Bayesian concept learning

Kevin P. Murphy

Last updated September 11, 2006

1 Introduction

This chapter, which is a summary of Josh Tenenbaum's PhD thesis [Ten99], 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.

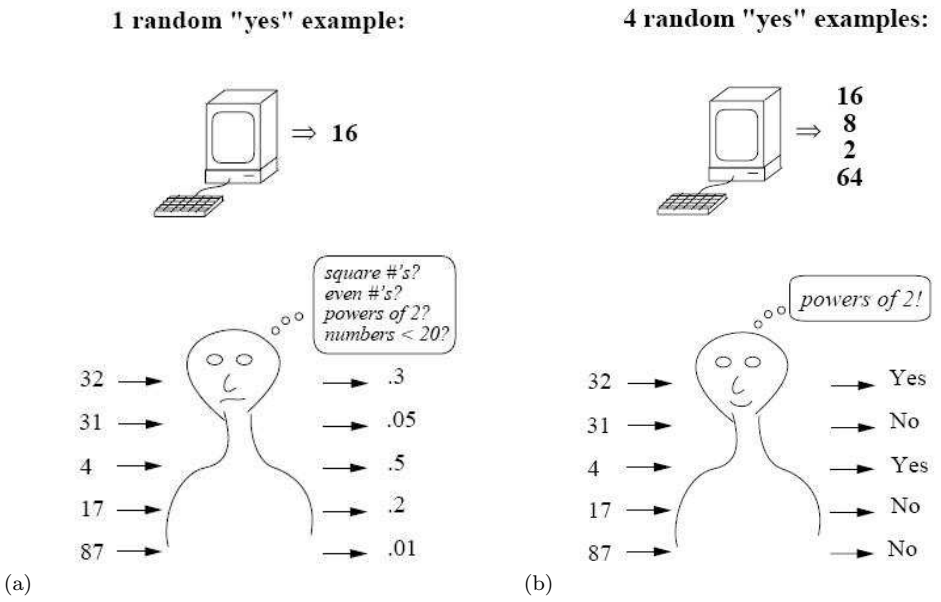


Figure 1: The number game. Belief state after seeing (a) 1 example, (b) 4 examples.

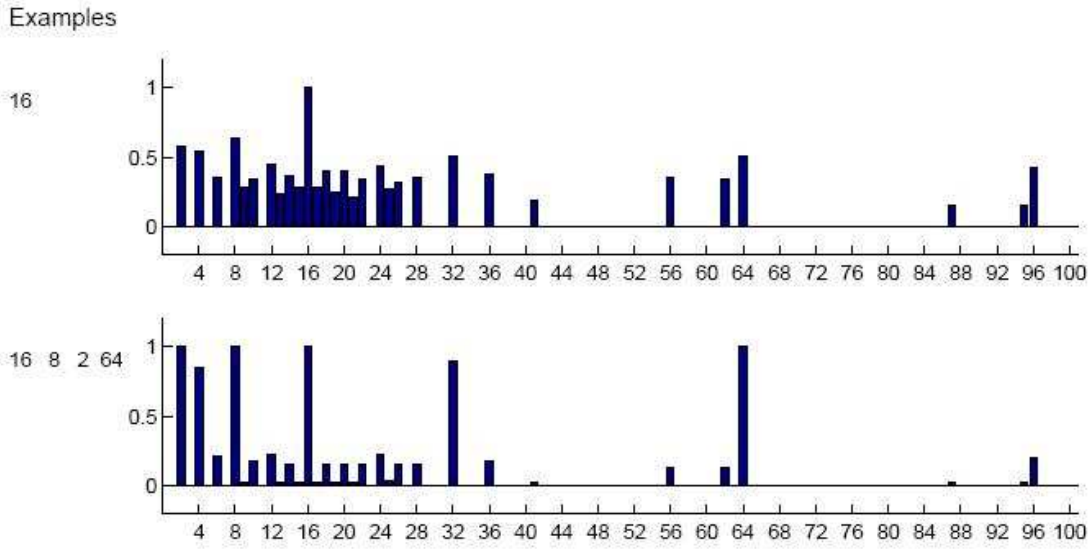


Figure 2: Empirical predictive distribution of humans in the number game. Top: after seeing one example. Bottom: after seeing 4 examples.

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

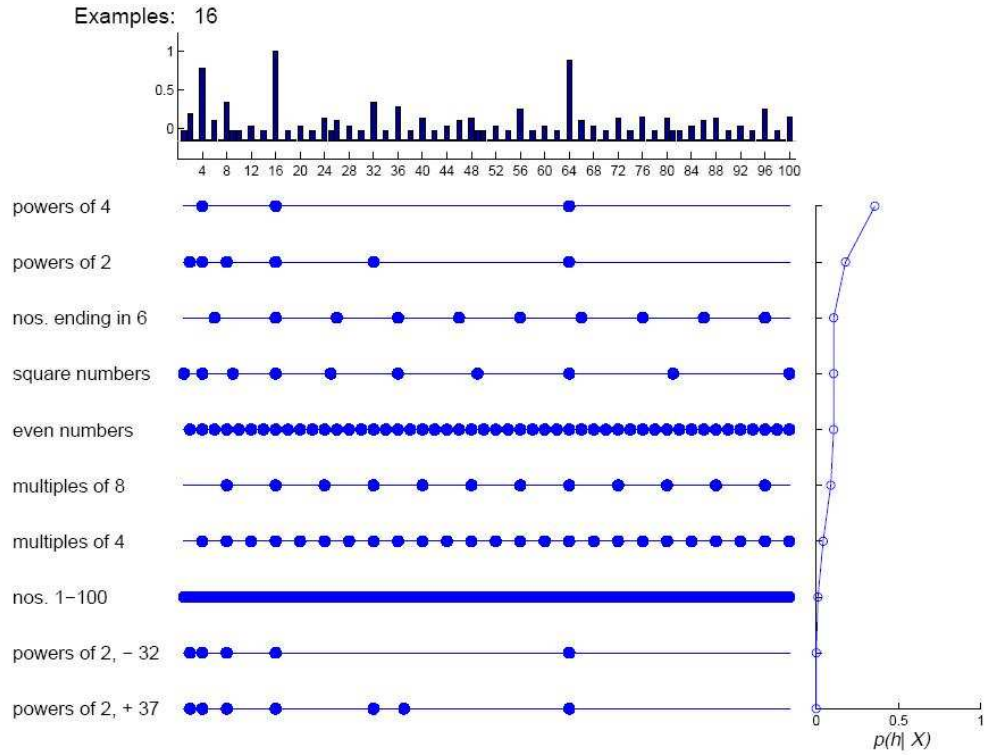


Figure 3: Posterior over hypotheses and the corresponding predictive distribution after seeing one example. A dot means this number is consistent with this hypothesis. The graph $p(h|X)$ on the right is the weight given to hypothesis h. By taking a weighed sum of dots, we get $p(y \in C|X)$ (top).

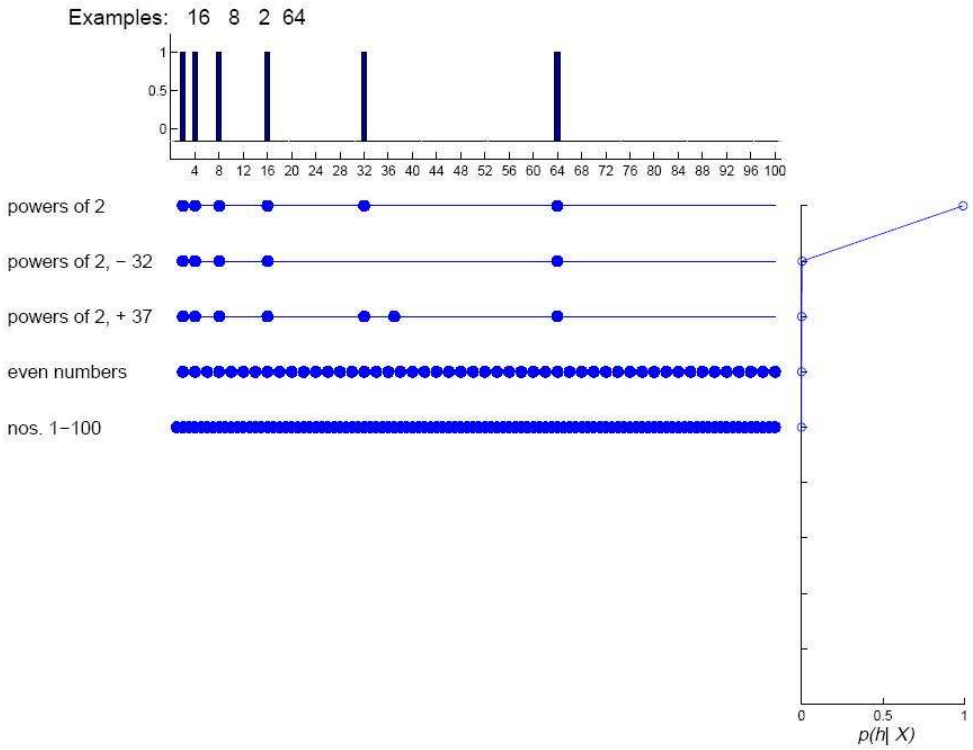


Figure 4: Posterior over hypotheses and the corresponding predictive distribution after seeing four examples.

