Machine Learning Hw2

r03922145 Yi Huang

Question 1

when $y = f(x)$, the probability of hypothesis h makes an error is μ, and $P(y|x) = \lambda$. when $y \neq f(x)$, the probability of hypothesis h makes an error $1 - \mu$, $P(y|x)$ equals to $1 - \lambda$. Thus, the probability of error this h makes in approximating the noisy target y is $\lambda\mu + (1 - \lambda)(1 - \mu)$.

Question 2

when λ equals to 0.5, then chance of choosing $y = f(x)$ is the same with not choosing $y = f(x)$, the probability of error this h makes in approximating the noisy target y would always be 0.5, the performance of h is independent of μ.

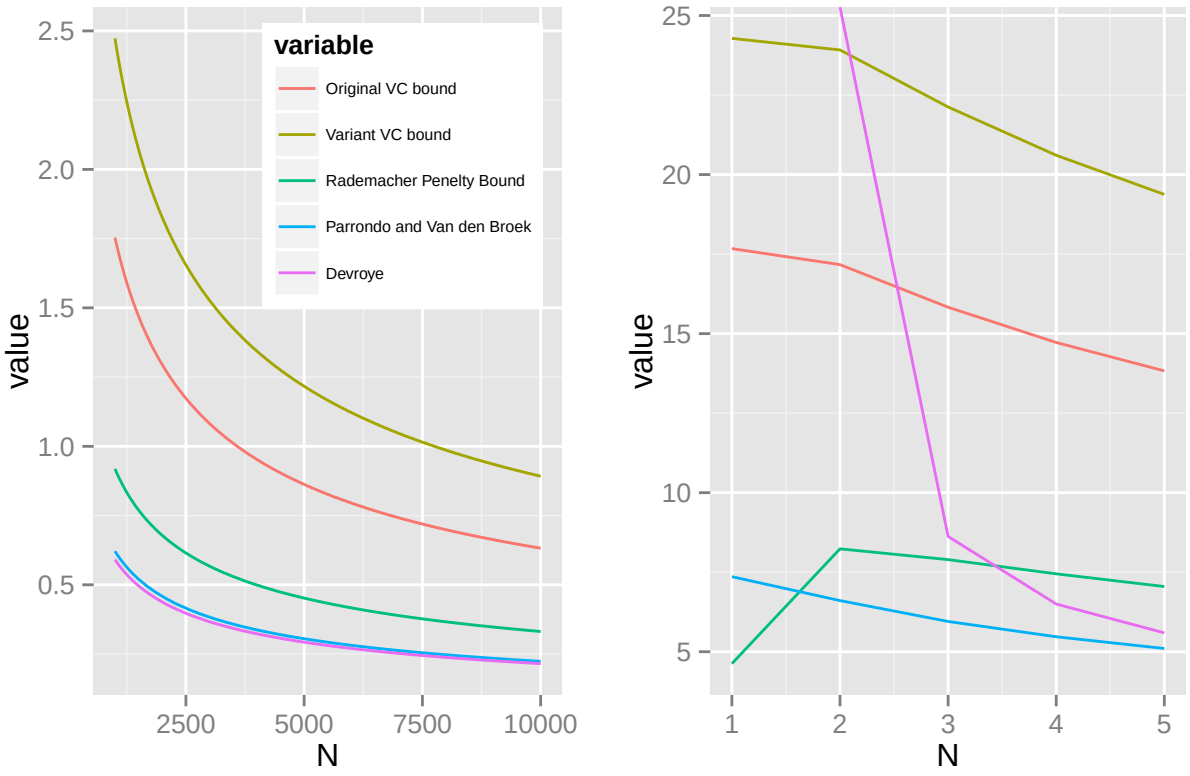
Question 3

Given specs $\epsilon = 0.05$,$\delta = 0.05$ and $d_{vc} = 10$, want $4(2N)^{d_{vc}} exp(-\frac{1}{8}\epsilon^2 N) \leq \delta$, then N should be closet to 460000.

