# Big Data
# Exercises: PAC learning with finite hypothesis classes

## Part 1: The Realizable Case

We first assume that the true labeling function is in the hypothesis set, that is, $f \in \mathcal{H}$.

### Exercise 1

On slide 17 of lecture 5, the following result was stated (in slightly different notation):

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

where $h_D$ is any hypothesis output by a consistent learner, that is, $L_D(h_D) = 0$, and $\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}$ denotes the set of bad hypotheses. Hence, another way to state the result is

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

Let's build up this result step by step. We have to find the probability after seeing $m$ samples from $\mathbb{P}(X)$, that the version space still contains a bad hypothesis (if it doesn't, then our consistent learner will certainly output a hypothesis with true error less than $\epsilon$). For concreteness, let's list the bad hypotheses as $h_b^1, h_b^2, \ldots, h_b^k$.

(a) First, we consider some fixed bad hypothesis, say $h_b^1$.

1. Bound the probability that $h_b^1$ classifies the first training example correctly.
2. Bound the probability that $h_b^1$ classifies all $m$ training examples correctly.

(b) Bound the probability that any of the $k = |\mathcal{H}_B|$ bad hypotheses classifies all $m$ training examples correctly.

(c) Give an upper bound for $|\mathcal{H}_B|$.

(d) Using the fact that for $0 \leq \epsilon \leq 1$, $(1-\epsilon) \leq e^{-\epsilon}$, show that if we want

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

to be at most $\delta$, then

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

training examples will suffice.

(e) Verify that qualitatively, the dependence of $m$ on $|\mathcal{H}|$, $\epsilon$ and $\delta$ makes sense:

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

(f) Use a similar argument to show that for any fixed sample size $m$ and confidence parameter $\delta$, with probability at least $1-\delta$ any consistent learner returns a hypothesis $h_D$ with:

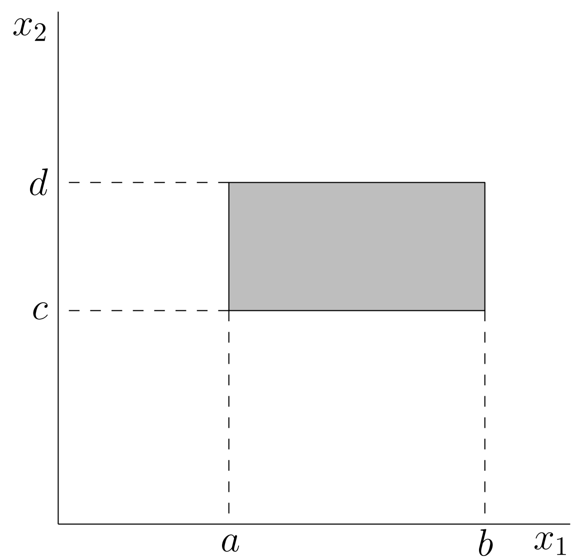
$$L_{\mathcal{D},f}(h_D) \leq \frac{1}{m}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)$$

**Exercise 2: Axis-aligned rectangles**

For any integers $a \leq b, c \leq d \in [0, n-1]$, let

$$h(x_1, x_2) = \begin{cases} 1 & \text{if } a \leq x_1 \leq b \text{ and } c \leq x_2 \leq d \\ 0 & \text{otherwise} \end{cases}$$

In a picture:




Let $\mathcal{H}$ denote the class of all such axis-aligned rectangles.

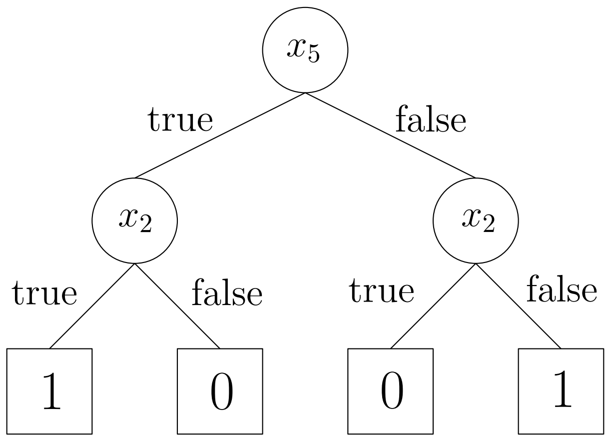
(a) As a function of $n$, how many distinct rectangles are there in $\mathcal{H}$?

(b) Let $n = 100$. How many training examples suffice to ensure that for any $f \in \mathcal{H}$, any consistent learner that uses $\mathcal{H}$ will, with probability at least 95% output a hypothesis with error at most 0.15?

(c) Describe a consistent learner for the hypothesis class of axis-aligned rectangles.

