Big Data
Solutions: PAC learning with finite hypothesis classes

Part 1: The Realizable Case

Exercise 1

Recall that $L_{\mathcal{D},f}(h)$ is the probability that h misclassifies an example drawn at random from the population:

$$L_{\mathcal{D},f}(h) = \mathbb{P}_{x\sim\mathbb{P}(X)}[h(x) \neq f(x)],$$

where $f(x)$ is the true class label of x, and $h(x)$ is the class label assigned to x by h.

In contrast, $L_D(h)$ is the fraction of training examples misclassified by h:

$$L_D(h) = \frac{1}{m}\sum_{i=1}^{m} I[h(x_i) \neq f(x_i)],$$

where I denotes the indicator function for the truth of its argument, i.e., $I[A] = 1$ if A is true, and $I[A] = 0$ if A is false. A set of m training examples is generically denoted as:

$$D = \{(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_m, f(x_m))\}$$

(a)
1. Since h_b^1 is a bad hypothesis, we know that $L_{\mathcal{D},f}(h_b^1) > \epsilon$, that is, $\mathbb{P}_{x\sim\mathbb{P}(X)}[h_b^1(x) \neq f(x)] > \epsilon$. The first training example, like any other training example, has been drawn at random from $\mathbb{P}(X)$ (and labeled according to f), so we may conclude that $\mathbb{P}\left[h_b^1(x_1) = f(x_1)\right] \leq 1 - \epsilon$.
2. The training examples have been drawn independently from $\mathbb{P}(X)$, so the events $h_b^1(x_i) = f(x_i)$ and $h_b^1(x_j) = f(x_j)$ are independent as well (for $i \neq j$ of course). Hence the joint probability of these m events is equal to the product of their individual probabilities:

$$\mathbb{P}\left[h_b^1(x_1) = f(x_1) \wedge h_b^1(x_2) = f(x_2) \wedge \ldots \wedge h_b^1(x_m) = f(x_m)\right] \leq (1-\epsilon)^m$$

(b) Using the union bound we can guarantee that this probability is $\leq |\mathcal{H}_B|(1-\epsilon)^m$

(c) $|\mathcal{H}_B| \leq |\mathcal{H}| - 1$. Since we're in the realizable case, we know that there is at least one hypothesis in $\mathcal{H}$ with zero training error. For notational simplicity we'll throw in that single hypothesis we were able to exclude as well. So we'll just use $|\mathcal{H}_B| \leq |\mathcal{H}|$.

(d) We want

$$\mathbb{P}_{D\sim\mathbb{P}(X)^m}[\exists h \in \mathcal{H} : L_D(h) = 0 \wedge L_{\mathcal{D},f}(h) > \epsilon] \leq \delta \tag{1}$$

We have

$$\mathbb{P}_{D\sim\mathbb{P}(X)^m}[\exists h \in \mathcal{H} : L_D(h) = 0 \wedge L_{\mathcal{D},f}(h) > \epsilon] \leq |\mathcal{H}|e^{-\epsilon m} \tag{2}$$

So for $\delta \geq |\mathcal{H}|e^{-\epsilon m}$ inequality (1) is satisfied. But we want to know for which values of m inequality (1) is satisfied. Here are the detailed steps:

$$\begin{aligned}
\delta &\geq |\mathcal{H}|e^{-\epsilon m} && \text{(Divide by } |\mathcal{H}|\text{)}\\
\frac{\delta}{|\mathcal{H}|} &\geq e^{-\epsilon m} && \text{(Take natural log)}\\
\ln\left(\frac{\delta}{|\mathcal{H}|}\right) &\geq -\epsilon m && (\ln \tfrac{a}{b} = \ln a - \ln b)\\
\ln\delta - \ln|\mathcal{H}| &\geq -\epsilon m && \text{(Multiply by } -1\text{)}\\
\ln|\mathcal{H}| - \ln\delta &\leq \epsilon m && \text{(Divide by } \epsilon\text{)}\\
m &\geq \frac{1}{\epsilon}\left(\ln|\mathcal{H}| - \ln\delta\right) && (-\ln a = \ln \tfrac{1}{a})\\
m &\geq \frac{1}{\epsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)
\end{aligned}$$

This many training examples will suffice.

(e)

1. For bigger hypothesis sets, we need more training examples.
2. If we want the true error of the classifier to be smaller, we need more training examples.
3. If we want bigger confidence that we achieve the required true error, we need more training examples.

