EE 660 **Homework 3 (Week 6)** Posted: Sat., 10/1/2022

v1 Due: Tues., 10/11/2022, 11:59 PM

Please note: in AML, **Exercises** are embedded in the text of each section; **Problems** are at the end of each chapter. Please be sure you work the assigned problem or exercise to get credit.

1. AML Problem 1.7a,b (p. 36), plus the following:

   For part (b), add: also plot for $N = 60$.

   Hint: in this problem, tossing one coin *N* times corresponds to one draw of a dataset of size *N*. Doing this with 1000 coins independently, corresponds to drawing 1000 different datasets, each of size *N*.

2. Suppose you have trained a model in a binary classification problem, and you want to estimate its error based on a test set.

   You want to estimate this error empirically for the case where labeled data is expensive. So you decide to first do some simulations to see how your test-set error might vary with the "luck of the draw" of the test set.

   Let the true probability of error on your model be $E_{out}(h) = \mu$. Because μ is unknown to you, for purposes of simulation you will try different values of μ.

   Method: Conceptually, a test set $\mathcal{D}^{(N)}$ is created by drawing *N* data points randomly from the input space, with replacement. An expert could correctly label each data point, and then you could color each data point as "correct" or "incorrect" depending on the ML classifier's prediction.

   You decide to simulate this by drawing (colored) data points randomly, with replacement, from a bin of "correct" and "incorrect" data points, with $P(\text{incorrect}) = \mu$.

   (a) Let $\mu = 0.30$ and $N = 10$.

   (i) Draw a colored dataset $\mathcal{D}^{(10)}$ of size $N = 10$. From your 10 drawn data points, compute the error rate $E_{\mathcal{D}^{(10)}}(h)$. Is it equal to 0.30? Explain why or why not.

   (ii) Calculate theoretically, using the given values of μ and *N*, the probability $P\left(E_{D^{(10)}}(h) = \mu\right)$ for any draw of *N* data points, rounded to 2 decimal places (0.*xx*).

   (iii) Give a theoretical expression for the probability $\mathrm{P}\left(E_{D^{(N)}}(h) = \mu\right)$ in terms of variables $\mu,\ N$. Assume that μN is an integer.

(b) Repeat the experiment of part (a)(i) 100 times. Keep a record of the error rate $E_{\mathcal{D}^{(10)}}(h)$ from each run (no need to turn in these 100 values). Give the following statistics and answer these questions:

(i)

$$\max\left\{E_{\mathcal{D}^{(10)}}(h)\right\},$$
$$\min\left\{E_{\mathcal{D}^{(10)}}(h)\right\},$$
$$\text{sample mean}\left\{E_{\mathcal{D}^{(10)}}(h)\right\},$$
$$\text{sample standard deviation}\left\{E_{\mathcal{D}^{(10)}}(h)\right\}.$$

(ii) How many of your 100 runs had an error rate different than μ? Does this agree with the value of $P\left(E_{D^{(10)}}(h) = \mu\right)$ from item (a)(ii) ?

(iii) From your results of the 100 runs, give an estimate for the probability

$$P\left(\left|E_{D^{(N)}}(h) - \mu\right| < 0.05\right)$$

(c) Repeat parts (a)(ii) and (b) for the following parameters:

(i) $\mu = 0.10,\ N = 10$

$\mu = 0.10,\ N = 100$

(ii) $\mu = 0.30,\ N = 100$

(iii) $\mu = 0.50,\ N = 10$

$\mu = 0.50,\ N = 100$

Tip: present your results in a table (including values given in (a)) for easy interpretation. A sample table is as below.

μ	N	$\max\{E_D\}$	$\min\{E_D\}$	sample mean $\{E_D\}$	sample std $\{E_D\}$	# runs $E_D \neq \mu$	$P(\lvert E_D - \mu\rvert < 0.05)$
0.1	10						
0.1	100						
0.3	10						
0.3	100						
0.5	10						
0.5	100						

(d) Comment on your results, in the following:

(i) How does the accuracy of your error-rate estimate from a test dataset, vary with N? **Hint:** look at min, max, std, and $\mathrm{P}\left(\left|E_{D^{(N)}}(h) - \mu\right| < 0.05\right)$.

(ii) For the classifier of (c)(iii):

⇒ based on the given true error rate (P(incorrect)), did the classifier learn anything?

⇒ How many test datasets of your draws in (c)(iii) gave an error rate indicating that the classifier did learn something, assuming that $E_{D^{(N)}}(h) \le 0.45$ or $E_{D^{(N)}}(h) \ge 0.55$ means it learned something?