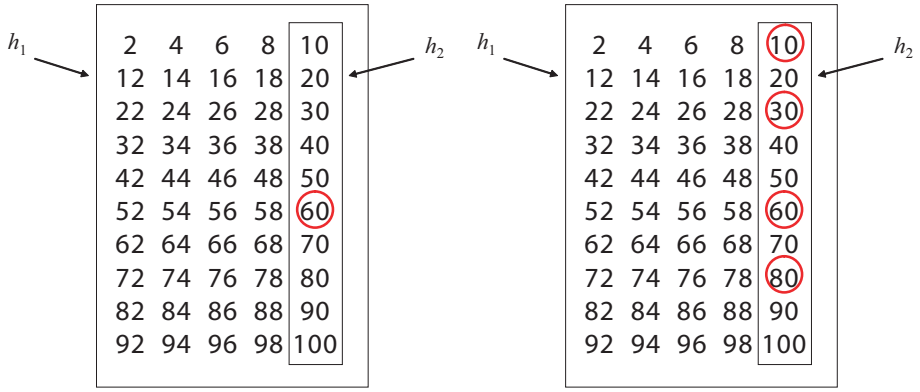


Figure 5: Illustration of the size principle. Consider h_1 = even numbers and h_2 = multiples of 10. If $X = \{60\}$, it is slightly more of a coincidence under h_1; but if $X = \{10, 30, 60, 80\}$, it is much more of a coincidence under h_1, i.e., $p(X|h_1) \ll p(X|h_2)$. Thus the more data we get,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

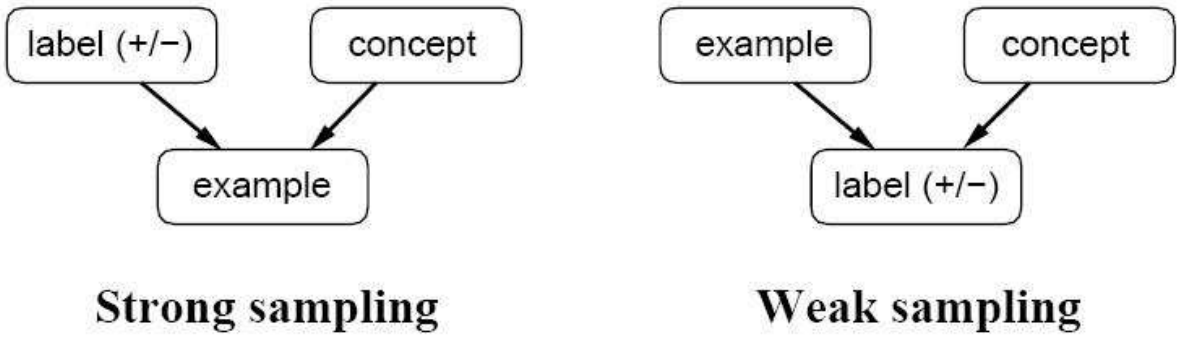


Figure 6: Strong vs weak sampling.

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

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

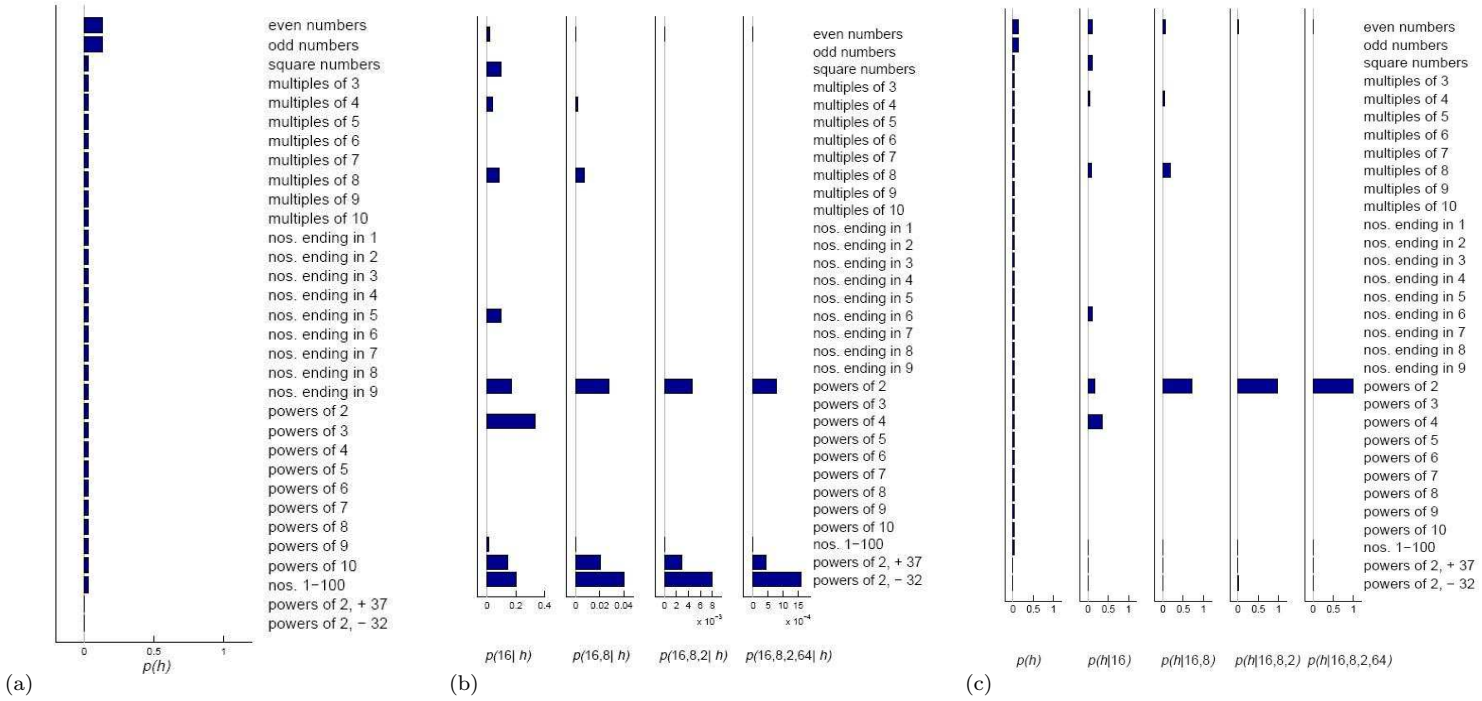
(We will consider more realistic likelihood models, than can handle noise, outliers, etc. later.) This crucial equation is called the **size principle**, and is a form of **Ockham's razor**, which says one should pick the simplest explanation that is consistent with the data. To see how it works, let $X = 16$. 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 [Ten99, p208] for details).

Hypothesis space for number game

Mathematical properites:

- Odd numbers
- Even numbers
- Square numbers
- Cube numbers
- Primes
- Multiples of n: $3 \leq n \leq 12$
- Powers of n: $2 \leq n \leq 10$
- Numbers ending in n: $0 \leq n \leq 9$

Magnitude properties:

- Intervals between n and m: $1 \leq n \leq 100$; $n \leq m \leq 100$

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 6.

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.)

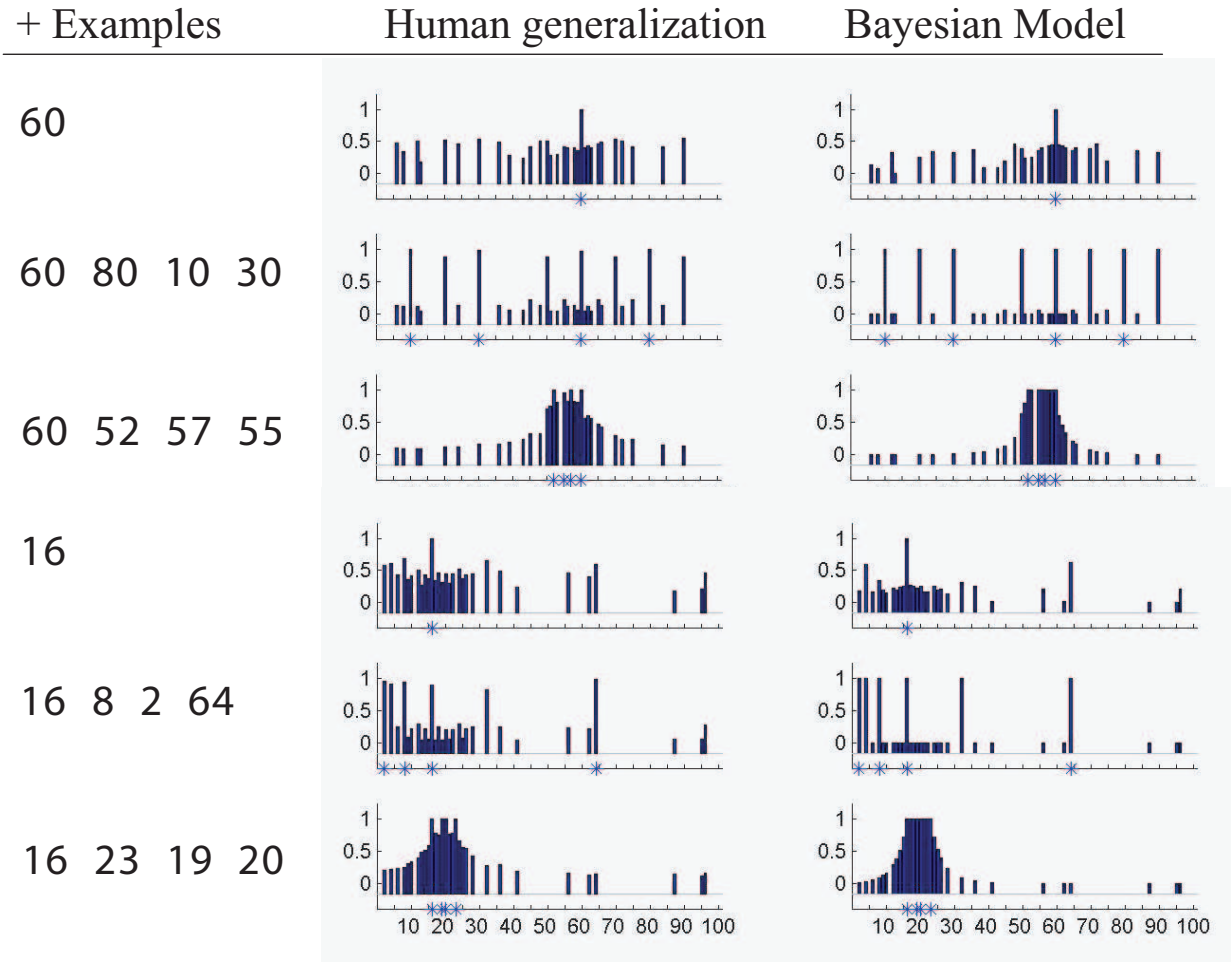


Figure 9: Predictive distributions for people and model using the full hypothesis space. We either get rule-like generalization or similarity-like generalization, depending on which hypotheses have higher posterior probability.

