EE660 Week4 HW Pin-Hsuan Lee 1139093874

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

Problem 1.7 A sample of heads and tails is created by tossing a coin a number of times independently. Assume we have a number of coins that generate different samples independently. For a given coin, let the probability of heads (probability of error) be μ. The probability of obtaining k heads in N tosses of this coin is given by the binomial distribution:

$$P[k \mid N, \mu] = \binom{N}{k} \mu^k (1-\mu)^{N-k}.$$

Remember that the training error ν is $\frac{k}{N}$.

(a) Assume the sample size (N) is 10. If all the coins have $\mu = 0.05$ compute the probability that at least one coin will have $\nu = 0$ for the case of 1 coin, 1,000 coins, 1,000,000 coins. Repeat for $\mu = 0.8$.

(b) Take the scenario of part (a), for the case of 1,000 coins, and $\mu = 0.05$. Consider the following interpretation in applying it in a machine learning setting.

There is one hypothesis that is given (one decision boundary and corresponding set of decision regions, or one decision rule); call it h. The out of sample error is $E_{out}(h) = 0.05$, and the in-sample error depends on the dataset drawn.

Hint: The number of tosses of a coin, $N = 10$, corresponds to the size of a dataset.

Complete the machine-learning interpretation by answering the following:

(i) What do the 1000 coins represent?

(ii) What does the calculation in part (a), for 1000 coins and $\mu = 0.05$, represent?

(iii) In this interpretation, take the most general version of the Hoeffding inequality in Ch. 1:

$$P[|\nu - \mu| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

Give values (or expressions) for μ, ν, and M.

$P[K \mid N, \mu] = \binom{N}{K} \mu^K (1-\mu)^{N-K}$

(a) $P[\text{no heads}] = \binom{10}{0}(0.05)^0(1-0.05)^{10-0} = 0.95^{10} \doteqdot 0.5987 \rightarrow$ 1 coin case (when $\mu = 0.05$)

$P\left[\begin{matrix}\text{at least 1 coin have } \nu=0 \\ \text{for the case of 1000 coins}\end{matrix}\right] = 1-(1-0.5987)^{1000} \doteqdot 1$

$P\left[\begin{matrix}\text{at least 1 coin have } \nu=0 \\ \text{for the case of } 10^6 \text{ coins}\end{matrix}\right] = 1-(1-0.5987)^{10^6} \doteqdot 1$ (when $\mu = 0.05$)

when $\mu = 0.8$:

$P[\text{no heads}] = \binom{10}{0}(0.8)^0(1-0.8)^{10-0} = 1.024 \times 10^{-7} \rightarrow$ 1 coin case

$P\left[\begin{matrix}\text{at least 1 coin have } \nu=0 \\ \text{for the case of 1000 coins}\end{matrix}\right] = 1-(1-1.024\times10^{-7})^{1000} = 1.0239\times10^{-4}$

$P\left[\begin{matrix}\text{at least 1 coin have } \nu=0 \\ \text{for the case of } 10^6 \text{ coins}\end{matrix}\right] = 1-(1-1.024\times10^{-7})^{10^6} = 0.0973$

(b)

(i) 1000 coins represent the total number of hypothesis.

(ii) It means that in 1000 hypothesis, the probability of at least one hypothesis will have in-sample error $E_{in}(h) = 0$

(iii) $P[|\nu - \mu| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$

$P[|E_{in}(h) - \underbrace{E_{out}(h)}_{0.05}| > \epsilon] \leq 2\underbrace{M}_{1000}e^{-2\epsilon^2 N}$, $N = N = 10$

2. AML Problem 2.1 (p. 69).

Problem 2.1 In Equation (2.1), set $\delta = 0.03$ and let

$$\epsilon(M, N, \delta) = \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}.$$

(a) For $M = 1$, how many examples do we need to make $\epsilon \le 0.05$?

(b) For $M = 100$, how many examples do we need to make $\epsilon \le 0.05$?

(c) For $M = 10{,}000$, how many examples do we need to make $\epsilon \le 0.05$?

$\epsilon(M, N, \delta) = \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$ $\delta = 0{,}03$

(a) $M = 1$

$\epsilon^2 = \frac{1}{2N} \ln \frac{2\times 1}{0{,}03} \leqslant 0{,}05^2 \Rightarrow \ln \frac{2}{0{,}03} = 4{,}1997 \leqslant 0{,}05^2 \times 2N$

$\Rightarrow N \geqslant \frac{4{,}1997}{0{,}05^2 \times 2} = 839{,}94$

$\Rightarrow$ we need at least 840 samples.