3. Consider a hypothesis set $\mathcal{H}$ consisting of all positive-interval hypotheses and all negative-interval hypotheses. (See AML pp.43-44, Example 2.2, *positive intervals*, for a definition of positive interval hypotheses.) Negative-interval hypotheses return $h = -1$ within the interval and $h = +1$ outside the interval. Thus, any hypothesis h in $\mathcal{H}$ contains one interval, and that interval can be positive or negative.

Hint: before answering the questions below, you may find it helpful to read or review AML Example 2.2, cases 1 and 2, or Discussion 5 notes.

(a) Find the growth function $m_{\mathcal{H}}(\mathrm{N})$ as a function of N. Show your work or reasoning.

Tip: you may use the AML book's result for negative intervals, and build on top of that.

(b) Give the smallest break point k of $\mathcal{H}$. Briefly justify your answer.

(c) Give the VCdim of $\mathcal{H}$. Briefly justify your answer.

4. Consider the WiFi localization dataset used in Homework 2: it had $N = 2000$ data points and 7 features, split into a training set of 1500 data points and a test set of 500 data points. Suppose you intend to use a linear perceptron classifier on that data instead of random forest. In the parts below, 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 without proof the relation that if $\mathcal{H}$ is a linear perceptron classifier in D dimensions (D features), $d_{VC}(\mathcal{H}) = D+1$.

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

b) Expressing the upper bound on the out-of-sample error as

$$E_{out}(h_g) \le E_{in}(h_g) + \varepsilon_{\mathrm{vc}}$$

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 $D = 4$. Now what is ε_{vc} ?

d) Suppose that you had control over the number of training samples N_{Tr} (by collecting more WiFi 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 (4 features)?
Hint: if you're not sure how to solve for *N*, see an example in AML Sec. 2.2.1 ("Sample Complexity", pp.57).

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:
$E_{out}\left(h_g\right) \leq E_{test}\left(h_g\right) + \varepsilon$
Use the original feature set and the original test set, so that $N_{Test} = 500$. Use the version of ε that gives the tightest bound. Give the appropriate expression for ε and calculate it numerically.
Does the number of features have any effect on this ε ?

5. You are given a regression problem in 1D. The target function is

$$\mathrm{f}(x) = \sigma(3x) = \frac{e^{3x}}{1 + e^{3x}}$$

and feature space extends over $-1 \leq x \leq +1$. $p(x)$ is a uniform distribution over the feature space.

Each draw of the dataset consists of $N = 2$ points:

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2)\} = \{(x_1, \sigma(3x_1)), (x_2, \sigma(3x_2))\} .$$

The hypothesis set consists of all lines $h(x) = ax + b$.

In this problem you will explore bias and var numerically.

You may use built-in python functions, including functions in NumPy to draw random numbers, such as numpy.random.uniform and numpy.random.normal.

(a) Give an algebraic expression for the best hypothesis $h_g^{(\mathcal{D})}$, in terms of x_1, x_2, y_1, y_2 . This is the hypothesis that minimizes the MSE on $\mathcal{D}$.
Hint: no need to take derivatives and do a formal minimization.

(b) Find the mean best hypothesis $\overline{h_g}$ numerically. **Tip**: average over many draws of $\mathcal{D}$. Give resulting a and b. Draw a plot containing curves of $f(x)$, $\overline{h_g}$, and several $h_g^{(D)}$ over the range of $-1 \leq x \leq +1$. The number of $h_g^{(D)}$ curves can be determined yourself by best visibility.

(c) Numerically compute bias and var. Also numerically compute $\mathrm{E}_{\mathrm{D}}\{E_{out}\left(h_g^{(\mathcal{D})}\right)\}$.

(d) Now let there be sensor noise on the data, so that

$$y(x) = f(x) + \epsilon_n, \quad \epsilon_n \sim N(\epsilon_n \mid \mu_n = 0, \sigma_n = 0.004)$$

Repeat parts (a)-(c) for this noisy data. **Tip:** if there is no change in (a), just state so.

Is var of the learned system very sensitive to σ_n ? Conjecture a reason why or why not.

(e) Now let there also be some sensor bias on the data, so that

$$y(x) = f(x) + \epsilon_n, \quad \epsilon_n \sim N(\epsilon_n \mid \mu_n = 0.1, \sigma_n = 0.004)$$

Repeat parts (a)-(c) for this biased and noisy data. **Tip:** if there is no change in (a), just state so.

What is the effect of the sensor bias on the bias and var of the learned system (i.e., does it increase, decrease, or have no effect, on bias and on var)?