N	δ
420000	697.7536261
440000	2.1448427
460000	0.0064581
480000	0.0000191
500000	0.0000001

Question 4

Given specs $\delta = 0.05$,$d_{vc} = 50$ and $N = [1 : 10000]$, the error ϵ bounds is showed below as a function of N. When N is very large, like N = 10000(showed in the left graph), the Devroye bound is the tightest.



Question 5

When N is small, like N = 5(showed in the right graph), the Parrondo and Van den Broek bound is the tightest.

Question 6

The growth function $m_H(N)$ of 'positive-and-negative intervals on $\mathbb{R}$' can be deducted from the growth function of 'positive intervals'. We know that the growth function of 'positive intervals' is $\binom{N+1}{2}+1$, this growth function can be explained by choosing 2 positions among N+1 positions plus one situation of all 'x'. 'negative intervals' produces many situation that is already in 'positive intervals', only when the 2 choosing positions locate in the middle, 'negetive intervals' can add new situations to 'positive intervals'. For example, when N=3, choosing the middle two position would add $\circ \times \circ$, when N=4, choosing 2 position in the middle would add $\circ \times \times \circ$, $\circ \circ \times \circ$ and $\circ \times \circ \circ$. Therefore, the growth function $m_H(N)$ of 'positive-and-negative intervals on $\mathbb{R}$' would be $\binom{N+1}{2}+1+\binom{N-1}{2}$.

Question 7

Since the growth function $m_H(N)$ of 'positive-and-negative intervals on $\mathbb{R}$' is N^2-N+2, when $N \leq 3$,$m_H(N)=2^N$,when N = 4, $m_H(N)<2^N$, the VC-dimension of 'positive-and-negative intervals on $\mathbb{R}$' is 3.

Question 8

In 'positive donuts of $\mathbb{R}^2$', we assume $0<a<b<\infty$, since each hypothesis is +1 within a 'donut' region of $a^2 \leq x_1^2+x_2^2 \leq b^2$, the label of one points is determined by its distance to the origin, which is $\sqrt{x_1^2+x_2^2}$.

This means that the 'positive donuts in $\mathbb{R}^2$' is the same as 'positive interval in $\mathbb{R}$'. If the distance of one point to origin is within the interval$[a, b]$, this point would be labeled $+1$. So the growth function for 'positive donuts' is the same as 'positive interval', is $\binom{N+1}{2} + 1$.

Question 9

For the set of **ordered** points $\{x_1, x_2, ..., x_{D+1} \mid x_i < x_{i+1}\}$, and a set of thresholds $\{l_1, l_2, ..., l_D\}$ in which $x_i < l_i < x_{i+1}$ and $\{a_1, a_2, ..., a_D \mid a_i \in \{0, 1\}\}$, $s \in \{-1, +1\}$, we can reformat the $\mathcal{H}$ into such form:

$$\begin{aligned}\mathcal{H} &= \{h_l | h_l(x) = s \cdot sign((l_1 \cdot (-\frac{1}{l_1})^{1-a_1} - a_1 x) \cdot (l_2 \cdot (-\frac{1}{l_2})^{1-a_2} - a_2 x) \cdot \ ... \cdot (l_D \cdot (-\frac{1}{l_D})^{1-a_D} - a_D x))\} \\ &= \{h_l | h_l(x) = s \cdot sign(\prod_{i=1}^{D} (l_i \cdot (-\frac{1}{l_i})^{1-a_i} - a_i x))\}\end{aligned}$$

By changing the value of a and s, the $D+1$ points can be shattered. Choosing any labeling $\{y_1, y_2, ..., y_{D+1}\}$, let

$$\begin{aligned} s &= sign(y_1), \\ a_i &= \begin{cases} 0 & , y_i = y_{i+1} \\ 1 & , y_i \neq y_{i+1} \end{cases}\end{aligned}$$

