Intro to Statistical Learning Theory
Exercise 1

1) Binary classification - finite realizable case:
A distribution $\mathcal{D}$ is realizable by $\mathcal{H}$ if there exists some $h^* \in \mathcal{H}$ such that $L_{\mathcal{D}}(h^*) = 0$.

Assume $\mathcal{Y} = \{\pm 1\}$, 0-1 loss, and $\mathcal{H}$ is a finite hypothesis class. Prove that $\mathcal{H}$ can PAC learn any realizable distribution $\mathcal{D}$ with $\mathfrak{M}(\epsilon, \delta) = \mathcal{O}\left(\frac{\log\left(\frac{|\mathcal{H}|}{\delta}\right)}{\epsilon}\right)$. You can use the inequality $1 - x \leq e^{-x}$.

2) Bayes optimal predictor: We define the Bayes-optimal predictor h_b as $h_b(x) = \arg\min_{\bar{y} \in \mathcal{Y}} \mathbb{E}_y[\ell(\bar{y}, y)|x]$.

   a) Show that for classification with $\mathcal{Y} = \{1, ..., k\} = [k]$ and $0 - 1$ loss, $h_b(x) = \arg\max_{y \in [k]} P(y|x)$.

   b) Show that for regression, $\mathcal{Y} = \mathbb{R}$ with squared loss $\ell(y, \bar{y}) = (y - \bar{y})^2$, that $h_b(x) = \mathbb{E}[y|x]$.

3) From bounded expected risk to agnostic PAC learning:
Let A be an algorithm that guarantees the following: If $m > \mathfrak{M}(\epsilon)$ then for every distribution $\mathcal{D}$ it holds that $\mathbb{E}_S[L_{\mathcal{D}}(A(S))] < \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$.

   a) Show that for every $\delta \in (0, 1)$, if $m > \mathfrak{M}(\epsilon \cdot \delta)$ then with probability of at least $1 - \delta$ it holds that $L_{\mathcal{D}}(A(S)) < \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$ (hint: Markov's inequality).

   b) For every $\delta \in (0, 1)$ let $k = \lceil \log_2(1/\delta) + 1 \rceil$ and $\bar{\mathfrak{M}}(\epsilon, \delta) = \mathfrak{M}(\epsilon/2)k + \left\lceil 2\frac{\ln(2/\delta) + \ln(k)}{\epsilon^2} \right\rceil$ Suggest a procedure that PAC learns the problem with sample complexity of $\bar{\mathfrak{M}}(2\epsilon, \delta)$ assuming that the loss function is bounded by 1. Hint: Divide the data into k+1 chunks, where each of the first k chunks is of size $\mathfrak{M}(\epsilon/2)$ examples.

4) Show that there exists a hypothesis space $\mathcal{H}$ with $|\mathcal{H}| = 2$ and an unbounded loss ℓ such that $\mathcal{H}$ is not PAC learnable.