not consistent, i.e., it does not converge on the true

[2] This is the delta function and is defined as $\delta(u) = 1$ if $u = 0$ and $\delta(u) = 0$ otherwise. An alternative would be to write $I(h == \hat{h})$, where $I(u)$ is the indicator function, which is 1 if u is true and 0 otherwise.

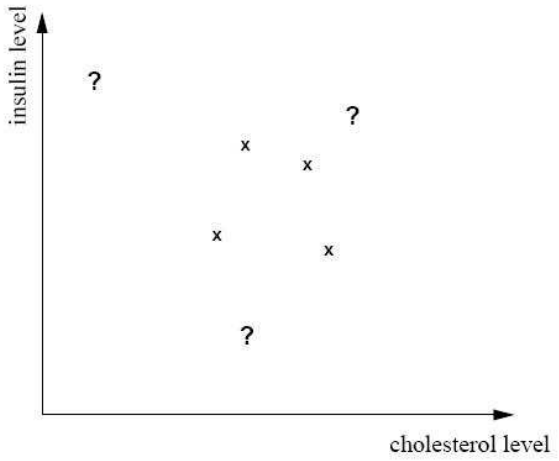


Figure 11: The healthy levels concept

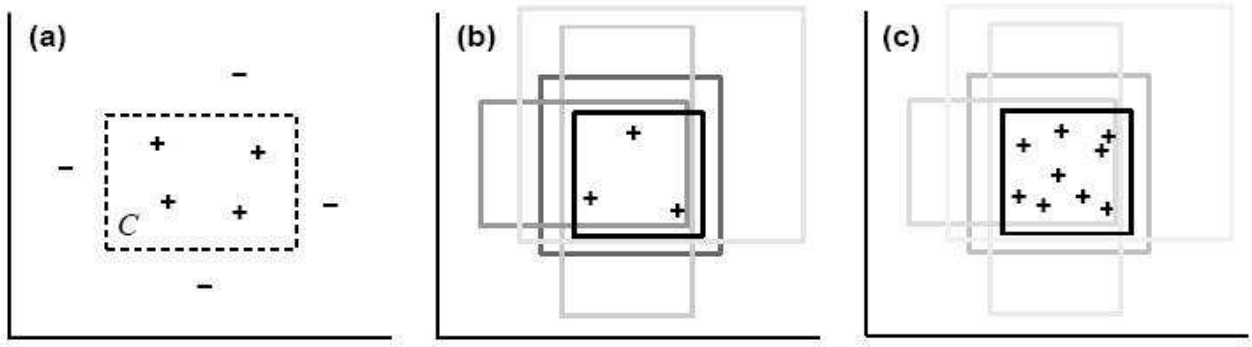


Figure 12: Axes parallel rectangles

hypothesis even as the sample size increases, since the posterior weights are independent of sample size. One can add ingredient 2 (informative prior), which amounts to putting weights on the features when measuring similarity, but this does not solve the consistency problem. So we see that strong sampling is crucial to ensure consistency, as well as rapid learning from small samples.

3 A continuous domain (the healthy levels concept)

We now consider modeling real-valued data, which complicates the mathematics, although the basic ideas are the same. Suppose we measure two continuous variables, the cholestrol and insulin levels of some randomly chosen healthy patients. We would like to know what range of values correspond to a healthy range. As usual, we want to learn the "healthy levels" concept from positive data alone: see Figure 11.

