Problem 2-1

Let the domain be $X = \mathbb{R}$, and let $C = C_s$ be the class of concepts defined by unions of $s$ intervals. That is, each concept $c$ is defined by real numbers $a_1 \leq b_1 \leq \cdots \leq a_s \leq b_s$ where $c(x) = 1$ if and only if $x \in [a_1, b_1] \cup \cdots \cup [a_s, b_s]$.

a. Compute the VC-dimension of $C_s$ exactly.

b. Describe an efficient algorithm that learns the class $C_s$ for every $s$, assuming that $s$ is known ahead of time to the learner. You should describe a single algorithm that works for all $C_s$, provided that $s$ is known so that the learner can choose the number of examples needed as a function of $\epsilon$, $\delta$ and $s$. Prove that your algorithm is PAC (i.e., produces a hypothesis with error at most $\epsilon$ with probability at least $1 - \delta$), and argue that both the running time and the required number of examples are polynomial in $1/\epsilon$, $1/\delta$ and $s$.

Problem 2-2

For this problem, you need not be concerned about computational efficiency. Throughout this problem, as usual, $C$ and $H$ are classes of concepts defined on the domain $X$.

a. Prove or disprove the following statement: For every finite domain $X$, and for all classes $C$ and $H$, if $C$ is PAC learnable by $H$, then $C \subseteq H$. (To prove the statement, you of course need to give a proof showing that it is always true. To disprove the statement, you can simply provide a counterexample showing that it is not true in general.)

b. Repeat part (a) without the assumption that $X$ is finite. In other words, prove or disprove that: For every (not necessarily finite) domain $X$, and for all classes $C$ and $H$, if $C$ is PAC learnable by $H$, then $C \subseteq H$.
Problem 2-3

Let $D$ be a distribution over $X \times \{0, 1\}$, and let $S = ((x_1, y_1), \ldots, (x_m, y_m))$ be a random sample from $D$. Let

$$
\text{err}(h) = \Pr_{(x,y) \sim D} [h(x) \neq y]
$$

$$
\hat{\text{err}}(h) = \frac{1}{m} \sum_{i=1}^{m} 1\{h(x_i) \neq y_i\}.
$$

For simplicity, we will assume that $H$ is finite, although the results of this problem can be carried over to the infinite case. Note that none of the results depend on $|H|$.

Let $\hat{h}$ and $h^*$ be the hypotheses in $H$ with minimum training error and generalization error, respectively:

$$
\hat{h} = \arg \min_{h \in H} \hat{\text{err}}(h)
$$

$$
h^* = \arg \min_{h \in H} \text{err}(h).
$$

Be sure to keep in mind that, unlike $h^*$, $\hat{h}$ is a random variable that depends on the random sample $S$.

a. Prove that

$$
E \left[ \hat{\text{err}}(\hat{h}) \right] \leq \text{err}(h^*) \leq E \left[ \text{err}(\hat{h}) \right].
$$

b. Prove that, with probability at least $1 - \delta$,

$$
|\hat{\text{err}}(\hat{h}) - E \left[ \hat{\text{err}}(\hat{h}) \right]| \leq O \left( \sqrt{\frac{\ln(1/\delta)}{m}} \right).
$$

Give explicit constants, and be sure to end up with a bound that is independent of $|H|$.

c. Explain in words the meaning of what you proved in both of the preceding parts, and how we would expect training error to compare to test error when using a machine learning algorithm on actual data.