(f) We start from

$$\mathbb{P}_{D\sim\mathbb{P}(X)^m}\left[L_{\mathcal{D},f}(h_D) > \epsilon\right] \leq |\mathcal{H}|e^{-\epsilon m}.$$

This implies that

$$\mathbb{P}_{D\sim\mathbb{P}(X)^m}\left[L_{\mathcal{D},f}(h_D) \leq \epsilon\right] \geq 1 - |\mathcal{H}|e^{-\epsilon m}. \tag{3}$$

We want to guarantee that

$$\mathbb{P}_{D\sim\mathbb{P}(X)^m}\left[L_{\mathcal{D},f}(h_D) \leq \epsilon\right] \geq 1 - \delta. \tag{4}$$

By equation (3) this holds for $\delta \geq |\mathcal{H}|e^{-\epsilon m}$. Solving this inequality for ϵ, we obtain

$$\epsilon \geq \frac{1}{m}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right) \tag{5}$$

Finally, we conclude that for $\epsilon \geq \frac{1}{m}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)$, equation (4) is satisfied. The strongest conclusion we can draw now is that

$$\mathbb{P}_{D\sim\mathbb{P}(X)^m}\left[L_{\mathcal{D},f}(h_D) \leq \frac{1}{m}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)\right] \geq 1-\delta. \tag{6}$$

This type of bound is called a generalization bound.

Exercise 2: Axis-aligned rectangles

(a) The number of distinct hypotheses is:

$$|\mathcal{H}| = \left(\binom{n}{1} + \binom{n}{2}\right)^2 = \left(\frac{n(n+1)}{2}\right)^2$$

Explanation: in each of the x_1 and x_2 dimensions, we can pick any pair of (not necessarily different) values from $[0, n-1]$ and assign the lower of the two values to the lower bound, and the higher of the two values to the upper bound of the rectangle in that dimension. So for each dimension we have $\binom{n}{1} + \binom{n}{2}$ different choices. Each choice for x_1 can be combined with each choice for x_2 to yield a different rectangle.

(b) For $n = 100$, we have:

$$|\mathcal{H}| = \left(\frac{100\times 101}{2}\right)^2 = 25,502,500$$

For $\epsilon = 0.15$ and $\delta = 0.05$ we require

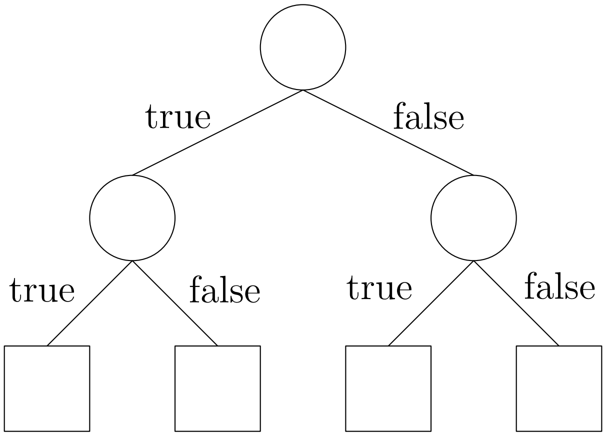
$$m \geq \frac{1}{\epsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right) = \frac{1}{0.15}\left(\ln(25,502,500) + \ln\frac{1}{0.05}\right) = 133.67$$

So $m = 134$ will do.

(c) Return the smallest rectangle that includes all the positive (i.e. class label = 1) examples present in the training sample.

Exercise 3: Regular depth-2 decision trees

(a) We have to find the number of ways to complete the picture below according to the rules for regular depth-2 decision trees:



We can choose any of the n variables for the root node, and then there are $n-1$ variables left to choose from for the children of the root node. Finally, we can assign class labels to the leaf nodes in $2^4 = 16$ ways. So the total number of different trees is

$$16n(n-1)$$

(b)

$$m \geq \frac{1}{\epsilon}\left(\ln(16n) + \ln(n-1) + \ln\frac{1}{\delta}\right)$$

The "sufficient sample size" grows as the logarithm of the number of variables.

(c) Not all trees that look different express different hypotheses. For example, the order of the two variables that are used actually doesn't matter. For every tree with x_i in the root, and x_j to follow, there is an equivalent tree (i.e. expressing the same hypothesis) with x_j in the root and x_i to follow. So we could have counted just $8n(n-1)$ hypotheses.

Since we counted *too many* hypotheses, the bound given at (b) is still valid, even though it is a bit more "loose" then necessary.