Let our hypothesis space be **axis-parallel rectangles**, as in Figure 12. This is reasonable, since we know (from prior domain knowledge) that healthy levels of both insulin and cholestrol must fall between (unknown) upper *and* lower bounds. (If the problem were to learn healthy levels of some chemical polutant,

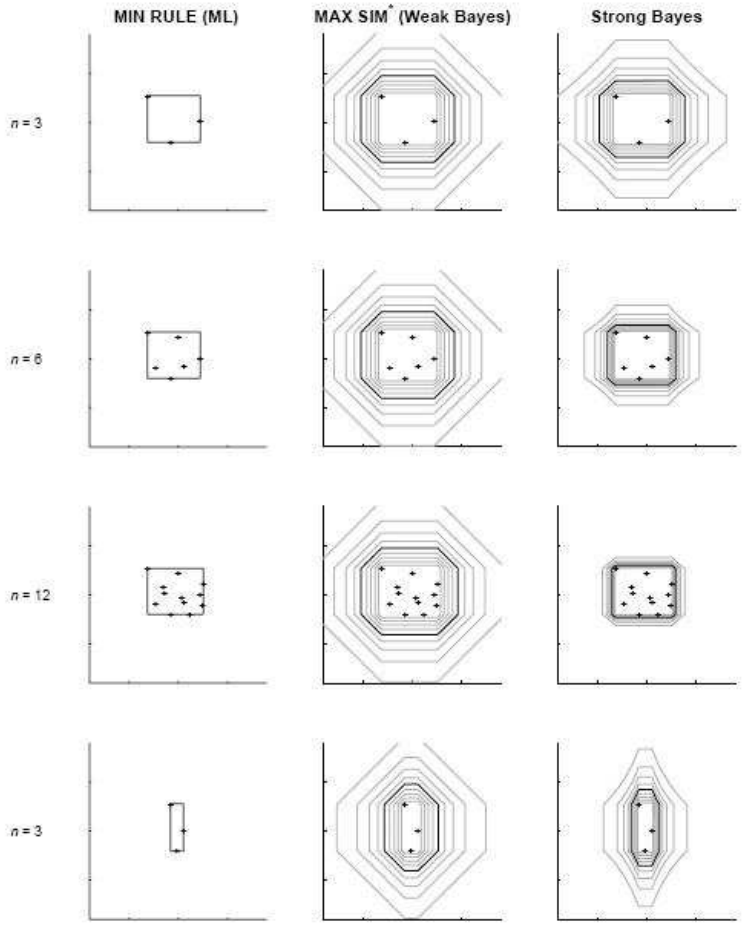


Figure 13: Generalization functions for three different methods on the healthy levels game.

we would use a different hypothesis space, since presumably zero is the healthiest.) Using the strong Bayes framework, we get the generalization behavior shown in Figure 13. We will explain this below.

3.1 Likelihood

We can represent a rectangle hypothesis as $h = (\ell_1, \ell_2, s_1, s_2)$, where $\ell_i \in [-\infty, \infty]$ are the coordinates of the upper right, and $s_i \in [0, \infty]$ are the lengths of the two sides. If we assume each example is independently sampled from the concept, the likelihood is given by

$$p(X|h) = 1/|h|^n \text{ if } \forall i.x_i \in h \quad (12)$$
$$= 0 \text{ otherwise} \quad (13)$$

where $|h| = s_1 s_2$ is the size of the rectangle. If we have negative examples, we simply set $p(X|h) = 0$ if h covers any of them.

3.2 Prior

Since one may have many different kinds of prior belief, the definition of $p(h)$ is subjective. We will proceed to make a variety of assumptions, mostly to simplify the mathematics. However, we will see that this results in qualitatively sensible conclusions.

First let us assume the prior factorizes as follows

$$p(h) = p(\ell_1)p(\ell_2)p(s_1)p(s_1) \quad (14)$$

We will assume $p(\ell_i) \propto 1$; this is called an **uninformative** or **uniform** prior, since we have no particular preference where the coordinates of the upper right occurs. (This is an example of a **translation invariant prior**.) Note also that this is an **impproper prior**, since it does not integrate to 1.

We might try to use a uniform prior for the scale, as well:

$$p(s_i) \propto 1 \quad (15)$$

However, Jeffrey's showed that the "right" way to get an uniformative prior about a scale quantity such as s is to use

$$p(s_i) \propto 1/s_i \quad (16)$$

This is called a **scale invariant prior**. We will explain this later.

