Solutions to Homework 1
CS-526 Learning Theory

Exercise 3.1

The hypothesis class $\mathcal{H}$ being PAC learnable with sample complexity $m_{\mathcal{H}}(\cdot,\cdot)$ means that there is a learning algorithm A such that when running A on $m \geq m_{\mathcal{H}}(\epsilon,\delta)$ i.i.d. samples generated by $\mathcal{D}$, with probability at least $1-\delta$, A returns a hypothesis $h \in \mathcal{H}$ with $L_{\mathcal{D}}(h) \leq \epsilon$.

Given $0 < \epsilon_1 \leq \epsilon_2 < 1$, consider $m \geq m_{\mathcal{H}}(\epsilon_1,\delta)$. We have that, with probability at least $1-\delta$, A returns a hypothesis $h \in \mathcal{H}$ satisfying $L_{\mathcal{D}}(h) \leq \epsilon_1 \leq \epsilon_2$. This implies that $m_{\mathcal{H}}(\epsilon_1,\delta)$ is a sufficient number of samples for accuracy ϵ_2. Therefore, $m_{\mathcal{H}}(\epsilon_1,\delta) \geq m_{\mathcal{H}}(\epsilon_2,\delta)$.

The proof of $m_{\mathcal{H}}(\epsilon,\delta_1) \geq m_{\mathcal{H}}(\epsilon,\delta_2)$ for $0 < \delta_1 \leq \delta_2 < 1$ follows analogously from the definition.

Exercise 3.3

We can simplify our task by realizing that this is equivalent of thinking of a threshold on a line. Imagine that all points with label 0 are on the left of a threshold and all points with label 1 are on the right of this threshold. We are given m samples. Consider the interval between the maximum sample of label 0 and the minimum sample of label 1. Let κ be the probability mass under the true distribution of samples falling into this interval. The chance that we get no samples in this interval is $(1-\kappa)^m$. Assume that we choose our threshold anywhere in this interval. The risk of the resulting classifier is then upper bounded by κ. We want that the risk is no more than ϵ with probability at least $1-\delta$. If $\epsilon \geq \kappa$ we are done. If $\epsilon \leq \kappa$ then $(1-\kappa)^m \leq (1-\epsilon)^m \leq \delta$. We conclude that as long as $m \geq \log(1/\delta)/\log(1/(1-\epsilon))$. Since $\log(1/\delta)/\epsilon \geq \log(1/\delta)/\log(1/(1-\epsilon))$ a valid choice $m \geq \lceil \log(1/\delta)/\epsilon \rceil$.

Below is an alternative proof. The realizability assumption for $\mathcal{H} = \{h_r : r \in \mathbb{R}_+\}$ implies that there is a circle of radius r^* such that, almost surely, any x inside it has label $y = 1$ and any x outside it as label $y = 0$. The learning task here is to distinguish this circle.

We now consider the ERM algorithm which, given a training sequence $S = \{(x_i, y_i)\}_{i=1}^m$, returns the hypothesis $h_S \in \mathcal{H}$ corresponding to the *tightest* circle which contains all of the positive (meaning $y_i = 1$) instances in S and none of the negative ones. We denote r_S the radius of this tightest circle. Under the realizability assumption, $r_S \leq r^*$ and $\forall S \in (\mathcal{X} \times \mathcal{Y})^m$:

$$L_{\mathcal{D}}(h_S) = \mathbb{P}_{(x,y)\sim\mathcal{D}}\big(r_S < \|x\| \leq r^*\big) .$$

Let $\epsilon_0 = \mathbb{P}_{(x,y)\sim\mathcal{D}}(0 < \|x\| \leq r^*)$. Note that $r \in [0, r^*] \mapsto \mathbb{P}_{(x,y)\sim\mathcal{D}}(r < \|x\| \leq r^*)$ is non increasing so $\forall r \in [0, r^*] : \mathbb{P}_{(x,y)\sim\mathcal{D}}(r < \|x\| \leq r^*) \leq \epsilon_0$. Therefore, for any $\epsilon \in (\epsilon_0, 1]$, $\{L_{\mathcal{D}}(h_S) \geq \epsilon\}$ is the empty set and $\mathbb{P}_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(h_S) \geq \epsilon) = 0 \leq e^{-\epsilon m}$. We now look at the more interesting case of $\epsilon \in [0, \epsilon_0]$. Define $r_\epsilon = \sup\big\{r \in [0, r^*] : \mathbb{P}_{(x,y)\sim\mathcal{D}}(r < \|x\| \leq r^*) \geq \epsilon\big\}$.

Assume for a moment that $r \mapsto \mathbb{P}_{(x,y)\sim\mathcal{D}}(r < \|x\| \leq r^*)$ is continuous on $[0, r^*]$. Then $\mathbb{P}_{(x,y)\sim\mathcal{D}}(r_\epsilon < \|x\| \leq r^*) = \epsilon$ and $L_\mathcal{D}(h_S) \geq \epsilon$ if, and only if, $r_S \leq r_\epsilon$. It directly follows that:

$$\begin{aligned}
\mathbb{P}_{S\sim\mathcal{D}^m}(L_\mathcal{D}(h_S) \geq \epsilon) &= \mathbb{P}_{S\sim\mathcal{D}^m}(r_S \leq r_\epsilon) \\
&= \mathbb{P}_{S\sim\mathcal{D}^m}\big(\text{no points in } S \text{ belongs to } \{x \in \mathbb{R}^2 : r_\epsilon < \|x\| \leq r^*\}\big) \\
&= (1-\epsilon)^m \\
&\leq e^{-\epsilon m} \ .
\end{aligned}$$