Assuming there are k times of sign conversion before y_j,

$$\begin{aligned} h_l(x_j) &= sign(y_1) \cdot sign((l_1 \cdot (-\frac{1}{l_1})^{1-a_1} - a_1 x_j) \cdot (l_2 \cdot (-\frac{1}{l_2}-)^{1-a_2} - a_2 x_j) \cdot \ ... \cdot (l_D \cdot (-\frac{1}{l_D})^{1-a_D} - a_D x_j)) \\ &= sign(y_1) \cdot sign((+1)^{j-k}(-1)^k(+1)^{D-j}) \\ &= y_j \end{aligned}$$

Thus $h_l(x_j) = y_j$ for all j, so $\{x_1, x_2, ..., x_{D+1}\}$ can be shattered by $\mathcal{H}$. And the times of sign conversion before y_j cannot be more than D times, $\mathcal{H}$ cannot shatter $D+2$ points because the last point will not longer be independent from the previous ones.So the VC dimension of "polynomial discriminant" hypothesis set of degree D on $\mathbb{R}$ is $D+1$.

Question 10

The number of hyper-rectangular regions $= 2^d$. Whether a hyper-rectangular region should be +1 or -1 is independent from each other and decided by 2^d free parameters. So the VC-dimension of "simplified decision trees" is 2^d.

Question 11

The 'triangle waves' hypothesis set on $\mathbb{R}$ is actually a square-waves function after taking the sign of argument. By changing the value of α, the width of each wave can be changed to produce any labeling for x. It can be showed that for any l, the set of points$\{x_1, x_2, ..., x_l\}$ can be shattered. Consider the set of points given by $x_j = 4^{-j}$, choosing any labeling $\{y_1, y_2, ..., y_l\}$, let

$$\alpha = \sum_{i=1}^{l}(1-y_i)\cdot 4^i$$

.Then,

$$\begin{aligned} h(x_j) &= sign(|(\sum_{i=1}^{l}(1-y_i)\cdot 4^i\cdot 4^{-j})\ mod\ 4-2|-1) \\ &= sign(|(\sum_{i=1}^{l}(1-y_i)\cdot 4^{i-j})\ mod\ 4-2|-1) \end{aligned}$$

For any y_i =1, the corresponding term in the summation will be zero and if $i > j$, we will add multiple of 4 to the mod function which causes no change in value. Therefore, these terms can be dropped from the summation. Now,

$$\begin{aligned} h(x_j) &= sign(|(\sum_{i:i\le j, y_i=-1}(1-y_i)\cdot 4^{i-j})\ mod\ 4-2|-1) \\ &= sign(|(1-y_j+\sum_{i:i<j, y_i=-1}2\cdot 4^{i-j})\ mod\ 4-2|-1) \end{aligned}$$

It is easy to show that the summation term is always less than 1. As a result, if $y_j = 1$, the value of αx is between 0 and 1, then $h(x_j) = 1 = y_j$. If $y_j = -1$, the value of αx is between 2 and 3, then $h(x_j) = -1 = y_j$. Thus $h(x_j) = y_j$ for all j. So the set$\{4^{-1}, 4^{-2}, ..., 4^{-l}\}$ can be shattered for any value of l, the VC dimension for 'triangle waves' hypothesis set is infinite.