An alternative is to use an **informative prior**. For scale parameters, it is common to use the **Gamma** distribution

$$Ga(s|\alpha, \beta) \propto s^{\alpha-1} e^{-s/\beta} \quad (17)$$

where α controls the shape and β controls the scale. If we know the expected size σ of the scale parameter, and that is all we know, then the **principle of maximum entropy** says the prior should have the form

$$p(s) \propto e^{-s/\sigma} = Ga(s|\alpha = 1, \sigma) \quad (18)$$

This is called an **exponential prior**. If we know a typical size σ and that sizes much smaller ($s \approx 0$) or larger ($s \gg \sigma$) are unlikely, then we should use an **Erlang density**

$$p(s) \propto s e^{-s/\sigma} = Ga(s|\alpha = 2, \sigma) \quad (19)$$

If we consider the limit $\alpha \to 0$, $\sigma \to \infty$, we recover the uninformative prior

$$p(s) \propto 1/s = Ga(s|0, \infty) \quad (20)$$

See Figure 14.

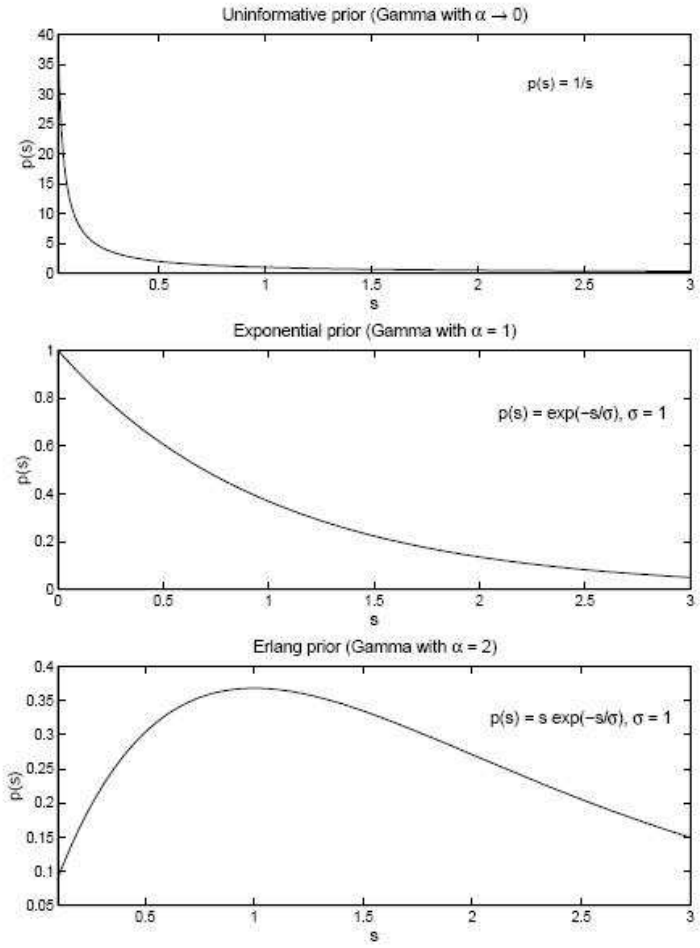


Figure 14: Some gamma distributions.

3.3 Posterior

The posterior is given by

$$p(h|X) = \frac{p(X|h)p(h)}{p(X)} \tag{21}$$

where

$$p(X) = \int_{h'} p(X|h')p(h')dh' = \int_{h' \in \mathcal{H}_X} p(h')/|h'|^n dh' \tag{22}$$

Similarly, the posterior predictive is given by

$$\begin{aligned} p(y \in C|X) &= \int_{h \in H} p(y \in C|h)p(h|X)dh & (23)\\ &= \int_{h \in H} p(y \in C|h)\frac{p(X|h)p(h)}{p(X)} & (24)\\ &= \frac{\int_{h \in \mathcal{H}_{X,y}} p(h)/|h|^n dh}{\int_{h' \in \mathcal{H}_X} p(h')/|h'|^n dh'} & (25) \end{aligned}$$

It turns out there is a simple closed form expression for this if $n \geq 2$ and if we use the Jeffrey's prior $p(s) \propto 1/s$.