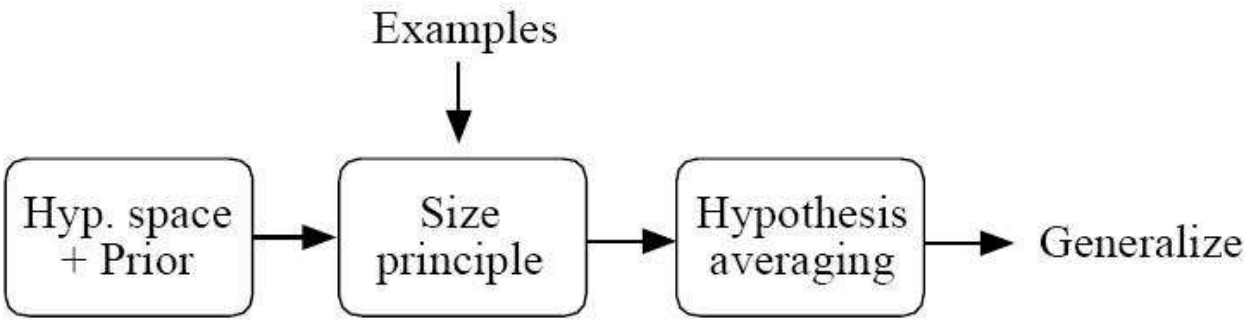


Figure 10: Summary of the Bayesian approach to concept learning.

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 $p(h|X) \approx \delta(h, \hat{h})$, 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 "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 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.

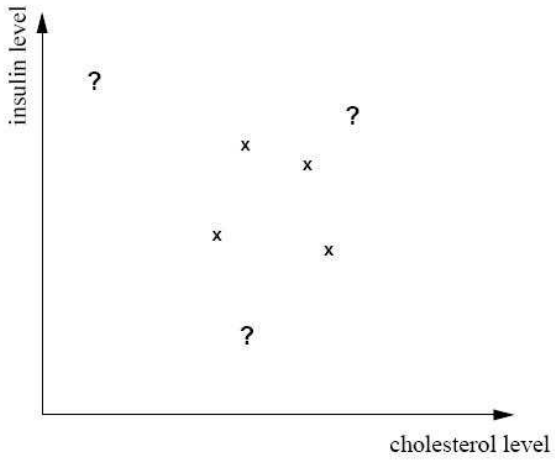


Figure 11: The healthy levels concept

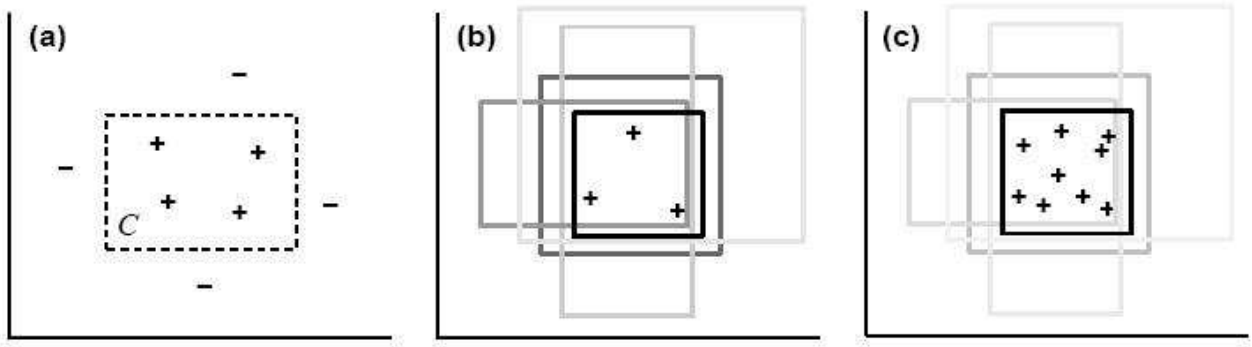


Figure 12: Axes parallel rectangles

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

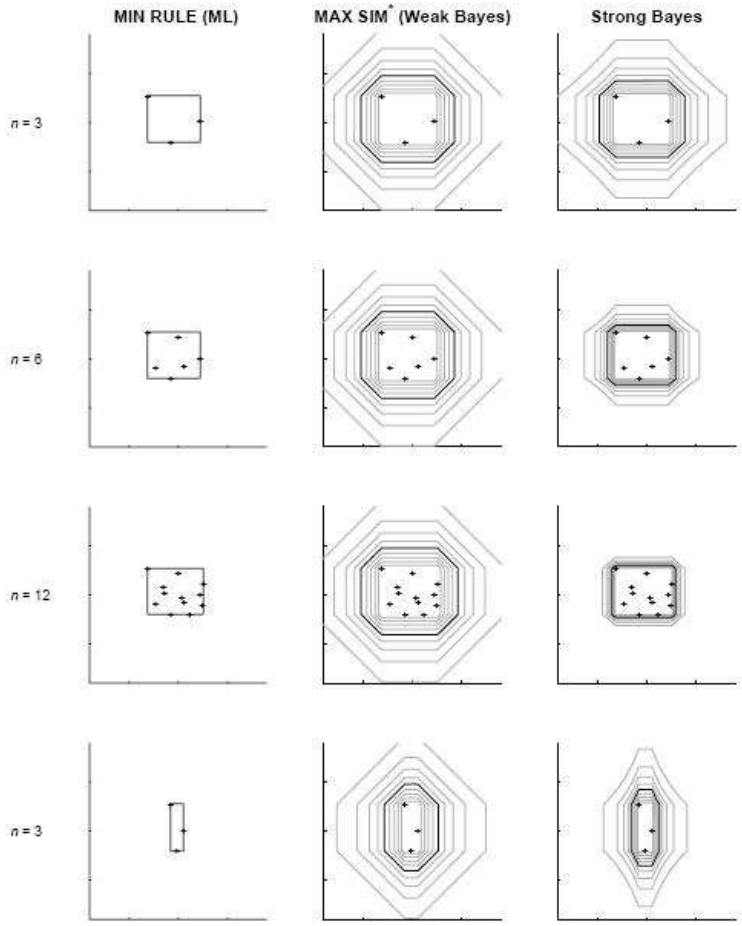


Figure 13: Generalization functions for three different methods on the healthy levels game.

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 called a **translation invariant prior**.

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

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

Note, however, that a uniform prior is an **improper prior**, since it does not integrate to 1. This causes a problem when comparing models (rectangles) of different size. 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**.

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.

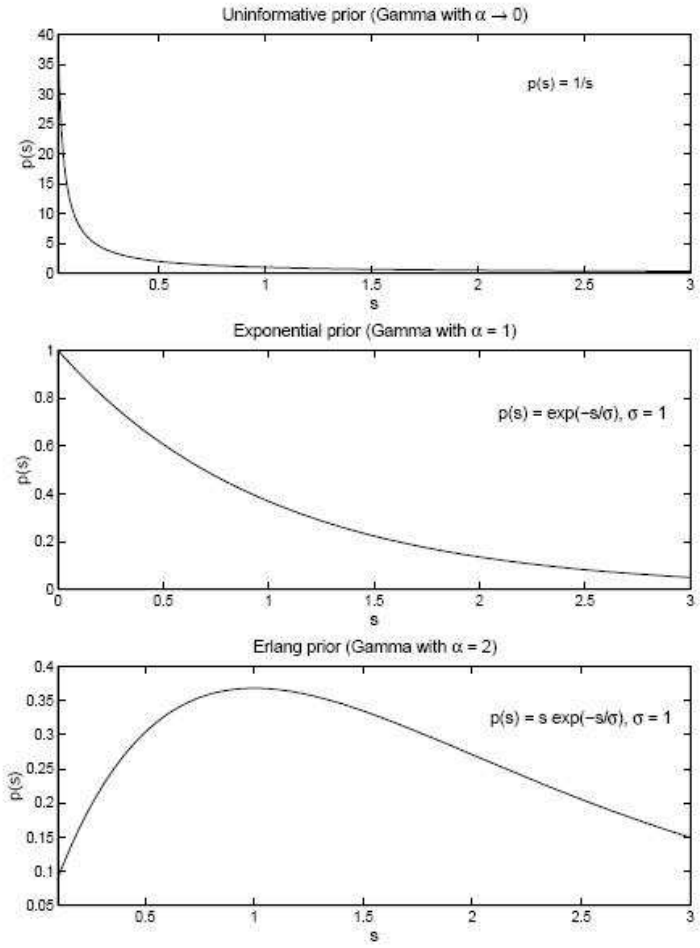


Figure 14: Some gamma distributions.

