Name:

SI221 or MICAS911? (circle)

Points: / 70 (SI221) and 80 (MICAS911)

Time: 2h10 for SI221 and 2h30 for MICAS911

ExamSI221/MICAS911

Exercise 1 (4pts) *Give a class of functions that is not learnable (justify precisely).*

Exercise 2 (4pts) *Let $\mathcal{H}$ be a class of functions and let $\mathcal{B} \subset \mathcal{H}$. Prove or disprove the inequality*

$$L_S(ERM_{\mathcal{H}}) \leq \frac{1}{|\mathcal{B}|} \sum_{h \in \mathcal{B}} L_S(h).$$

Exercise 3 (4pts) *Show that the sample complexity $m_{\mathcal{H}}(\varepsilon, \delta)$ of a PAC learnable class of functions $\mathcal{H}$ is non-increasing in each of its arguments.*

Exercise 4 (7pts) *Let $\mathcal{X} = \{0,1\}^8$ and consider the class of functions*

$$\mathcal{H} = \{h : \mathcal{X} \to \{0,1\}\}.$$

- *(3pts) Show that $\mathcal{H}$ is learnable (you may rely on a result seen in class).*

- *(4pts) Let $m_{\mathcal{H}}(\varepsilon, \delta)$ denote the sample complexity of PAC learning $\mathcal{H}$. Prove that*

  $$m_{\mathcal{H}}(1/10, 1/10) \geq 2^7$$

  (You may rely on a result seen in class).

Exercise 5 (Segment classifiers, 33pts) *Given real numbers $a \leq b$ define the segment classifier $h_{(a,b)}$ as*

$$h_{(a,b)}(x) = \begin{cases} 1 & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}. \tag{1}$$

The class of segment classifiers is defined as

$$\mathcal{H}_{\text{seg}} = \{h_{(a,b)} : a \leq b, -\infty < a < b < \infty\}$$

Throughout this exercise we rely on the realizability assumption.

1. *(4pts) Compute the VC dimension of $\mathcal{H}_{\text{seg}}$.*

2. *(4pts) Let A be the algorithm that returns the smallest segment enclosing all "1" examples in the training set. Show that A is an ERM.*

3. *(3pts) Let $R^* = [a^*, b^*]$ be the segment that generates the labels and let $h^* \in \mathcal{H}$ be the corresponding function. Let $R(S^m)$ be the segment returned by A. Show that $R(S^m) \subseteq R^*$ and that $R^* \backslash R(S^m)$ is composed of two segments (which could be degenerate).*

4. *(4pts) Let us denote the two segments by $R_L(S^m)$ and $R_R(S^m)$. Show that if the probability under P of each of these segments is at most $\varepsilon/2$, then the hypothesis returned by $A(S^m)$ has error at most ϵ, that is $L_{P,h^*}(A(S^m)) \leq \varepsilon$.*

5. *(2pts) Deduce that if* $L_{P,h^*}(A(S^m)) > \varepsilon$ *then* $P(R_i(S^m)) > \varepsilon/2$ *for some* $i \in \{L, R\}$.

6. *(4pts) Define* $I(S^m)$ *as the set of indices* $i \in \{L, R\}$ *such that* $P(R_i(S^m)) > \varepsilon/2$. *Show that* $P^m(i \in I(S^m)) \leq (1 - \varepsilon/2)^m$.

7. *(4pts) Deduce that*

$$P^m(L_{P,h^*}(A(S^m)) > \varepsilon) \leq 2(1 - \varepsilon/2)^m \leq 2e^{-\varepsilon m/2}$$

8. *(4pts) Deduce that* $\mathcal{H}_{\text{seg}}$ *is PAC learnable with sample complexity* $m_{\mathcal{H}_{\text{seg}}}(\varepsilon, \delta) \leq \frac{2\ln(2/\delta)}{\epsilon}$.

9. *(4pts) Show that algorithm A can be implemented so that the computational complexity is polynomial in* $1/\epsilon$ *and in* $1/\delta$.

Exercise 6 (6pts) *Consider the class*

$$\mathcal{H}_{poly(k)} = \{p(x)\}$$

where the $p(x)$*'s denote degree* k *polynomials of the form* $p(x) = \sum_{j=0}^{k} a_j x^j$. *Show that performing a degree-k polynomial regression over* $\mathbb{R}$ *reduces to performing linear regression over* $\mathbb{R}^{k+1}$.

Exercise 7 (4pts) *Eve has a collection of points in* $\mathbb{R}^d$ *with binary labels. These points are non-separable but Eve nevertheless decides to run the Perceptron claiming that, depending on the initialization the Perceptron could achieve zero empirical error. Is she correct? (Justify)*

Exercise 8 (4pts) *Consider the k-NN algorithm and a given training set which consists of a collection of m points in $\mathbb{R}^d$ with binary labels. Suppose we set $k = m$ and test the algorithm over 10 fresh points. How many different labels do we get over these fresh points?*

Exercise 9 (4pts) *Consider the k-Means cost*

$$\sum_{i=1}^{k} \sum_{z \in C_i} ||z - \mu_i||_2^2$$

for a given set of clusters $C_1, C_2, \ldots, C_k$. Show that the update rule for each of the k centroids, that is $\mu_i \to \mu_i = \frac{1}{|C_i|} \sum_{z \in C_i} z$, achieves

$$\operatorname*{argmin}_{\mu} \sum_{z \in C_i} ||z - \mu||_2^2.$$

Exercise 10 (5pts, MICAS only) *After T iterations Adaboost produces some function*

$$\tilde{h}_T(x) \in L(B,T) = \{x \to sign(\sum_{i=1}^{T} w_t h_t(x)), w_t \in \mathbb{R}, h_t \in B\}$$

where B denotes some baseline class of functions.

Suppose we are in $\mathbb{R}^2$ and suppose that B consists of just two linear classifiers. Show through a simple example that $L(B,2)$ contains functions which are not in B.

Exercise 11 (5pts, MICAS only) *Explain why in general the error probability (test error) of AdaBoost does not tend to zero as its number of iterations tends to infinity.*