Since we assume a separable prior, $p(\ell_1, \ell_2, s_1, s_2) = p(\ell_1, s_1)p(\ell_2, s_2)$, and since the likelihood also factors across dimensions, we can consider the case of one dimensional "rectangles" (i.e., lines) , and then just multiply the results to get the general case.

Since we assume a translation invariant prior, we can assume an arbitrary maximal value for the examples; suppose we choose 0 to be the maximum. Then the right edge of the rectangle must lie past the data, so $\ell \geq 0$. Also, if r is the range spanned by the examples, then the left most data point is at $-r$, so the left side of the rectangle must satisfy $l - s \leq -r$, where s is size of the rectangle. Hence

$$\begin{aligned} p(X) &= \int_{h \in \mathcal{H}_X} \frac{p(h)}{|h|^n} dh & (26)\\ &= \int_{s=r}^{\infty} \int_{l=0}^{s-r} \frac{p(s)}{s^n} dl ds & (27)\\ &= \int_{s=r}^{\infty} \left[\int_{l=0}^{s-r} \frac{1}{s^{n+1}} dl \right] ds & (28)\\ &= \int_{s=r}^{\infty} \frac{1}{s^{n+1}} [l]_0^{s-r} ds & (29)\\ &= \int_{s=r}^{\infty} \frac{s-r}{s^{n+1}} ds & (30) \end{aligned}$$

Now, using **integration by parts**

$$I = \int_a^b f(x)g'(x)dx = [f(x)g(x)]_a^b - \int_a^b f'(x)g(x)dx \tag{31}$$

with the substitutions

$$\begin{aligned} f(s) &= s - r & (32)\\ f'(s) &= 1 & (33)\\ f'(s) &= s^{-n-1} & (34)\\ g(s) &= \frac{s^{-n}}{-n} & (35) \end{aligned}$$

we have

$$p(X) = \left[\frac{(s-r)s^{-n}}{-n}\right]_r^\infty - \int_r^\infty \frac{s^{-n}}{-n} ds \tag{36}$$

$$= \left[\frac{s^{-n+1}}{-n} + \frac{rs^{-n}}{n} - \frac{-1}{n}\frac{s^{-n+1}}{-n+1}\right]_r^\infty \tag{37}$$

$$= \frac{r^{-n+1}}{n} - \frac{rr^{-n}}{n} + \frac{r^{-n+1}}{n(n-1)} \tag{38}$$

$$= \frac{1}{nr^{n-1}} - \frac{r}{nr^{n-1}r} + \frac{1}{n(n-1)r^{n-1}} \tag{39}$$

$$= \frac{1}{n(n-1)r^{n-1}} \tag{40}$$

To compute the generalization function, let us suppose y is outside the range spanned by the examples (otherwise the probability of generalization is 1). Without loss of generality assume $y > 0$. Let d be the distance from y to the closest observed example. Then we can compute the numerator in Equation 25 by replacing r with $r + d$ in the limits of integration (since we have expanded the range of the data by adding y), yielding

$$p(y \in C, X) = \int_{h \in \mathcal{H}_{X,y}} \frac{p(h)}{|h|^n} dh \tag{41}$$

$$= \int_{r+d}^\infty \int_0^{s-(r+d)} \frac{p(s)}{s^n} dl ds \tag{42}$$

$$= \frac{1}{n(n-1)(r+d)^{n-1}} \tag{43}$$

Hence the posterior predictive is

$$p(y \in C|X) = \frac{\int_{h \in \mathcal{H}_{X,y}} \frac{p(h)}{|h|^n} dh}{\int_{h \in \mathcal{H}_X} \frac{p(h)}{|h|^n} dh} \tag{44}$$

