CSCI 5822 Assignment 1

Evan Sidrow

February 1, 2018

1 Prior Distribution

The prior distribution used for each hypothesis class was Tenenbaum's expected-size prior:

$$P(h) \propto \exp\left[-\left(\frac{s_1}{\sigma_1}+\frac{s_2}{\sigma_2}\right)\right]$$

Where s_1 and s_2 are the lengths of the sides of the hypothesis class and σ_1 and σ_2 are parameters related to the spread of the distribution. Using this prior on square hypotheses with side lengths $2i$, $i \in \{1, 2, ..., 10\}$, we get the following:

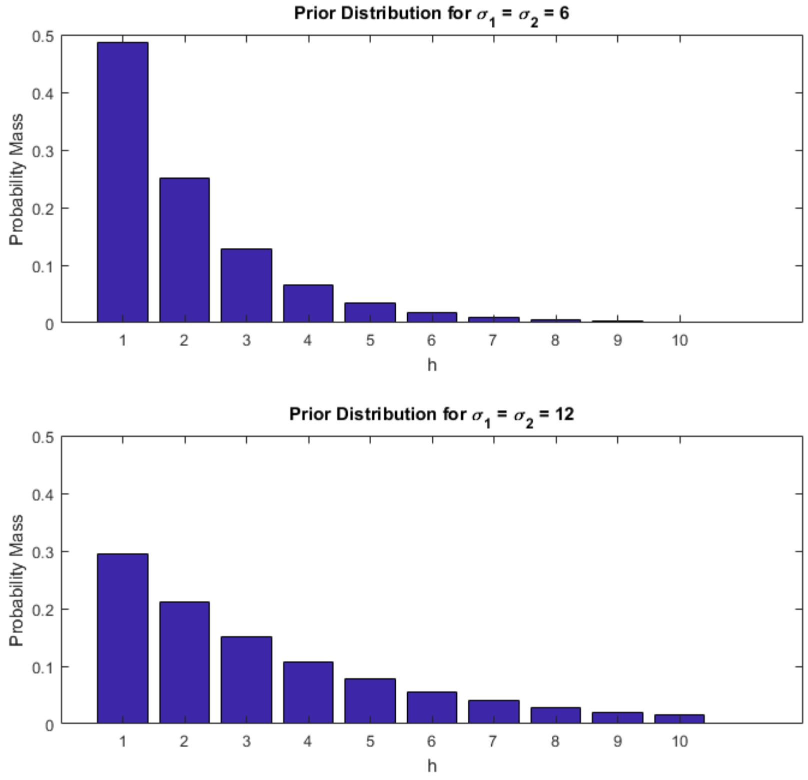

Figure 1: Expected size priors for square hypothesis classes

Note that the distribution becomes more shallow as σ increases. This is because the exponential in the expected size-prior varies less with s when σ is large.

2 Posterior Distribution

The likelihood function used for this assignment was Tenenbaum's Size Principle:

$$P(X|h) = \left(\frac{1}{s_1 s_2}\right)^n * \mathbb{1}_{\forall j (x_j \in h)}$$

Where $\mathbb{1}$ is the indicator function, and X is a vector of n data points.

Once data ($X = \{(1.5, 0.5)\}$) had been observed, the posterior distribution was obtained by multiplying the expected size prior with the likelihood function, as shown below:

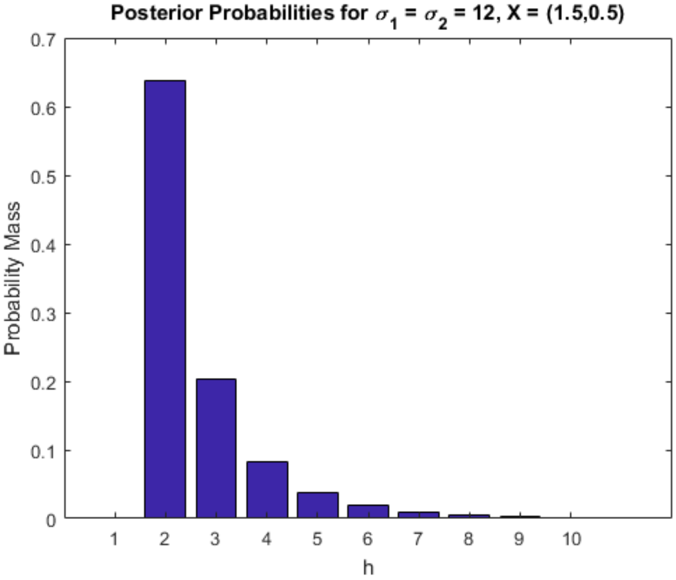

Figure 2: Posterior distribution for square hypothesis classes

Note that the probability of hypothesis class 1 is zero, since the data fell outside of the square associated with it. However, other hypothesis classes with smaller areas became more likely, since the likelihood function is larger when the hypothesis class is smaller.