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.

4 Word learning

Let us return to the original example that motivated the discussion of learning from positive-only data. Suppose, as a child, you see a dog (Rover, a black labrador) run by, and your mother says, "Look, a dog!" What can you infer about the meaning of the word "dog" from this? It could refer to this particular dog (Rover), to all dogs, to all mammals, to all animals, to all labradors, to all black labradors, to all dogs and horses, etc.

People are able to rapidly learn the **extension** of a word from very small numbers of examples. What hypothesis space are they using? When learning the meaning of nouns, it is reasonable to suppose they are

using a **tree-structured hypothesis space**, i.e., they have a **taxonomic bias**. This is because nouns can be arranged into hierarchies, such as subordinate, basic and superordinate level categories.

4.1 Experimental setup

To study this in the lab, Tenenbaum presented test images of novel objects, such as Figure 15, to adult humans, and asked them to predict which other objects belonged to the same category. When presented with a single subordinate level object, such as a green pepper, people tended to generalize to other objects at the same subordinate level in the hierarchy (i.e., other green peppers), and, to a lesser extent, to objects at the next level up (i.e., the basic level category of all peppers), and occasionally two levels up (i.e., the superordinate level category of all vegetables), but never to objects on a different branch of the tree (the set of all test objects is in Figure 17). Thus there appeared to be a gradient of generalization based on similarity.

However, when presented with three examples from a given level (e.g., three same colored peppers, three differently colored peppers, or three different vegetables), people did not generalize above the level of the given object. For example, when shown three green peppers, and told that they are called "pogs", people thought pog meant green pepper, not any kind of pepper; similarly, when shown three diffently colored peppers, they thought pog meant any kind of pepper, but not any kind of vegetable. Thus the generalization behavior appeared to be rule-like: they interpet the word to be the most specific category in the hierarchy that contains all the examples. See Figure 16.

4.2 Bayesian analysis

It should be clear that this transition from similarity-based to rule-like generalization behavior can be modeled by a Bayesian model, using the strong sampling assumption and a suitable prior. To test this empirically, Tenenbaum asked the subjects to estimate the similarity between all pairs of 39 objects (13 animals, 13 vegetables, and 13 vehicles). He then applied **average-link agglomerative clustering** to this **similarity matrix**, with the result shown in Figure 18. The dark-ringed nodes correspond to concepts used in the experiment (e.g., 34 = animals), but other nodes represent plausible concepts, too (e.g., 33=mammals, 36=toys). Let us assume that the hypothesis space is that the extension of a word corresponds to a node in this tree, and all the nodes below it. (We do not want to specify the tree by hand, since different people may have different ideas about what words represent **natural kinds**, especially if they have different mother tongues.)

Next we need to define the likelihood and the prior. The likelihood should be $p(X|h) = 1/|h|^n$. Intuitively, we can define the size of a hypothesis to be one minus its average similarity; thus small classes have high similarity and vice-versa. For example, the class of green peppers is smaller than the class of all peppers, since the similarity between green peppers is higher than between any pair of random peppers. Agglomerative clustering always puts the most similar groups at the leaves, so the higher up in the tree, the lower the within-class similarity, the larger the class, and hence the lower the likelihood: See Figure 19. Hence we define

$$p(X|h) = \left[\frac{1}{\text{height}(h)}\right]^n \tag{50}$$

or $p(X|h) = 0$ if any $x \in X$ is not in h. Hence all classes above the smallest consistent hypothesis in the tree will have non-zero likelihood, and hence will be given some probability mass.

Note that n is the number of *distinct individuals* we have seen with the given name. Seeing the same object being called "pog" on three separate occasions is not equivalent to seeing three separate objects, each being called "pog". Also, note that the size principle, implicit in the above equation, assumes that the teacher is giving random examples of the concept in question (the strong sampling assumption), as opposed to naming something given to it by some other process (such as the learner's asking "what is this"?). For the latter, one should use a weak sampling model: $p(X|h) = 1$ if X is consistent with h, and 0 otherwise.

For the prior, it seems reasonable that nodes (classes) whose members are much more similar to each other than to objects outside the class should be reasonable candidates for being named with words. In

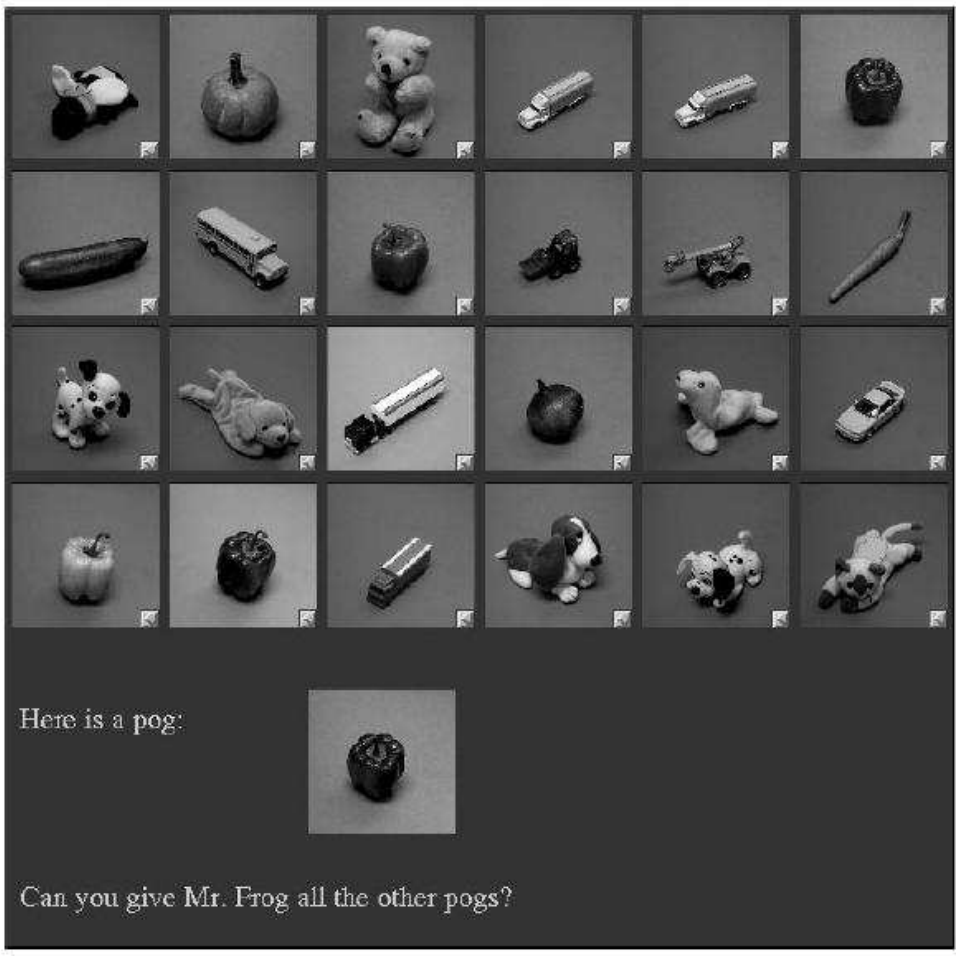


Figure 15: The word learning game.

