1. PAC learnability. Consider the concept class $C$ of all conjunctions (allowing negations) over $n$ boolean features. Prove that this concept class is PAC learnable. Note that we have gone through this example in class but not in full detail. For this problem, you should work out the full detail, including showing that there exists a poly-time algorithm for finding a consistent hypothesis.

2. VC dimension. Consider the hypothesis space $H_c = \text{circles in the } (x, y) \text{ plane}$. Points inside the circle are classified as positive examples. What is the VC dimension of $H_c$? Please fully justify your answer.

3. Consider the class $C$ of concepts of the form $(a \leq x \leq b) \land (c \leq y \leq d)$, where $a, b, c,$ and $d$ are integers in the interval $[0, 0.99]$. Note that each concept in this class corresponds to a rectangle with integer-valued boundaries on a portion of the $(x, y)$ plane. Hint: Given a region in the plane bounded by the points $(0, 0)$ and $(n-1, n-1)$, the number of distinct rectangles with integer-valued boundaries within this region is $\left(\frac{n(n-1)}{2}\right)^2$.

   (a) Give an upper bound on the number of randomly drawn training examples sufficient to assure that for any target concept $c$ in $C$, any consistent learner using $H = C$ will, with probability $95\%$, output a hypothesis with error at most $0.15$.

   (b) Now suppose the rectangle boundaries $a, b, c,$ and $d$ take on real values instead of integer values. Update your answer to the first part of this question.

4. In the discussion of learning theory, a key assumption is that the training data has the same distribution as the test data. In some cases, we might have noisy training data. In particular, we consider the binary classification problem with labels $y \in \{0, 1\}$, and let $D$ be a distribution over $\{(x, y)\}$ that we think of as the original uncorrupted distribution. However, we don’t directly observe from this distribution for training. Instead, we observe from $D_{\tau}$, a corrupted distribution over $\{(x, y)\}$, which is the same as $D$, except that the label $y$ has some probability $0 < \tau < 0.5$ to be flipped. In other words, our training data is generated by first sampling from $D$, then with probability $\tau$ (independent of the observed $x$ and $y$) replace $y$ with $1 - y$. Note that $D_0 = D$.

The distribution $D_{\tau}$ models the setting in which an unreliable labeler is labeling the training data for you, and each example has a probability $\tau$ of being mislabeled. Even though the training data is corrupted, we are still interested in evaluating our hypotheses w.r.t. the original uncorrupted distribution $D$.

We define the generalization error with respect to $D_{\tau}$ to be:

$$\epsilon_{\tau} = P_{(x, y) \sim D_{\tau}}[h(x) \neq y]$$

Note that $\epsilon_0$ is the generalization error with respect to the clean distribution. It is with respect to $\epsilon_0$ that we wish to evaluate our hypothesis.

   a. For any hypothesis $h$, the quantity $\epsilon_0(h)$ can be calculated as a function of $\epsilon_{\tau}(h)$ and $\tau$. Write down a formula for $\epsilon_0(h)$ in terms of $\epsilon_{\tau}(h)$ and $\tau$.

   b. Let $H$ be finite and suppose our training set $S = \{(x^i, y^i) : i = 1, ..., m\}$ is obtained by sampling IID from $D_{\tau}$. Suppose we pick the hypothesis $\hat{h} \in H$ that minimizes the training error: $\hat{h} = \arg\min_{h \in H} \epsilon_{\tau}(h)$. Also, let $h^* = \arg\min_{h \in H} \epsilon_0(h)$, i.e., the optimal hypothesis in $H$. Let any $\delta, \gamma > 0$ be given, prove that for

$$\epsilon_0(\hat{h}) \leq \epsilon_0(h^*) + 2\gamma$$

to hold with probability $1 - \delta$, we need to have training data size

$$m \geq \frac{1}{2(1 - 2\tau)^2 \gamma^2} \log \frac{2|H|}{\delta}$$
Hint: you will need to use the following fact derived from Hoeffding bound:

\[ \forall h \in H, |\tilde{\epsilon}_\tau(h) - \hat{\epsilon}_\tau(h)| \leq \hat{\gamma} \text{ with probability } (1 - \delta), \delta = 2|H|\exp(-2\hat{\gamma}^2m) \]

where \( \tilde{\epsilon}_\tau(h) \) is the training error of \( h \) on training data \( S \) generated from corrupted distribution \( D_\tau \). You will also need to use the answer to subproblem (a).

5. Boosting. Please show that in each iteration of Adaboost, the weighted error of \( h_i \) on the updated weights \( D_{i+1} \) is exactly 50%. In other words, \( \sum_{j=1}^{N} D_{i+1}(j) I(h_i(X_j) \neq y_j) = 50\% \).