$$= \frac{n(n-1)r^{n-1}}{n(n-1)(r+d)^{n-1}} \tag{45}$$

$$= \frac{r^{n-1}}{(r+d)^{n-1}} \tag{46}$$

$$= \frac{1}{(1+d/r)^{n-1}} \tag{47}$$

For a general y, we replace d with $\tilde{d}$, which is 0 if y is inside the range of values spanned by X, and otherwise is just d, which is the distance of y from the nearest example. Finally, for the 2D rectangle case, we get

$$p(y \in C|X) = \left[\frac{1}{(1+\tilde{d}_1/r_1)(1+\tilde{d}_2/r_2)}\right]^{n-1} \tag{48}$$

where r_i measures he size of the smallest rectangle containing X.

Note that if $n = 1$, this is undefined (since d is undefined). This seems reasonable, since if we have no prior information and only one example, we cannot determine anything about the shape of the rectangles.

Similar results can be obtained using the Gamma prior, but various approximations must be made to get an analytic solution.

3.4 Intuition

Figure 13 plots the predictive distribution using an exponential prior with $\sigma_1 = \sigma_2 =$ half the width of the axes; other priors produce qualitatively similar results. The thick line represents the **decision boundary** $p(y \in C|X) = 0.5$. What we see is that there is a broad gradient of generalization for $n = 1$ (row 1) that rapidly sharpens up to the smallest consistent hypothesis as n increases (rows 2-3).

The reason for this behavior is as follows. The size principle dictates that the smallest enclosing rectangle has the highest likelihood. However, there are many other rectangles that are slightly larger with only slightly smaller likelihood; these all get averaged together to give a smooth generalization gradient. But when we have a lot of data, the larger hypotheses get penalized more, and thus contribute less to the posterior; so the generalization gradient is dominated by the most likely hypothesis.

In Figure 13 we also see that the generalization extends further along the dimension with the broader range r_i of observations (row 4). This is because the generalization function contains the term $\tilde{d}_i/r_i$ in the denominator, so if the range on dimension i is small, then the denominator is big, so $p(y \in C|X)$ is very small unless y falls inside X (in which case $\tilde{d} = 0$). This also follows from the size principle: it would be a suspicious coincidence if the rectangle is large in dimension i but r_i is small.

3.5 Special cases of the Bayesian framework

Figure 13 also plots the predictive distribution of two special cases. The first one, MIN RULE, is just maximum likelihood. By the size principle, the ML rectangle is the smallest rectangle than contains all the positive examples. However, similar results hold for the MAP model. The key missing ingredient is hypothesis averaging. MIN-RULE works well when n is large or r_i is small (tightly clustered examples), since then it provides a good approximation of the strong Bayes model (since the posterior is peaky, so averaging has little effect).

The second method, MAX SIM*, is the weak Bayes model, i.e. it uses the weak sampling likelihood that all consistent examples receive a likelihood of 1 instead of $1/|h|^n$. In this case, with an exponential prior, the generalization function is

$$p(y \in C|X) = \exp(-[\frac{\tilde{d}_1}{\sigma_1} + \frac{\tilde{d}_2}{\sigma_2}]) \tag{49}$$

where $1/\sigma_j$ is a weighting factor for dimension j, and $\tilde{d}_j$ is the distance from y to the nearest positive example along dimension j, or is zero if y is inside the range of examples. (This is like a **nearest neighbor classifier**, but only uses positive examples, and returns a probability rather than a class label.) MAX-SIM* works well when n is small or r_i is large, since then it provides a good approximation of the strong Bayes model. (If n is small, the weak sampling likelihood will be similar to the strong one; if r_i is large, then $1/(1 + d_i/r_i)^{n-1} \approx 1$, which results in the weak Bayes generalization function.)

The question of how to learn the similarity metric (i.e., the weights σ_i) in MAX-SIM* is a standard problem. However, in the strong Bayes framework, it does not matter so much, since these prior terms will be dominated exponentially fast by the likelihood. By Equation **??**, the effective weight of dimension i increases if the distance of y (along dimension i) is small relative to the range r_i.