3 Probability of Falling Inside the Concept

After calculating the total probability of each hypothesis, It is possible to find the probability that a new point y is a member of the concept class. In particular:

$$P(y \in concept|X) = \sum_{y \in h} P(X|h)P(h)$$

Using this equation, along with prior parameters of $\sigma_1 = \sigma_2 = 10$, I created a contour plot in 2-D space of the probability that point taken from that region would be a member of the concept class. The results are shown below:

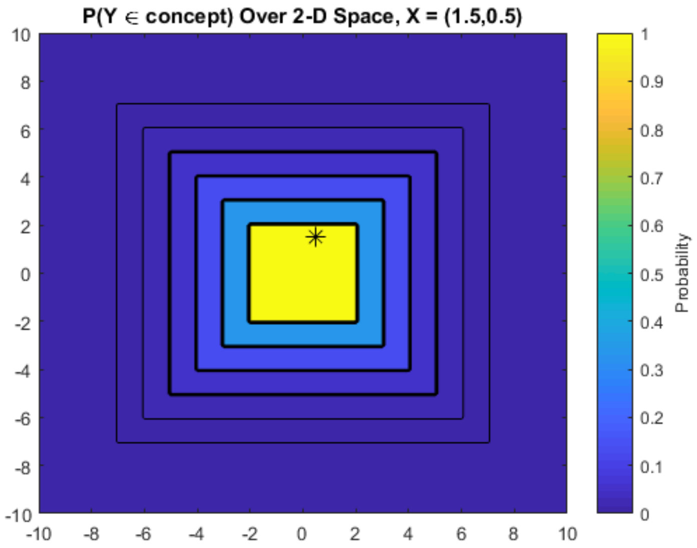

Figure 3: Generalization predictions of concept class membership, X = {(1.5,0.5)}

Note that the probability of any point in the second hypothesis square belonging to the concept is one. This is because h_1 is impossible (the observed point is not in the first hypothesis square) and the second hypothesis square is a subset of all other possible hypotheses.

4 Varying the Observed Data

If we repeat the exercise in task 3 above, but with a new value for X ($X = \{(4.5, 2.5)\}$), we get the following contour plot:

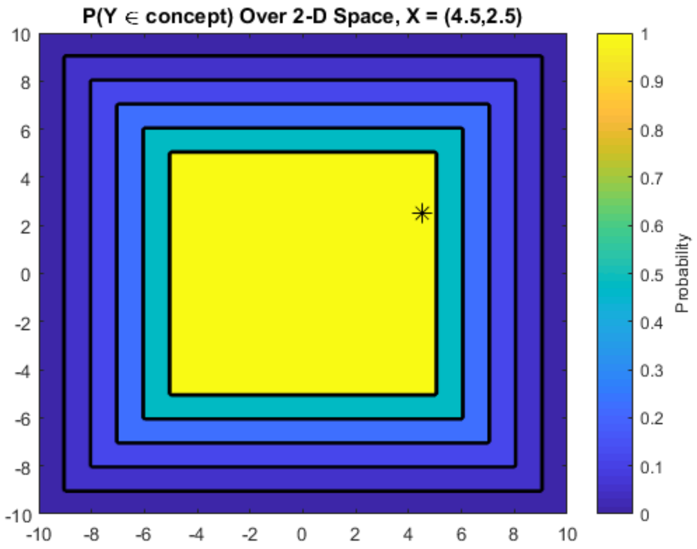

Figure 4: Generalization predictions of concept class membership, X = {(4.5,2.5)}

Now note that the probability of any point in the fifth hypothesis square belonging to the concept is one. This is for similar reasons as those described in task 3. In addition, the spread of likely concept classes is larger here than in task 3 partly because the relative difference between concept classes is smaller. For example, samples drawn entirely from concept square 8 fitting in concept square 5 is more likely than samples drawn from concept square 5 fitting completely in concept square 2)

5 Increasing the Number of Observed Points

Now, using $\sigma_1 = \sigma_2 = 30$ for the prior, I used the formulation used before to find the generalization predictions for $X = \{(2.2, -0.2)\}$, $X = \{(2.2, -0.2), (0.5, 0.5)\}$, and $X = \{(2.2, -0.2), (0.5, 0.5), (1.5, 1)\}$ The results of this are shown below:

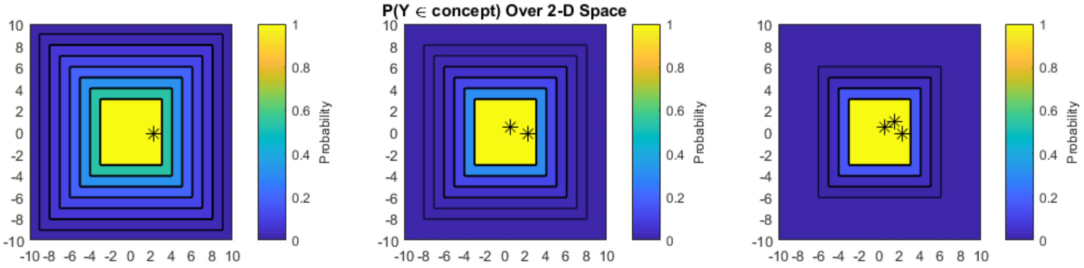

Figure 5: Generalization predictions of concept class membership

5.1 How the Posterior is Changing

As more data are observed, two things happen:

1. Consider the largest concept square such that all X data is inside of it. All points inside this square always have a probability of 1.
2. The spread of likely concept squares beyond this square shrinks in as more data is observed.

5.2 Why this Occurs

I will address both points above separately. However, it is important to note that, for any point y, its generalization prediction is the sum of the probabilities of all concepts that contain y.

1. Every hypothesis square that does NOT contain every point X necessarily has a probability of zero. In addition, these hypothesis squares are subsets of every other hypothesis square. Therefore, any point y inside the smallest hypothesis square that contains all X points must also be in EVERY possible hypothesis square. Therefore, these points MUST be in the concept class, so their generalization prediction is one.
2. As more and more points show up inside a certain hypothesis square, it becomes less and less likely that they were generated by a different hypothesis square, but *happened* to all fit inside a smaller one. This unlikeliness is captured by the model, and hypothesis squares outside of the smallest possible one become less likely. Namely, this is capture in the likelihood function:

$$P(X|h) = \left(\frac{1}{s_1 * s_2}\right)^n * \mathbb{1}_{\forall j(x_j \in h)}$$

as n (the number of points in X) increases, hypothesis classes with large s_1 and s_2 values will have their likelihoods decrease faster than smaller hypothesis classes, and so the smaller hypothesis classes will be favored (assuming all X data is inside this smaller hypothesis class)

6 Experimenting with Different Priors and Likelihoods

6.1 Uniform Prior

I repeated task 5, but instead of using the expected-size prior, I used a uniform prior:

$$P(h) = 1/10$$

Since there are a total of 10 hypothesis classes. The results of doing so are shown below:

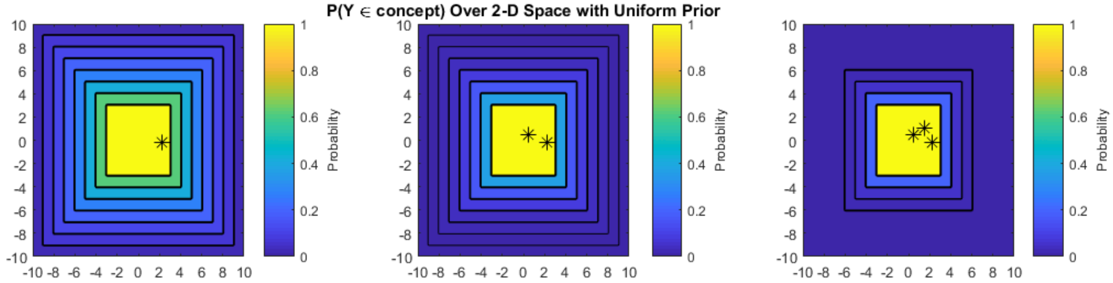

Figure 6: Generalization predictions of concept class membership with a uniform prior

This experiment exhibits similar trends to those in task 5, but points inside larger hypothesis squares are slightly more likely at each step. This is because the probability of each hypothesis class scales linearly with its prior. While the original prior punishes larger hypothesis classes, a uniform prior does no such thing, so larger hypothesis classes are more likely. As a result, points inside these larger hypothesis classes have a larger probability of being in the concept.