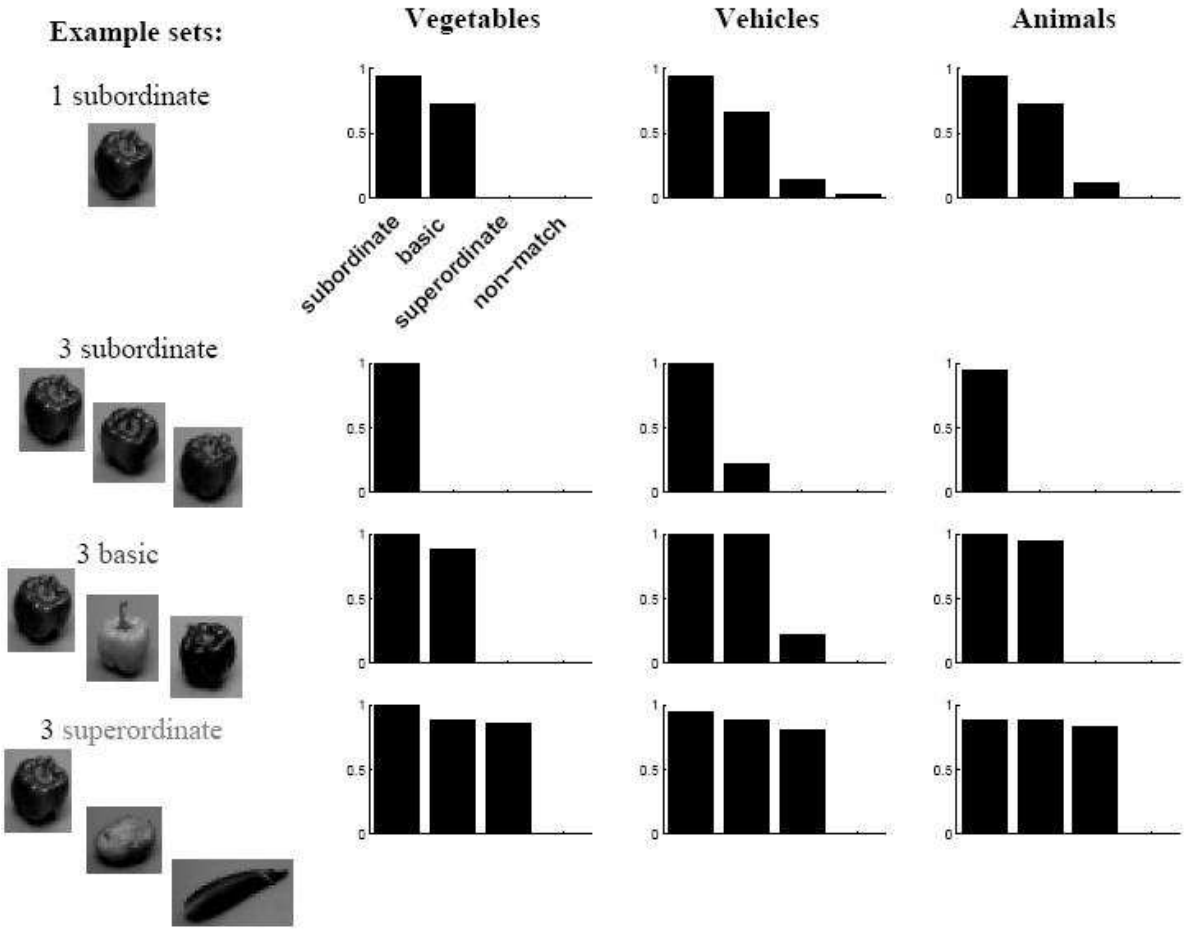


Figure 16: Human generalization performance on the word learning game.

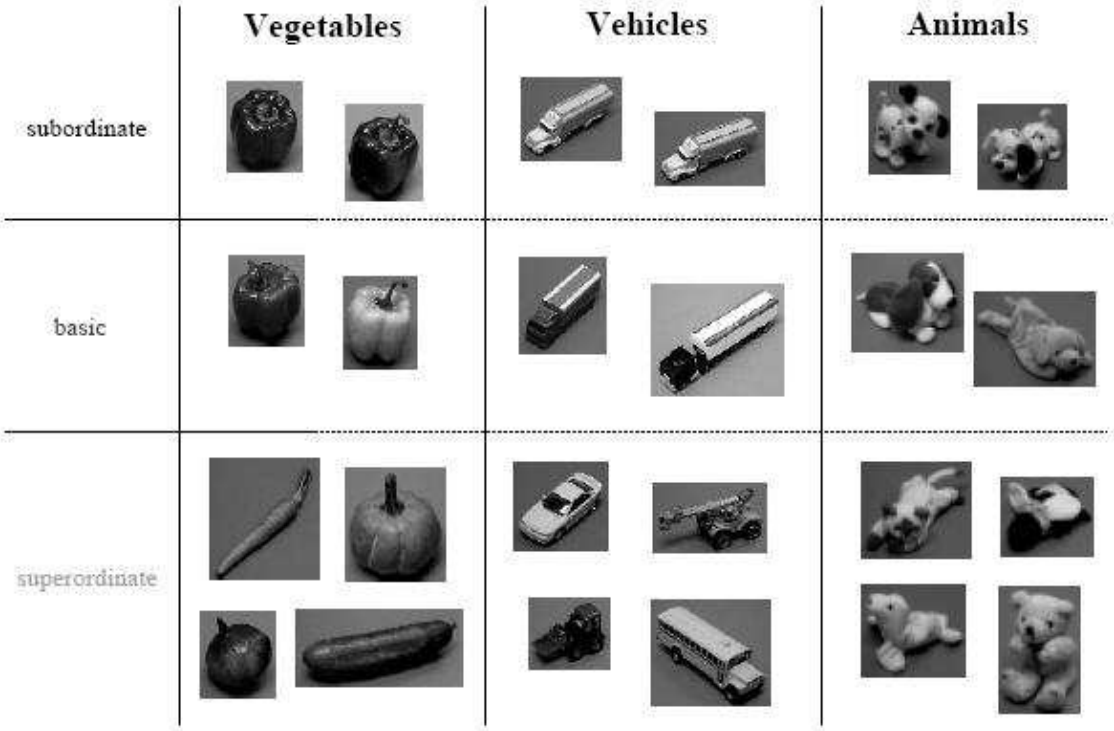


Figure 17: The 24 test objects arranged hierarchically.

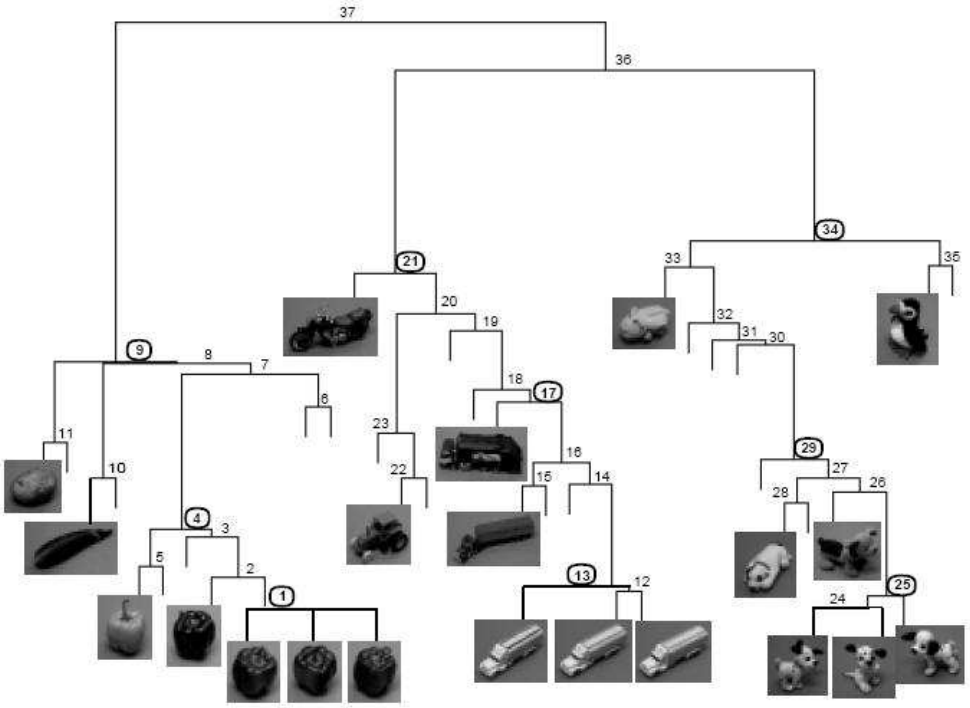


Figure 18: A hierarchical clustering inferred from the empirical similarity matrix.

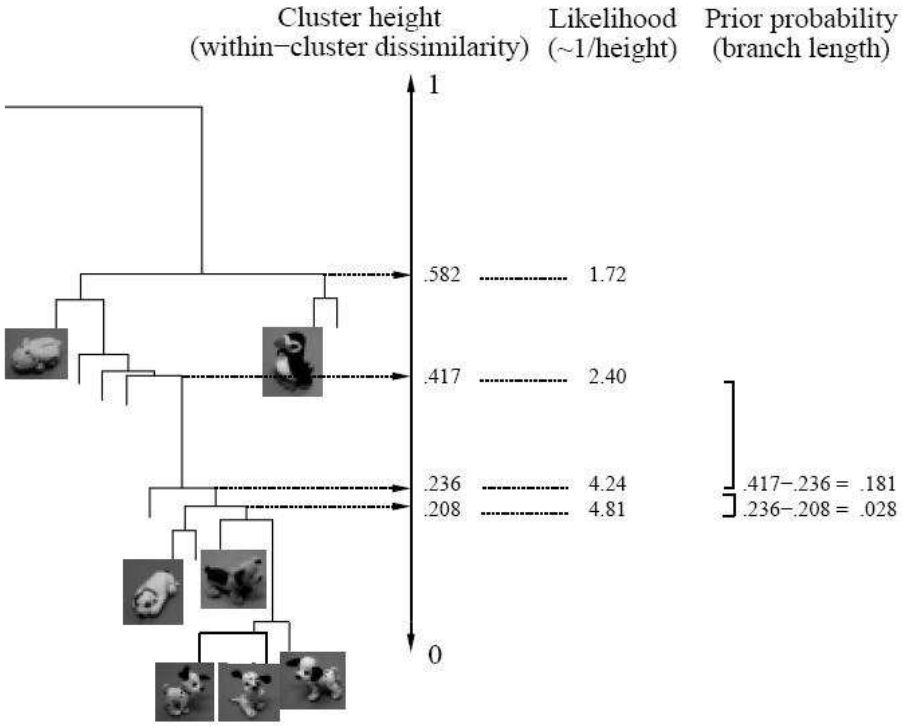


Figure 19: An illustration of how the prior and likelihood are derived from the hierarchical clustering.