If $r \mapsto \mathbb{P}_{(x,y)\sim\mathcal{D}}(r < \|x\| \leq r^*)$ is not continuous, we have to consider two cases:

1. If $\mathbb{P}_{(x,y)\sim\mathcal{D}}(r_\epsilon < \|x\| \leq r^*) \geq \epsilon$ then $L_\mathcal{D}(h_S) \geq \epsilon$ if, and only if, $r_S \leq r_\epsilon$. Similarly to the continuous case:

$$\begin{aligned}
\mathbb{P}_{S\sim\mathcal{D}^m}(L_\mathcal{D}(h_S) \geq \epsilon) &= \mathbb{P}_{S\sim\mathcal{D}^m}(r_S \leq r_\epsilon) \\
&= \mathbb{P}_{S\sim\mathcal{D}^m}\big(\text{no points in } S \text{ belongs to } \{x \in \mathbb{R}^2 : r_\epsilon < \|x\| \leq r^*\}\big) \\
&\leq (1-\epsilon)^m \\
&\leq e^{-\epsilon m} \ .
\end{aligned}$$

2. If $\mathbb{P}_{(x,y)\sim\mathcal{D}}(r_\epsilon < \|x\| \leq r^*) < \epsilon$ then $L_\mathcal{D}(h_S) \geq \epsilon$ if, and only if, $r_S < r_\epsilon$. Therefore:

$$\begin{aligned}
\mathbb{P}_{S\sim\mathcal{D}^m}(L_\mathcal{D}(h_S) \geq \epsilon) &= \mathbb{P}_{S\sim\mathcal{D}^m}(r_S < r_\epsilon) \\
&= \mathbb{P}_{S\sim\mathcal{D}^m}\big(\text{no points in } S \text{ belongs to } \{x \in \mathbb{R}^2 : r_\epsilon \leq \|x\| \leq r^*\}\big) \\
&\leq (1-\epsilon)^m \\
&\leq e^{-\epsilon m} \ ,
\end{aligned}$$

   where the first inequality uses that $\mathbb{P}_{(x,y)\sim\mathcal{D}}(r_\epsilon \leq \|x\| \leq r^*) \geq \epsilon$.

The desired bound on the sample complexity follows from requiring $e^{-\epsilon m} \leq \delta$.

Exercise 3.7

Let g be any *potentially probabilistic* classifier from $\mathcal{X}$ to $\{0, 1\}$. Note that for the 0-1 loss:

$$\begin{aligned}
L_\mathcal{D}(g) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{1}_{g(x)\neq y}] = \mathbb{E}_{x\sim\mathcal{D}_x}\big[\mathbb{E}_{y\sim\mathcal{D}_{y|x}}[\mathbb{1}_{g(x)\neq y}]\big] = \mathbb{E}_{x\sim\mathcal{D}_x}\big[\mathbb{P}(g(x) \neq y|x)\big] \ ; \\
L_\mathcal{D}(f_\mathcal{D}) &= \mathbb{E}_{x\sim\mathcal{D}_x}\big[\mathbb{P}(f_\mathcal{D}(x) \neq y|x)\big] \ .
\end{aligned}$$

We will compare the two conditional probabilities inside the expectations over $x \sim \mathcal{D}_x$. Let $x \in \mathcal{X}$ and $a_x := \mathbb{P}(y = 1|x)$. Using the conditional independence of $g(x)$ and y given x, we have:

$$\begin{aligned}
\mathbb{P}(g(x) \neq y|x) &= \mathbb{P}(g(x) = 0|x) \cdot \mathbb{P}(y = 1|x) + \mathbb{P}(g(x) = 1|x) \cdot \mathbb{P}(y = 0|x) \\
&= \mathbb{P}(g(x) = 0|x) \cdot a_x + \mathbb{P}(g(x) = 1|x) \cdot (1 - a_x) \\
&\geq \mathbb{P}(g(x) = 0|x) \cdot \min\{a_x, 1 - a_x\} + \mathbb{P}(g(x) = 1|x) \cdot \min\{a_x, 1 - a_x\} \\
&= \min\{a_x, 1 - a_x\} \ .
\end{aligned}$$

If $g = f_\mathcal{D}$ then $\mathbb{P}(g(x) = 0|x) = \mathbb{1}_{a_x<1/2}$ and $\mathbb{P}(g(x) = 1|x) = \mathbb{1}_{a_x\geq1/2}$, and the above inequality is tight:

$$\mathbb{P}(f_\mathcal{D}(x) \neq y|x) = \mathbb{1}_{a_x<1/2} \cdot a_x + \mathbb{1}_{a_x\geq1/2} \cdot (1 - a_x) = \min\{a_x, 1 - a_x\} \ .$$

Therefore, we have $L_\mathcal{D}(f_\mathcal{D}) \leq L_\mathcal{D}(g)$.

Exercise 3.8

1. Solved already in Exercise 3.7.

2. We have shown in Exercise 3.7 that the Bayes optimial predictor $f_{\mathcal{D}}$ is optimal w.r.t. $\mathcal{D}$; in other words, $f_{\mathcal{D}}$ is always better than any other learning algorithm w.r.t. $\mathcal{D}$.

3. Take $\mathcal{D}$ to be any probability distribution and $B = f_{\mathcal{D}}$.