Part 2: Agnostic PAC-learning

Exercise 4: Sample complexity of agnostic PAC-learning

(a) Starting with $\delta \geq 2|\mathcal{H}|e^{-\frac{1}{2}\epsilon^2 m}$ and solving for m you should find:

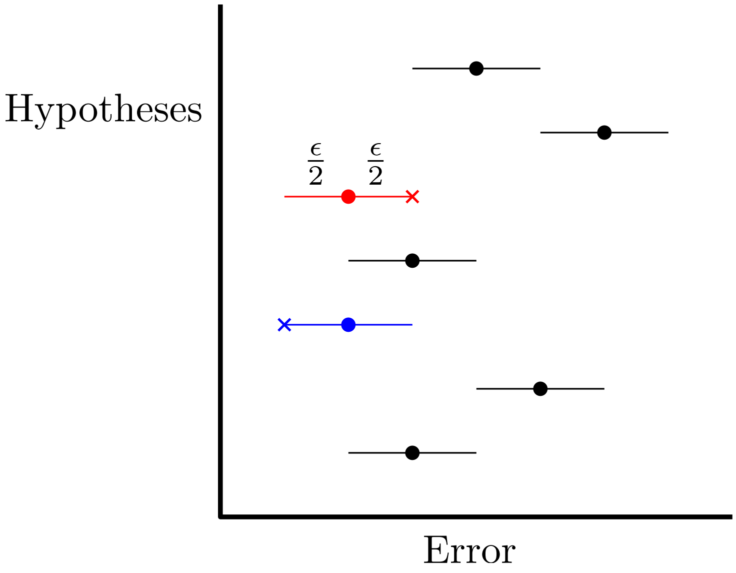
$$m \geq \frac{2}{\epsilon^2}\left(\ln 2|\mathcal{H}| + \ln\frac{1}{\delta}\right).$$

In the realizable case we got:

$$m \geq \frac{1}{\epsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right).$$

The most important difference is that in the agnostic case we have the factor $\frac{2}{\epsilon^2}$, instead of $\frac{1}{\epsilon}$ in the realizable case. Since ϵ is between 0 and 1, and typically close to 0, we have that ϵ^2 is smaller than ϵ, and hence $\frac{1}{\epsilon^2}$ is bigger than $\frac{1}{\epsilon}$. Bottom line is we need more data in the agnostic case.

(b) There are formal proofs on the lecture slides and in the book. But the idea can be seen from the following picture:



The dots indicate the training error of different hypotheses. Since the sample is $\frac{\epsilon}{2}$ representative, we know that the true error is within $\frac{\epsilon}{2}$ of the training error. This interval is indicated by the horizontal bar. Our ERM-algorithm will return a hypothesis with minimum training error. In the picture, both the red and the blue hypothesis achieve the minimum training error, so an ERM algorithm might return either one of them. Let's say our ERM-algorithm returns the red one. Worst thing that can happen is that its true error (the red cross) is $\frac{\epsilon}{2}$ higher than its training error, whereas for the hypothesis we did not select, the true error (the blue cross) is $\frac{\epsilon}{2}$ smaller than its training error. But even then the true error of the selected hypothesis is still within ϵ of the best true error.

Exercise 5: Learning threshold functions

(a) Filling in the numbers in the formula we found in exercise 4, we get:

$$m \geq \frac{2}{\epsilon^2}\left(\ln 2|\mathcal{H}| + \ln\frac{1}{\delta}\right) = \frac{2}{0.01^2}\left(\ln 18 + \ln\frac{1}{0.05}\right) = 117,722.1.$$

Rounding up to the nearest integer, we get $m \geq 117,723$.

Notice that we don't have to divide ϵ by 2, because we already filled in $\frac{\epsilon}{2}$ in the Hoeffding inequality.

(b) $m \geq 1178$.

(c) This doesn't make any difference. What matters is the number of hypothesis in $\mathcal{H}$, and that is still 9.

Exercise 6: Axis-aligned rectangles

For the number of hypotheses in this set, see the solution to exercise 2. Filling in the numbers in the formula we found in exercise 4, we get:

$$m \geq \frac{2}{\epsilon^2}\left(\ln 2|\mathcal{H}| + \ln\frac{1}{\delta}\right) = \frac{2}{0.15^2}\left(\ln 51,005,000 + \ln\frac{1}{0.05}\right) = 1,843.837.$$

Rounding up to the nearest integer, we get $m \geq 1,844$.