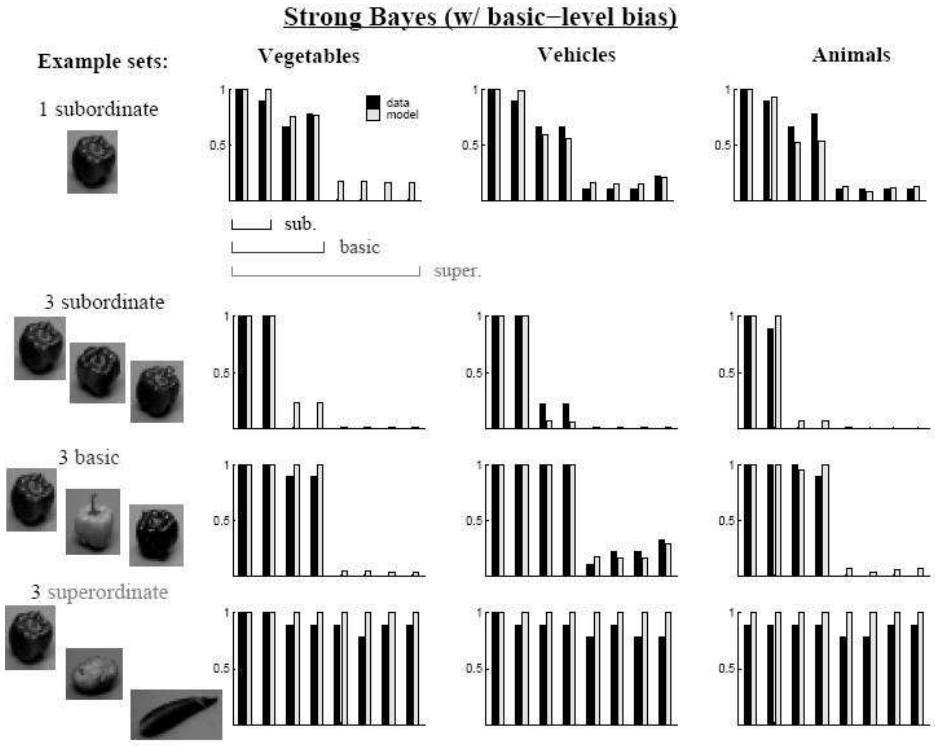


Figure 20: Comparison of human data and model fit for word learning game. The prior has an extra bias towards basic level categories.

contrast, a class whose members are no more similar to each other than to objects outside the class seems like just another random collection of objects, not worthy of a name. The distinctiveness of a cluster can be measured by the length of the branch connecting it to its parent: See Figure 19. Hence we define

$$p(h) = \text{height}(\text{parent}(h)) - \text{height}(h) \tag{51}$$

Tenenbaum showed that the model described above was reasonably accurate at predicting people's generalization behavior, in the sense that it could predict what fraction of the time they would generalize to another subordinate, basic or superordinate level category (see Figure 20).

4.3 Relevance to word learning from real data

It is natural to ask how relevant the above model is for learning the meaning of words in the "real world". The biggest deficiency of the model is that it does not explain how to compute the similarity between two objects. Instead, the similarity matrix is assumed to be known, and is used to derive the hypothesis space. Furthermore, when computing the likelihood, we say $p(X|h) = 0$ unless all the test images in X are contained in hypothesis h (in which case the likelihood is $1/|h|^n$); however, the mechanism for determining if X is within h is to see if X was contained in the cluster for h, but these cluster were created during training.

Nevertheless, one can build a more elaborate Bayesian model in which the classes are arranged in a tree, and each class has an (unknown) distibution over features. There are then two sources of uncertainty. However, this requires more advanced computational machinery, which we will get to later.

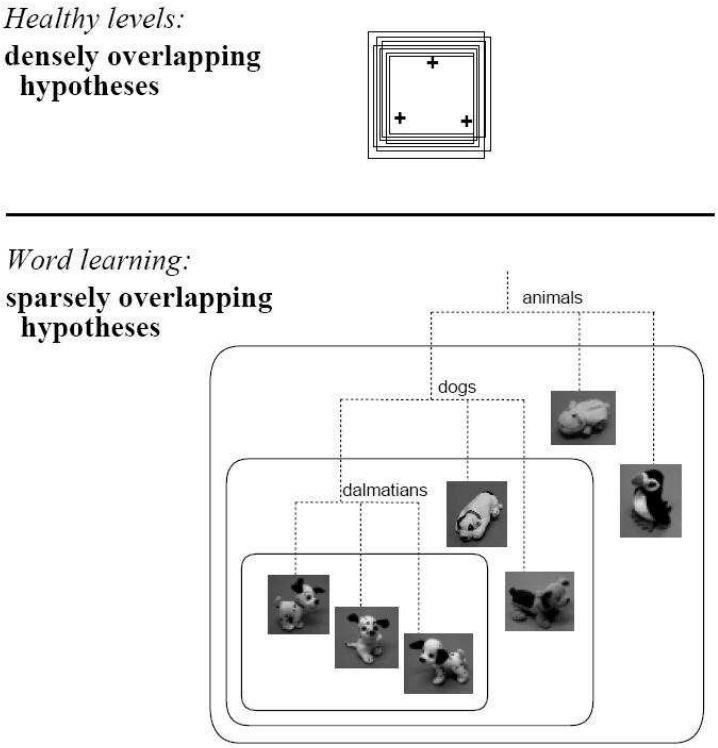


Figure 21: In the healthy levels task, the hypotheses are densely overlapping. In the word learning task, the hypotheses are nested and are only sparsely overlapping.

5 Rules vs similarity

What is remarkable about the word learning experiments is how fast people learn: after only 3 examples, they seem to have an "aha" moment, and figure out the rule that defines the concept. By contrast, in the healthy levels domain, it takes many tens of examples until the boundary that defines the concept becomes sharp.

The reason for this qualitatively different behavior has to do with differences in the two hypothesis spaces. In the healthy levels case, there are many densely overlapping hypotheses. Thus for any given data set, there are many rectangles with very similar size (and hence likelihood), all of which will be averaged over, to produce a smooth generalization gradient. (Negative examples can rapidly define the boundaries of the concept, however.)

In the word learning case, the hypotheses are only sparsely overlapping (see Figure 18). For any given data set, there are only a few hypotheses that cover the data, and they differ dramatically in size: each hypothesis is usually much smaller (and hence more likely) than its parent (especially since the prior favors well-separated classes). Thus the smallest consistent hypothesis becomes exponentially more probable, resulting in rule-like behavior.

In the number game, the hypothesis space is a mixture of both sparsely overlapping hypotheses (mathematical concepts) and densely overlapping hypotheses (interval concepts). Hence whether the generalization is similarity-like or rule-like depends on which hypotheses get "activated" by the data. This is illustrated in Figure 9, where we see that if $X = \{16, 8, 2, 64\}$, people (and the model) generalize in a rule-like way (the concept "powers of two" being by far the most probable), whereas if $X = \{6, 23, 19, 20\}$, people (and the model) generalize in a similarity-like way (since several concepts, corresponding to intervals around 6–23, all have posterior support).

6 More realistic models

There are various ways in which we can extend the models to handle such things as noisy data, concepts with two or more extensions (disjunctive concepts), and uncertainty about the relevant features. We sketch a few of these extensions below, in the contetx of the healthy levels game.

The basic approach is to add extra **hidden (latent) variables** to the model, to capture extra sources of uncertainty (orthogonal to uncertainty about the definition of the concept), and then to **marginalize them out**:

$$p(h|X) = \frac{p(X|h)p(h)}{p(X)} \tag{52}$$

$$= \frac{\sum_z p(X, Z=z|h)p(h)}{p(X)} \tag{53}$$

$$= \frac{\sum_z p(X|h,z)p(z)p(h)}{p(X)} \tag{54}$$

where h is the unknown hypothesis (concept) and z is the latent variable. (We have assumed that h and z are conditionally independent.) (In realistic situations, this marginalization can be quite expensive, but we leave discussion of efficient computational techniques until later.) Given the posterior, we can compute the generalization function in the using

$$p(y \in C|X) = \sum_h p(y \in C, h|X) \tag{55}$$

$$= \sum_h p(y \in C|h)p(h|X) \tag{56}$$

or

$$p(y \in C|X) = \sum_z \sum_h p(y \in C, h, z|X) \tag{57}$$

$$= \sum_z p(y \in C|z, X)p(z|X) \tag{58}$$

6.1 Handling outliers

Consider the situation in Figure 22, where one data point (top right) is far from the cluster of the others. There are (at least) two possible explanations: either the concept is a large rectangle (shown in Figure 22(a)), or the data point in question is an **outlier**, and was not in fact generated from the concept (see Figure 22(b)). The latter explanation may be more probable, depending on the number and location of the outliers, and the prior probability of an outlier.

Let us partition the data X into inliers, X_{in}, and outliers, $X_{out} = X \setminus X_{in}$. Let z denote this partition; hence it has $2^n - 1$ possible values (assuming we are not going to reject every example as an outlier). Alternatively, z could be represented as a vector of n bits, where $z_i = 1$ means x_i is an outlier, otherwise $z_i = 0$; this is called a **1-of-n (distributed) encoding**.

