EE 660 **Homework 4 (Week 5)** Posted: Fri., 9/25/2020
Due: Fri., 10/2/2020, 5:00 PM PDT

1. Consider the email spam classification problem of Murphy Problem 8.1. Suppose you intend to use a linear perceptron classifier on that data. In the parts below, unless stated otherwise, assume the dataset of $N = 4601$ samples is split into $N_{Tr} = 3000$ for training and $N_{Test} = 1601$ for testing. Also, for the tolerance δ in the VC generalization bound, use 0.1 (for a certainty of 0.9). The parts below have short answers.

   Hint: You may use the relation that if $\mathcal{H}$ is a linear perceptron classifier in D dimensions (D features), $d_{VC}\left(\mathcal{H}\right) = D+1$. (This will be proved in Problem 2.)

   a) What is the VC dimension of the hypothesis set?

   b) Expressing the upper bound on the out-of-sample error as $E_{out}\left(h_g\right) \le E_{in}\left(h_g\right) + \varepsilon_{\text{vc}}$

   For $E_{in}\left(h_g\right)$ measured on the training data, use d_{vc} from part (a) to get a value for ε_{vc} .

   c) To get a lower ε_{vc}, suppose you reduce the number of features to $D = 10$, and also increase the training set size to 10,000. Now what is ε_{vc} ?

   d) Suppose that you had control over the number of training samples N_{Tr} (by collecting more email data). How many training samples would ensure a generalization error of $\varepsilon_{\text{vc}} = 0.1$ again with probability 0.9 (the same tolerance $\delta = 0.1$), and using the reduced feature set (10 features)?

   e) Instead suppose you use the test set to measure $E_{in}\left(h_g\right)$, so let's call it $E_{test}\left(h_g\right)$. What is the hypothesis set now? What is its cardinality?

   f) Continuing from part (e), use the bound:
   $E_{out}\left(h_g\right) \le E_{test}\left(h_g\right) + \varepsilon$

   Use the original feature set and the original test set, so that $N_{Test} = 1601$. Give an appropriate expression for ε and calculate it numerically.

2. AML Exercise 2.4 (page 52). In addition to the hints given in the book, you can solve the problem by following the steps outlined below (on the next page).

For part (a):

i. Write a point $\underline{x}_i$ as a $d+1$ dimensional vector;

ii. Construct the $(d+1)\times(d+1)$ matrix suggested by the book;

iii. Write $\underline{h}(\underline{\underline{X}})$, the output of the perceptron, as function of $\underline{\underline{X}}$ and the weights $\underline{w}$ (note that $\underline{h}(\underline{\underline{X}})$ is a $d+1$ dimensional vector with elements +1 and -1);

iv. Using the nonsingularity of $\underline{\underline{X}}$, justify how any $\underline{h}(\underline{\underline{X}})$ can be obtained.

For part (b):

i. Write a point $\underline{x}_k$ as a linear combination of the other $d+1$ points;

ii. Write $h(\underline{x}_k)$ (output for the chosen point) and substitute the value of $\underline{x}_k$ by the expression just found on the previous item (**Hint**: use the $\text{sgn}\{\bullet\}$ function);

iii. What part of your expression in (ii) determines the class assignment of each point $\underline{x}_i$, for $i \neq k$?

iv. You have just proven (part (a)) that $\underline{h}(\underline{\underline{X}})$ with $\underline{\underline{X}}_{(d+1)\times(d+1)}$ can be shattered. When we add a $(d+2)^{\text{th}}$ line to $\underline{\underline{X}}$ can it still be shattered? In other words, can you choose the value of $h(\underline{x}_k)$? Justify your answer. **Hint:** you can choose the class label of the other $(d+1)$ points.

3. AML Problem 2.24 (page 75), except

>> Replace part (a) with:

(a.1) For a single given dataset, give an expression for $g^{(\mathcal{D})}(x)$. (AML notation)

(a.2) Find $\bar{g}(x)$ analytically; express your answer in simplest form.

>> For parts (b) and (c), obtain $\text{E}_{\mathcal{D}}\{E_{out}\}$ by direct numerical computation, not by adding bias and var.

>> For part (d), obtain bias(x), var(x), bias, var, and $\text{E}_{\mathcal{D}}\{E_{out}\}$, all by analytical (pencil and paper) techniques.

Reading

AML Sections 4.0, 4.1 (pp. 119-126): *Overfitting*
AML Section 4.2 (pp. 126-137): *Regularization (AML perspective)*

Problem based on the reading

4. AML Exercise 4.3 (p. 125). part (a) - first question only; part (b) - first question only. [Think about the second question of each part if you like; there isn't necessarily a single answer to each, though.]
5. AML Exercise 4.5 (p. 131)