(b) $M = 100$

$\epsilon^2 = \frac{1}{2N} \ln \frac{2\times 100}{0{,}03} \leqslant 0{,}05^2 \Rightarrow \ln \frac{200}{0{,}03} \leqslant 0{,}05^2 \times 2N$

$\Rightarrow N \geqslant (\ln \frac{200}{0{,}03}) / (0{,}05^2 \times 2) = 1760{,}975$

$\Rightarrow$ we need at least 1761 samples.

(c) $M = 10^4$

$\epsilon^2 = \frac{1}{2N} \ln \frac{2\times 10^4}{0{,}03} \leqslant 0{,}05^2 \Rightarrow \ln \frac{2\times 10^4}{0{,}03} \leqslant 0{,}05^2 \times 2N$

$\Rightarrow N \geqslant (\ln \frac{2\times 10^4}{0{,}03}) / (2 \times 0{,}05^2) = 2682{,}009$

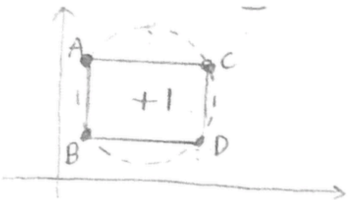
$\Rightarrow$ we need at least 2683 samples.

3. AML Problem 2.2 (p. 69)

Problem 2.2 Show that for the learning model of positive rectangles (aligned horizontally or vertically), $m_{\mathcal{H}}(4) = 2^4$ and $m_{\mathcal{H}}(5) < 2^5$. Hence, give a bound for $m_{\mathcal{H}}(N)$.

Assume feature space is 2D. A "positive rectangle" is a rectangle-shaped decision boundary, and has value (label) +1 inside and value (label) –1 outside. The sides of the rectangle are parallel to the coordinate axes.

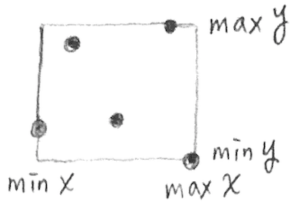
Hint for the last question: the bound is a polynomial in *N*.



$m_H(4) = 16$

A	B	C	D
+	+	+	+
–	–	–	–
+	–	–	–
–	+	+	+
–	+	–	–
+	–	+	+
–	–	+	–
+	+	–	+
+	+	+	–
–	–	–	+
+	+	–	–
–	–	+	+
+	–	+	–
–	+	–	+
+	–	–	+
–	+	+	–

$m_H(5)$



As we can see from the plot when N=5, all of the points will locate inside the retangle. Therefore, it's impossible for us to have 2^5 dichotomies which means that $m_H(5) < 2^5$.

From above, we can know the break point $K=5$. Since $d_{vc} = K-1 = 4$, from A.M.L (2.10) $\Rightarrow m_H(N) \leq N^{d_{vc}}+1$

$$\Rightarrow \underline{m_{\mathcal{H}}(N) \leq N^4+1}$$

4. [Based on AML Exercise 2.1, p. 45.]:

Exercise 2.1

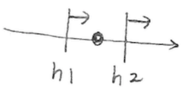
By inspection, find a break point k for each hypothesis set in Example 2.2 (if there is one). Verify that $m_{\mathcal{H}}(k) < 2^k$ using the formulas derived in that Example.

(a) Find the smallest break point *k* for the hypothesis set consisting of Positive Rays (defined in Example 2.2).

(b) Find the smallest break point *k* for the hypothesis set consisting of Positive Intervals (defined in Example 2.2).

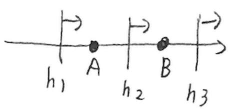
(a) Positive Rays

① when $N=1 \Rightarrow$ can be shattered



$h_1 : (+)$
$h_2 : (-)$

② when $N=2 \Rightarrow$ cannot be shattered

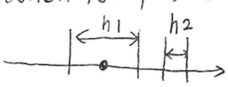


$h_1 : (+,+)$
$h_2 : (-,+)$
$h_3 : (-,-)$

$\Rightarrow$ the smallest break point $K = 2$

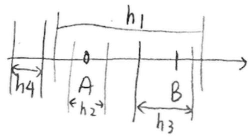
(b) positive intervals

① when $N=1 \Rightarrow$ can be shattered



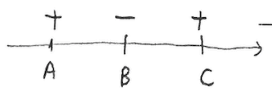
$h_1 : (+)$
$h_2 : (-)$

② when $N=2 \Rightarrow$ can be shattered



$h_1 : (+,+)$
$h_2 : (+,-)$
$h_3 : (-,+)$
$h_4 : (-,-)$

③ when $N=3 \Rightarrow$ cannot be shattered



$\rightarrow$ it can't happen

$\Rightarrow$ the smallest break point $K=3$

5. AML Exercise 2.6 (p. 60).

Exercise 2.6

A data set has 600 examples. To properly test the performance of the final hypothesis, you set aside a randomly selected subset of 200 examples which are never used in the training phase; these form a test set. You use a learning model with 1,000 hypotheses and select the final hypothesis g based on the 400 training examples. We wish to estimate $E_{\text{out}}(g)$. We have access to two estimates: $E_{\text{in}}(g)$, the in sample error on the 400 training examples; and, $E_{\text{test}}(g)$, the test error on the 200 test examples that were set aside.