Let us assume the outliers are generated independently from the inliers. Suppose outliers are generated uniformly at random with probability ϵ over a region of size L (the maximum possible range of values). Then the likelihood becomes

$$p(X_{out}|z) = \frac{1}{L}^{n_{out}} \tag{59}$$

and the prior is

$$p(z) = \epsilon^{n_{out}} \tag{60}$$

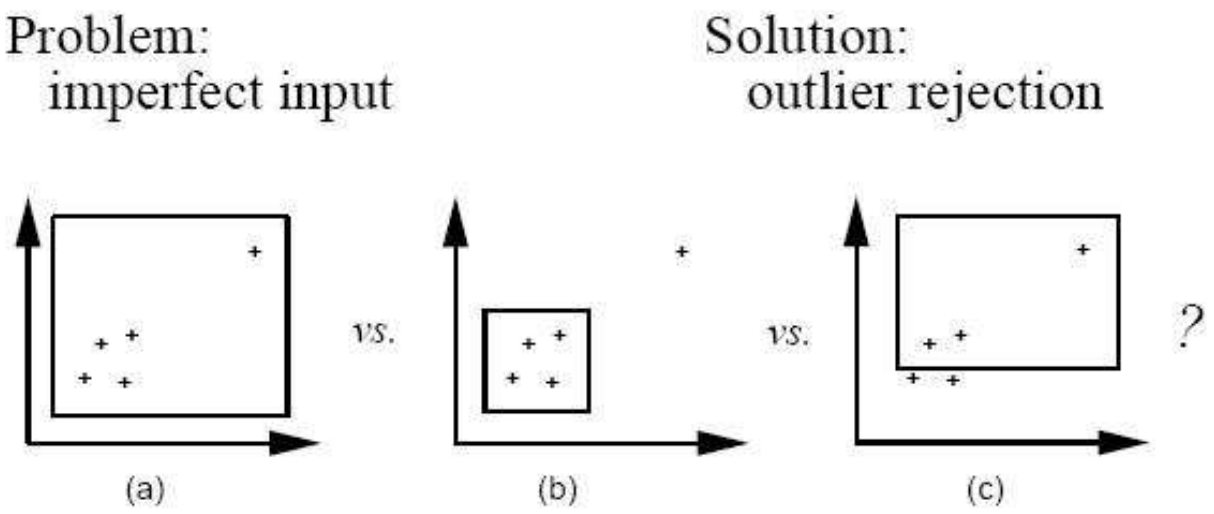


Figure 22: Rectangle hypotheses where data may contain outliers.

If we assume an uninformative (scale-invariant) prior for the rectangles, then the (marginal) likelihood of the inliers is a slight modification of Equation 40:

$$p(X_{in}|z) = \int p(X_{in}|h,z)p(h)p(z)dh \tag{61}$$

$$= \frac{1}{n_{in}(n_{in}-1)r_{in}^{n_{in}-1}} \tag{62}$$

where n_{in} is the number of inliers specified by z, and r_{in} is the size of the smallest interval containing all inliers. (Here we work in 1D for simplicity.) This equation is not defined if $n_{in} = 1/\ r_{in} = 0$. One solution is to use an informative prior. However, the solution adopted by Tenenbaum is to use $p(s) \propto 1/s^{1+\beta}$; the limit as $\beta \to 0$ is uninformative. In addition, he assumes some observation noise α, so the data gets "fattened out" at the edges by this amount. This amounts to replacing n_{in} by $n_{in} + \beta$, and r_{in} by $r_{in} + \alpha$, Tenenbaum sets α to 1% of the data range and $\beta = 0.2$. (This is an example of **empirical Bayes**, since the hyperparameters are being set based on the data.)

The posterior over z is given by

$$p(z|X) \propto p(X|z)p(z) \tag{63}$$

$$= p(X_{in}|z)p(X_{out}|z)p(z) \tag{64}$$

$$\propto \left(\frac{1}{r_{in}}\right)^{n_{in}-1} \left(\frac{\epsilon}{L}\right)^{n_{out}} \tag{65}$$

This equation makes intuitive sense. If we reject any example as an outlier, then n_{out} decreases, which decreases the second term, but if this rejection significantly reduces r_{in}, then the rejection inceases the first term, leading to an overall improvement. This is why the upper right point in Figure **??**(b) is more likely to be an outlier than any of the others (such as the bottom two in Figure **??**(c)).

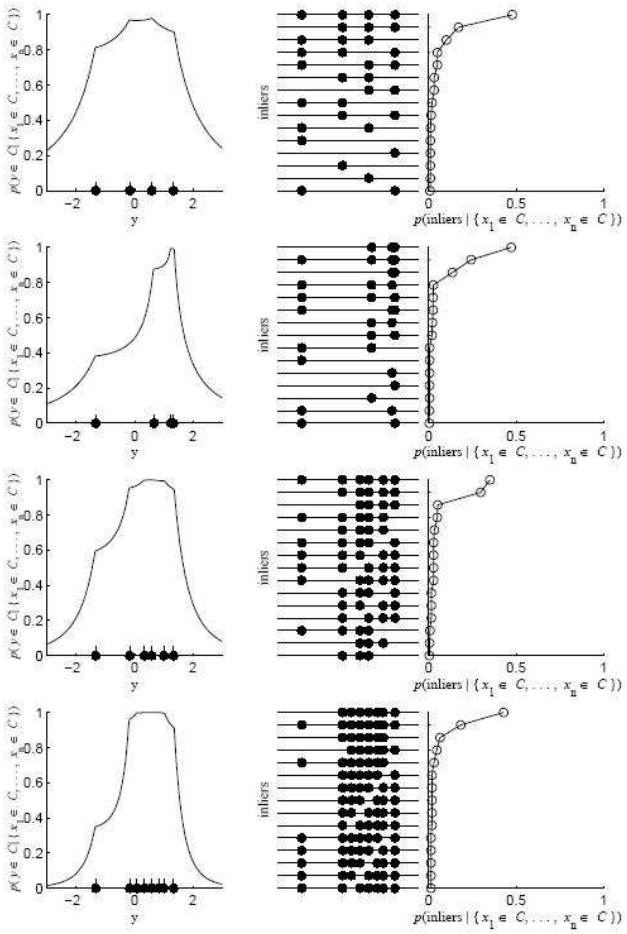


Figure 23: Generalization function for rectangle hypotheses where data may contain outliers.

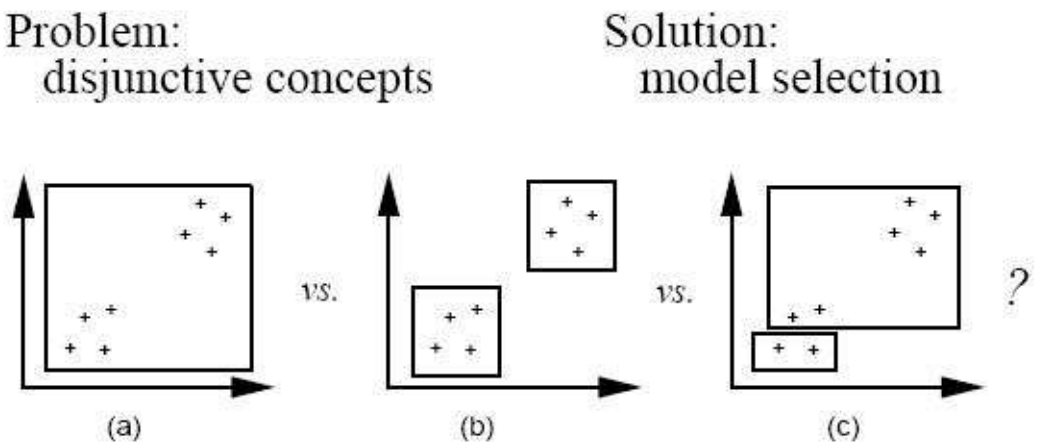


Figure 24: Does the data come from one rectangle or two?

Having computed the posterior over z, the generalization function is

$$p(y \in C|X) = \sum_z \sum_h p(y \in C, h, z|X) \quad (66)$$

$$= \sum_z p(y \in C|z, X)p(z|X) \quad (67)$$

$$= \sum_z p(y \in C|X_{in(z)})p(z|X) \quad (68)$$

where

$$p(y \in C|X_{in(z)}) = \frac{1}{(1 + \tilde{d}/r)^{n-1}} \quad (69)$$

where d,r and n are functions of z. (Recall that $\tilde{d}$ is the distance of y to the closest point in X_{in}, or 0 if y is inside the data.)

Figure 23 shows some examples of the generalization function for different sets X (working in 1D for simplicity). If most of the data is tightly clustered, but some points are well separated from the rest, the latter points will be detected as outliers, and will have little impact on the prediction performance. This figure also shows the 15 most probable inlier sets; most of these have neglibible probability beyond the first four of five. Furthermore, the top sets are all quite similar, suggesting that the average over all possible sets z can be well approximated by **Monte Carlo sampling**, or even by maximum likelihood.

6.2 Model selection for disjunctive concepts

Consider Figure 24 where some data is drawn from one extension of the concept, X_1, and other data is drawn from another extension, X_2. This could arise when learning a word with two distinct senses (polysemy), for example.

Let z denote a partition of X into X_1 and X_2. Hence z can take on $2^n - 1$ possible values. (Without loss of generality, we will assume $n_2 > 0$, so only X_1 can be empty.) Using the fact that $p(A \vee B) = p(A) + p(B) - p(A \wedge B)$, for events A and B, the generalization function is

$$p(y \in C|X) = \sum_z p(y \in C|z, X)p(z|X) \quad (70)$$

$$= \sum_z [p(y \in C|X_1(z)) + p(y \in C|X_2(z)) - p(y \in C|X_1(z))p(y \in C|X_2(z))]p(z|X) \quad (71)$$

If we assume the examples are drawn from each split independently, we have

$$p(z|X) \propto p(X|z)p(z) \tag{72}$$
$$\propto p(X_1|z)p(X_2|z)p(z) \tag{73}$$

If we use an uninformative (scale-invariant) prior for the rectangles, then the marginal likelihood (in 1D) is

$$p(X_1|z) = \int p(X_1|h,z)p(h)dh \tag{74}$$
$$= \frac{1}{n_1(n_1-1)r_1^{n_1-1}} \tag{75}$$

and similarly for $p(X_2|z)$. Again, when $n_1 = 1$, we can use the α/β fix as before. But if $n_1 = 0$, we are comparing two models of different dimensionalities: one has two rectangles and the other has one. Model comparison with improper priors can lead to problems (see the discussion of **Lindley's paradox** in Section **??**). Tenenbaum used the approximation $p(X_1|z) = 1/L$ if $n_1 = 0$ as a simple alternative to exact computation of the Bayes factor.