6.2 Beta Likelihood

I again repeated task 5 and changed the prior back to the expected-size prior. However, I did change the likelihood function from the size principle to a scaled beta distribution:

$$P(X|h) = \left(\frac{1}{s^2}\right)^n * \prod_{i=1}^{n} \frac{z_i^{\alpha-1}(1-z_i)^{\beta-1}}{\beta(\alpha,\beta)}$$

where s is the side length of the hypothesis square, α and β are parameters, and:

$$z_i = \frac{1}{2} + \frac{x_i}{s}$$

I did found the generalization prediction for a beta likelihood with $\alpha = \beta = 4$, and the results are shown below:

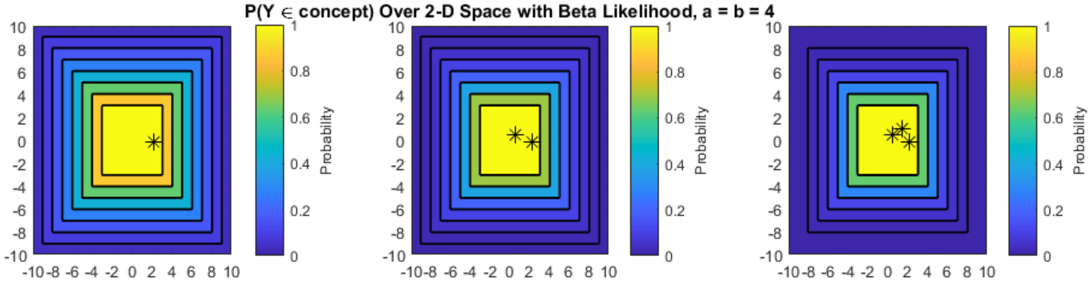

Figure 7: Generalization predictions of concept class membership with a Beta Likelihood

It is clear that now points inside larger concept classes are much more likely to be in the concept. This is because the beta distribution with $\alpha = \beta = 4$ looks like this:

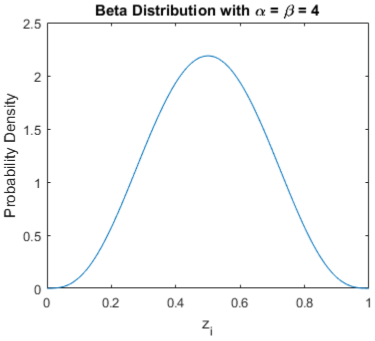

Figure 8: Beta Distribution with $\alpha = \beta = 4$

So it is much more likely that a larger hypothesis class will create data that *happen* to all exist inside another smaller hypothesis square. Therefore, the model can be less certain about the size of the concept.

In addition, the Beta distribution becomes more peaked as α and β increase, so it could be expected that this effect of giving more weight to larger concept classes could be tested by changing α and β. To this end, a plot of the posterior distribution of each concept class for different values of α and β is shown below:

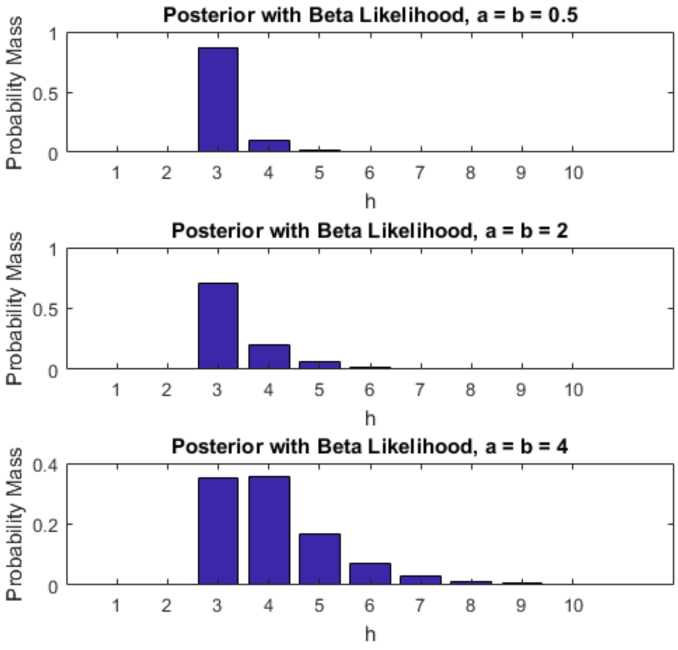

Figure 9: Posterior distributions for square hypothesis classes with different Beta likelihoods

Interestingly, we see that when the Beta distribution is most peaked at $\alpha = \beta = 4$, hypothesis class 4 is MORE likely than hypothesis class 3. This could be due to the fact that the data is spread too far within hypothesis square 3 to be feasible, given the sharp peak of the beta distribution.