(a) Using a 5% error tolerance ($\delta = 0.05$), which estimate has the higher 'error bar'?

(b) Is there any reason why you shouldn't reserve even more examples for testing?

600 examples < 200 examples : test data
400 examples : train data → select final g

1000 hypotheses ⇒ $M = |\mathcal{H}| = 1000$

5% error tolerance ⇒ $\delta = 0.05$

(a)

① $E_{in}(g)$:

$$E_{out}(g) \le \underbrace{E_{\mathcal{D}}}_{E_{in}}(g) + \sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}} \Rightarrow E_{out}(g) \le E_{\mathcal{D}}(g) + \sqrt{\frac{1}{2\times 400}\ln\frac{2\times 1000}{0.05}}$$

$$\Rightarrow E_{out}(g) \le E_{\mathcal{D}}(g) + 0.115$$

② $E_{test}(g)$:

$$E_{out}(g) \le E_{test}(g) + \sqrt{\frac{1}{2N'}\ln\frac{2M'}{\delta}} \Rightarrow E_{out}(g) \le E_{test}(g) + \sqrt{\frac{1}{2\times 200}\ln\frac{2\times 1}{0.05}}$$

$$\Rightarrow E_{out}(g) \le E_{test}(g) + 0.096$$

from ① & ②:

$E_{in}(g)$ has a higher "error bar"

(b) Since we can only use the training data to get the final hypothesis g. If we have more data on training set, we can have more probability to get a better final hypothesis g.

6. AML Exercise 2.8 (p. 63). Note that g in AML notation is h_g in our class notation (= best chosen hypothesis in $\mathcal{H}$).

Exercise 2.8

(a) Show that if $\mathcal{H}$ is closed under linear combination (any linear combination of hypotheses in $\mathcal{H}$ is also a hypothesis in $\mathcal{H}$), then $\bar{g} \in \mathcal{H}$.

(b) Give a model for which the average function $\bar{g}$ is not in the model's hypothesis set. *[Hint: Use a very simple model.]*

(c) For binary classification, do you expect $\bar{g}$ to be a binary function?

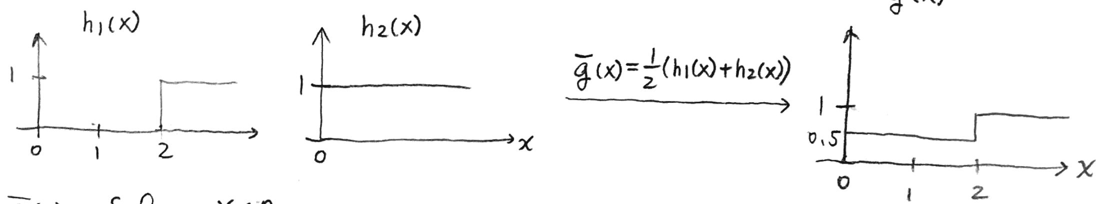
(a) $g_N(x) = \sum_{i=0}^{K} C_i X_i$ C_i : constant , K : integer and $K > 0$, $g_N \in \mathcal{H}$

$\mathcal{H}$ is closed under linear combination.

$\Rightarrow \bar{g} = \frac{1}{N} \sum_{k=1}^{N} g_N \Rightarrow \bar{g} \in \mathcal{H}$ QE.D

(b) $\mathcal{H} = \{h_1, h_2\}$, $\mathcal{H}$ only has the output $y = 0$ or 1.

$h_1(x) \begin{cases} 1 , & x \geq 2 \\ 0 , & \text{otherwise} \end{cases}$ $h_2(x) \begin{cases} 1 , & x \geq 0 \\ 0 , & \text{otherwise} \end{cases}$



$\bar{g}(x) = \begin{cases} 0 , & x < 0 \\ 0.5 , & 0 \leq x < 2 \\ 1 , & x \geq 2 \end{cases}$ $\Rightarrow \bar{g}$ is not in the hypothesis set.

(c) from $\bar{g}(x)$, we can't expect $\bar{g}$ to be a binary function.