**Exercise 3: Regular depth-2 decision trees**

Consider the hypothesis class of regular depth-2 decision trees over $n$ boolean variables $x_1, x_2, \ldots, x_n$. A regular depth-2 decision tree is a depth-2 decision tree (a tree with four leaves, all at distance 2 from the root) in which the left and right child of the root are required to split on the same variable. For instance, the following tree is a regular depth-2 decision tree:




Note that the decision tree may use any of the $n$ variables to split on; in this example it happened to be $x_2$ and $x_5$. The tree above represents the following prediction rule:

- If $x_5 =$ true and $x_2 =$ true then $h(x_5, x_2) = 1$
- If $x_5 =$ true and $x_2 =$ false then $h(x_5, x_2) = 0$
- If $x_5 =$ false and $x_2 =$ true then $h(x_5, x_2) = 0$
- If $x_5 =$ false and $x_2 =$ false then $h(x_5, x_2) = 1$

(a) As a function of $n$, how many different trees are there in $\mathcal{H}$?

(b) As a function of $\epsilon$, $\delta$ and $n$, how many training examples suffice to ensure that for any $f \in \mathcal{H}$, any consistent learner that uses $\mathcal{H}$ will, with probability at least $1 - \delta$ output a hypothesis with error at most $\epsilon$ ?
How does the "sufficient sample size" grow with the number of variables?

(c) Do all trees that look different really express different hypotheses? If not, does that mean your answer to question (b) is incorrect?

## Part 2: Agnostic PAC-learning

We drop the assumption that there is a hypothesis in $\mathcal{H}$ with zero true error (the realizability assumption), and move to agnostic PAC-learning. We are still considering only finite hypothesis classes.

### Exercise 4: Sample complexity of agnostic PAC-learning

If we require

$$\mathbb{P}_{D\sim\mathbb{P}(X,Y)^m}\left[L_{\mathcal{D}}(h_D) > \min_{h\in\mathcal{H}}\{L_{\mathcal{D}}(h)\} + \epsilon\right] \leq \delta,$$

where $h_D$ is any hypothesis output by an ERM learner, then it suffices to obtain a sample that is $\frac{\epsilon}{2}$ representative with probabability at least $1-\delta$. That is, we need

$$\mathbb{P}_{D\sim\mathbb{P}(X,Y)^m}\left[\exists h \in \mathcal{H} : \left|L_{\mathcal{D}}(h) - L_D(h)\right| > \frac{\epsilon}{2}\right] \leq \delta.$$

By the union bound and Hoeffding's inequality we have that

$$\mathbb{P}_{D\sim\mathbb{P}(X,Y)^m}\left[\exists h \in \mathcal{H} : \left|L_{\mathcal{D}}(h) - L_D(h)\right| > \frac{\epsilon}{2}\right] \leq 2|\mathcal{H}|e^{-\frac{1}{2}\epsilon^2 m}.$$

Hence, for $\delta \geq 2|\mathcal{H}|e^{-\frac{1}{2}\epsilon^2 m}$ we're good.

(a) Derive a formula for the sufficient sample size to meet given $(\epsilon, \delta)$ requirements. Compare this to the formula we obtained in the realizable case.

(b) Show that if the sample is $\frac{\epsilon}{2}$ representative with respect to $\mathcal{H}$, then

$$L_{\mathcal{D}}(h_D) \leq \min_{h\in\mathcal{H}}\{L_{\mathcal{D}}(h)\} + \epsilon,$$

for any ERM hypothesis $h_D$.

### Exercise 5: Learning threshold functions

Consider the class of threshold functions $\mathcal{H} = \{\frac{1}{10}, \frac{2}{10}, \ldots, \frac{9}{10}\}$, and let $x$ be a real number in the interval $[0,1]$. For example, one of the members of $\mathcal{H}$ is the function:

$$h(x) = \begin{cases} 1 & \text{if } x \geq \frac{1}{10} \\ 0 & \text{otherwise} \end{cases}$$

(a) How many examples suffice to agnostically PAC learn this hypothesis class for $\epsilon = 0.01$ and $\delta = 0.05$?

(b) What if $\epsilon = 0.1$?

(c) What if $x$ can be *any* real number?

**Exercise 6: Axis-aligned rectangles**

For any integers $a \leq b, c \leq d \in [0, 99]$, let

$$h(x_1, x_2) = \begin{cases} 1 & \text{if } a \leq x_1 \leq b \text{ and } c \leq x_2 \leq d \\ 0 & \text{otherwise} \end{cases}$$

Let $\mathcal{H}$ denote the class of all such axis-aligned rectangles.

How many training examples suffice to ensure that any ERM learner that uses $\mathcal{H}$ will, with probability at least 95% output a hypothesis with true error at most 0.15 worse than the hypothesis with lowest true error in $\mathcal{H}$? Compare this sample size, to the one that was sufficient in the realizable case.