Homework 3
CSCI4100

Han Hai
Rin:661534083
haih2@rpi.edu

September 24 2018

1. **[100 points] Exercise 1.13** Consider the bin model for a hypothesis h that makes an error with probability μ in approximating a deterministic target function f (both h and f are binary functions). If we use the same h to approximate a noisy version of f given by
   $P(y|x) = \lambda$ if y=f(x),
   $P(y|x) = 1 - \lambda$ if y $\neq$ f(x)

   (a) What is the probability of error that h makes in approximating y?
   $P(e) = \mu\lambda + (1-\lambda)(1-\mu)$

   (b) At what value of λ will the performance of h be independent of μ?
   When λ is 0.5
   $0.5\mu + 0.5 - 0.5\mu = 0.5$ it's independent of μ

2. **[100 points] Exercise 2.1** By inspection, find a break point k for each hypothesis set in Example 2.2(if there is one). Verify that $m_H k < 2^k$ using the forumlas derived in that example.

   (a) positive rays: the break points is k+1
   the minimum k for this to work is 2 2+1=3 $2^2 = 4$ so k+1 $< 2^k$ for k=2

   (b) positive intervals:
   for N=1 and N=2 $0.5N^2 + 0.5N + 1 = 2^N$
   for N=3 it's $7 < 8$ therefore, the break point is 3

   (c) convex sets:
   there isn't a break point for this one since 2^N is the formula

3. **[100 points Exercise 2.2]**

   (a) Verify the bound of Theorem 2.4 in the three cases of Example 2.2:

      i. positive rays: H consists of all hypotheses in one dimension of the form h(x)=sign(x-a).
      From 2.1 we know k=2, so
      $\sum_{i=0}^{2-1} Nci = 1 + N$ therefore the theorem hold in this case

ii. Positive intervals: H consists of all hypotheses in one dimension that are positive within some interval and negative elsewhere.
From 2.1 we know k=3,so

$\sum_{i=0}^{3-1} Nci = 1 + \frac{1}{2}N^2 + \frac{1}{2}N$
here, the theorem hold as well

iii. Convex sets
k doesn't exist here so the theorem doesn't apply

(b) Doe there exist a hypothesis set for which $m_H(N) = N + 2^{[N/2]}$ where N/2 is the largest integer $\leq N/2$
No, because $m_H(N)$ is bounded by polynomial, an exponential term seems unreasonable

4. **[200 points] Exercise 2.3** Compute the VC dimension of H for the hypothesis sets in parts 1 2 and 3 of exercise 2.2

(a) for this one d_{VC}= 1, since 2 is the break point

(b) d_{VC}=2, since 3 is the break point

(c) d_{VC} is ∞ since break point does not exist

5. **[100 points] 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 1000 hypotheses and select the final hypothesis g based on the 400 training examples. We wish to estimate $E_{out}(g)$. We have access to two estimates:$E_{in}(g)$, the in-sample error on the 400 training access to two estimates; and, Etest(g), the test error on the 200 test examples that were set aside.

(a) Using a 5 % error tolerance ($\lambda = 0.05$) which estimates has the higher error bar?
let$\lambda = 2Me^{-2\epsilon^2 N}$
then, $\epsilon = \sqrt{\frac{1}{2N} ln(\frac{2M}{\lambda})}$
plus in the numbers, Etraining with a smaller N though large M will have larger error bar

(b) Is there any reason why you shouldn't reserve even more examples for testing?
Yes, a even larger testing data will result in a even smaller training data, that might have a negative effect

6. **[200 points] Problem 1.11** The matrix which tabulates the cost of various error for the CIA and Supermarket applications in Example 1.1 is called a risk or loss matrix.
For the two risk matrices in Example 1.1, explicitly write down the in-sample error E_{in} that one should minimize to obtain g. This in-sample error should weight the different types of errors based on the risk matrix.

for the first matrix supermarket:
$E_{in} = 10 * \frac{1}{N} \sum_{n=1}^{N} [h(x_n) = -1, f(x_n) = +1] + \frac{1}{N} \sum_{n=1}^{N} [h(x_n) = +1, f(x_n) = -1]$

for the second matrix CIA:
$E_{in} = \frac{1}{N}\sum_{n=1}^{N}[h(x_n) = -1, f(x_n) = +1] + 1000 * \frac{1}{N}\sum_{n=1}^{N}[h(x_n) = +1, f(x_n) = -1]$

7. **[200 points]Problem 1.12** The problem investigates how changing the error measure can change the result of the learning process. You have N data points $y_1 \le ....y_N$ and wish to estimate the "representative" value.

   (a) If you algorithm is to find the hypothesis h that minimizes the in-sample sum of squared deviations,
   $E_{in}(h) = \sum_{n=1}^{N}(h - y_n)^2$,
   show that your estimate will be the in-sample mean,
   $h_{mean} = \frac{1}{N}\sum_{n=1}^{N} y_n$

   take the derivative of Ein $E'_{in} = 2\sum_{n=1}^{N}(h - y_n)$ when $E'_{in} = 0$ E_{in} is minimize, so $h_n = y_n$ would minimize, the hypothesis $h_{mean} = \frac{1}{N}\sum_{n=1}^{N} y_n$ can get close to y_n

   (b) If your algorithm is to find the hypothesis h that minimizes the in-sample sum of absolute deviations $E_{in} = \sum_{n=1}^{N}|h - y_n|$, show that the median will be the estimate
   again take the derivatie $E'_{in} = \sum_{n=1}^{N}|h - y_n|/h - y_n$ it's either 1 or -1, so when half of the number is positive above y_n and half of the number is below, E'_{in} 1+1-1+1-1... is zero and E_{in} is minimize, that is the the median number

   (c) Suppose y_n is perturbed to $y_N + \epsilon$ where $\epsilon \to \infty$ So, the single data point $y_N$becomes an outlier. What happens to your two estimators hmeans and hmedian?

   hmedian is unaffected, because median number is robust, while hmean $\to \infty$, so hmean is perturbed