We can set the prior for z as follows. Let $p(z = (0,\ldots,0)) \propto \kappa$ (the assignment such that $n_1 = 0$, so all points are assigned to X_2), and let all other assignments have probability proportional to $1-\kappa$. Thus κ is our prior belief that there is only one extension. Overall, the posterior is

$$p(z|X) \propto \begin{cases} \kappa\frac{1}{L}\frac{1}{n_2(n_2-1)r_2^{n_2-1}} & \text{if } n_1 = 0 \\ (1-\kappa)\frac{1}{n_1(n_1-1)r_1^{n_1-1}}\frac{1}{n_2(n_2-1)r_2^{n_2-1}} & \text{if } n_1 > 0 \end{cases} \tag{76}$$

This equation makes intuitive sense. If we switch a point from X_1 to X_2, we decrease (or leave unchanged) r_1 but increase (or leave unchaned) r_2. The optimal decision about which group to assign a particular example to is thus determined by which groups's range it will have a greater effect on. This is why the split in Figure 24(b) seems more natural than the one in Figure 24(c). In addition, while the special case of assigning all the examples to one group is the least favored on relative size grounds (since the likelihood becomes $1/L$), it is favored on grounds of simplicity (as controlled by κ). This is why a set of examples must be highly clustered in order to be a plausible candidate for being its own separate extension.

Figure 25 shows the generalization function for different X. The degree of bimodality depends on the support for the existence of two independent extensions, which in turn depends on several factors: the relative ranges spanned by the two clusters of data, the separation between the clusters, and the amount of data. It also shows the 8 most probable values of z. Againt,

6.3 Handling weak prior knowledge

Suppose you know the concept is a rectangle, but you do not its orientation (see Figure 26). Let $z \in [0, \pi/2]$ represent the unknown orientation. Let us use a uniform prior $p(z) = 2/\pi$. Let us assume points are sampled at random from the concept, so the marginal likelihood (marginalizing over the extent of the rectangle) is

$$p(X|z) \propto \frac{1}{(r_z r_{z+\pi/2})^{n-1}} \tag{77}$$

where r_z is the range spanned by the examples when projected onto a line at orientation z. This prefers orientations onto which the data project as small a range as possible.

Note that we can send this likelihood to infinity if n is less than or equal to the dimensionality of the space, e.g., in 2D, if we have $n = 2$ points, we can choose z to lie along the line joining them, so $r_{z+pi/2}$ is infinite. However, the posterior

$$p(z|X) = \frac{p(X|z)p(z)}{\int_z p(X,z)} \tag{78}$$

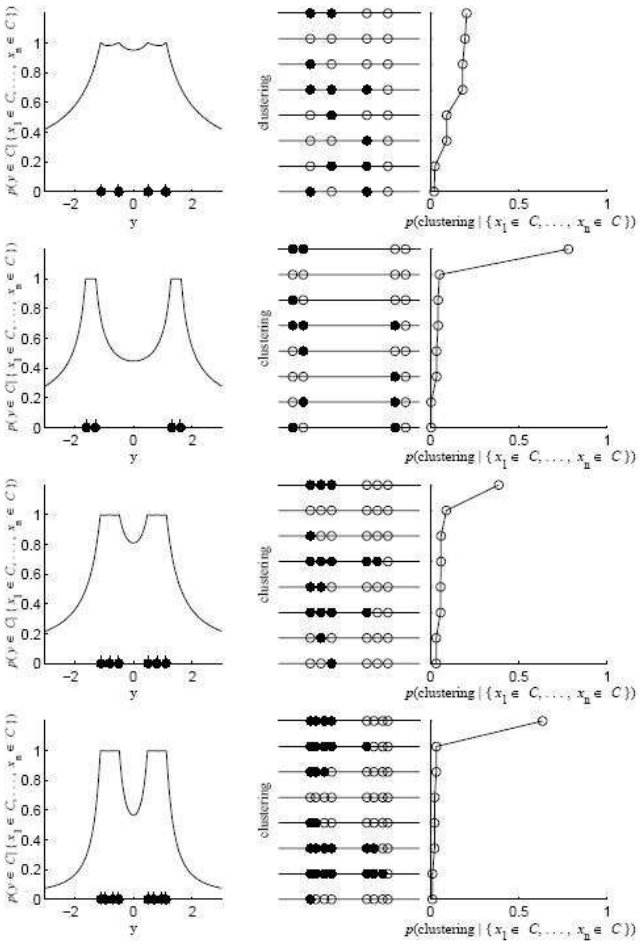


Figure 25: Generalization functions for when the data may come from one rectangle or two. A black dot means this data point came from cluster 1, a hollow dot means it came from cluster 2. For the first row, the top 4 hypotheses (not in order) are $(X_1 = \emptyset, X_2 = \{1,2,3,4\})$ $(X_1 = \{1\}, X_2 = \{2,3,4\})$ $(X_1 = \{1,2\}, X_2 = \{3,4\})$ $(X_1 = \{1,2,3\}, X_2 = \{4\})$. (The hypothesis $(X_1 = \{1,2,3,4\}, X_2 = \emptyset)$ is disallowed.) Thus there is some ambiguity about whether there is one cluster or two (the exact results will depend on the hyperparameters κ and L.) In the other cases, it is clear there are two clusters. Note that we can label the "left" cluster X_1 or X_2; thus there is some **label ambiguity.**

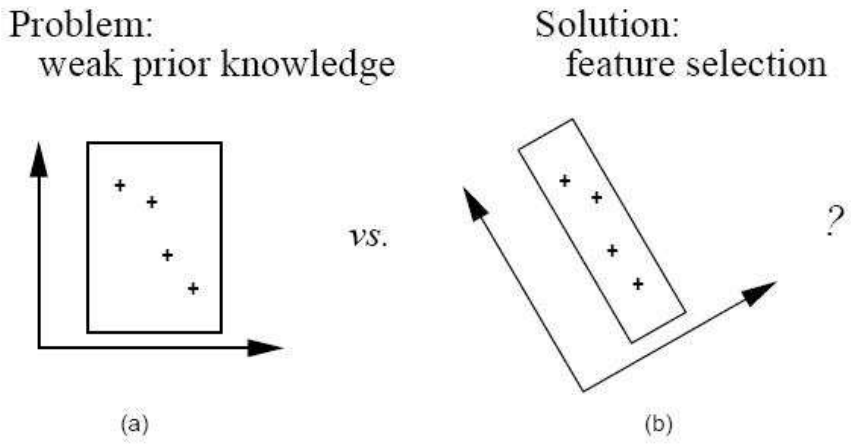


Figure 26: Which axes should we use?

should be finite for all z (since posteriors, unlike likelihoods, must integrate to 1). Similarly, the predictive distribution

$$p(y \in C|X) = \int_z p(y \in C|z, X)p(z|X) \tag{79}$$

should be well behaved.

In this example, since z is continuous, marginalizing out z may be hard analytically. In the 1D case, we can approximate this by discretization. Figure 27 shows some examples. The blocky and sharp-cornered generalization gradients of Section 3 have been replaced by smoother contours, as a consequence of averaging over all possible rectangle orientations. The extent to which the generalization gradients deciate from circularity is a function of how much evidence the data provide for a preferred set of axes, i.e., how tight the examples are clustered on any one direction. (This is related to the number of **principal components** of the data.)

References

[Ten99] J. Tenenbaum. *A Bayesian framework for concept learning.* PhD thesis, MIT, 1999.

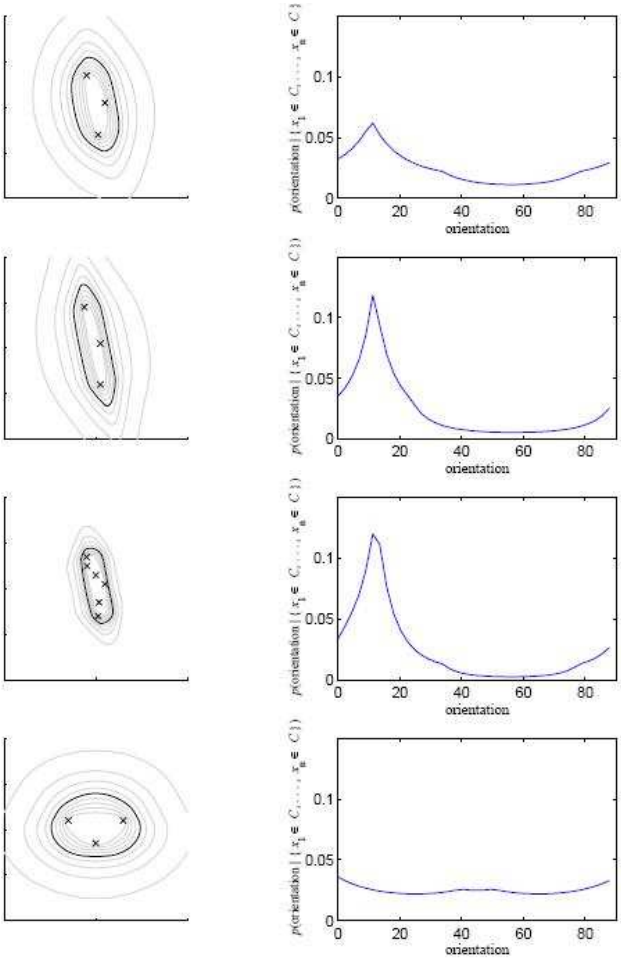


Figure 27: Generalization gradient when the orientation of the rectangle is unknown. We also show $p(z|X)$.