Question 12

For $N \ge d_{vc} \ge 2$, any bound greater than $N^{d_{vc}}$ would be a valid upper bound.

a. correct
$m_{\mathcal{H}}(1)\cdot m_{\mathcal{H}}(1) = 4 = m_{\mathcal{H}}(N)\ ,\ for\ N = 2$;
$m_{\mathcal{H}}(1)\cdot m_{\mathcal{H}}(2) = 2\cdot 4 = m_{\mathcal{H}}(N)\ ,\ for\ N = 3$;
$m_{\mathcal{H}}(\lfloor\frac{N}{2}\rfloor)\cdot m_{\mathcal{H}}(\lceil\frac{N}{2}\rceil) = (\frac{N}{2})^{d_{vc}}\cdot(\frac{N}{2})^{d_{vc}} = N^{d_{vc}}\cdot\frac{N^{d_{vc}}}{4^{d_{vc}}} \ge N^{d_{vc}}\ ,\ for\ N \ge 4\ and\ N\ is\ even$;
$m_{\mathcal{H}}(\lfloor\frac{N}{2}\rfloor)\cdot m_{\mathcal{H}}(\lceil\frac{N}{2}\rceil) = m_{\mathcal{H}}(\frac{N-1}{2})\cdot m_{\mathcal{H}}(\frac{N+1}{2}) = \frac{(N-1)^{d_{vc}}(N+1)^{d_{vc}}}{4^{d_{vc}}} > N^{d_{vc}}\cdot\frac{(N-1)^{d_{vc}}}{4^{d_{vc}}} \ge N^{d_{vc}}\ , for\ N \ge 5\ and\ N\ is\ odd$

b. wrong
$2^{d_{vc}} \le N^{d_{vc}}$ for $N \ge 2$

c. correct
$2^i m_{\mathcal{H}}(N-i) \sim 2^i(N-i)^{d_{vc}}$ which is monotonely increasing by i. So $\min\limits_{1\le i\le N-1} 2^i m_{\mathcal{H}}(N-i) = 2\cdot m_{\mathcal{H}}(N-1)$, that can be an upper bound. This can also be explained by taking away some points that we do not know whether provide full potential and adding back all the possiblity of these points. Thus, no matter what value of i is, $2^i m_{\mathcal{H}}(N-i)$ would be an upper bound.

d. correct
$N^{d_{vc}} + 1 > N^{d_{vc}}$

e. wrong
$m_{\mathcal{H}}(N) = m_{\mathcal{H}}(N-1) + N\cdot d_{vc} \le \frac{(N-1)(N+2)}{2}\cdot d_{vc} < N^{d_{vc}}$ for $d_{vc} \ge 3$

Question 13

Growth functions should satisfies such properties:

- $m_{\mathcal{H}}(N) \leq 2^N$ for all $N \geq 2$
- $m_{\mathcal{H}}(N)$ is monotonically increasing because when we add a new point which is independent from the previous ones we get at least one more dichtomies.

a. correct

2^N satisfies the above properties.

b.c.d. wrong

They are not monotonically increasing.

e. correct

The first derivative of **e.** is $1 + \frac{3N^2-6N+2}{6} > 0$ for all $N \geq 1$ so it is monotonically increasing. And $1 + N + \frac{N(N-1)(N-2)}{6} \leq 2^N$ for all $N \geq 1$

Question 14

The number of *hypothesis* of $\cap_{k=1}^{K} \mathcal{H}_k$ is in the range of $[0, min\{\#hypothesis\ in\ \mathcal{H}_k\}_{k=1}^{K}]$. So the tightest bound on the VC dimension of the intersection is:

$$0 \leq d_{vc}(\cap_{k=1}^{K} \mathcal{H}_k) \leq min\{d_{vc}(\mathcal{H}_k)\}_{k=1}^{K}$$

Question 15

The minimal number of *hypothesis* of the union of the sets $\cup_{k=1}^{K} \mathcal{H}_k$ is at least as many as the number of *hypothesis* of the largest set. So the lower bound of $d_{vc}(\cup_{k=1}^{K} \mathcal{H}_k)$ is $max\{d_{vc}(\mathcal{H}_k)\}_{k=1}^{K}$.

As for the upper bound of $d_{vc}(\cup_{k=1}^{K} \mathcal{H}_k)$, we can first consider the situation with two hypothesis $\mathcal{H}_1$ and $\mathcal{H}_2$ and $d_{vc}(\mathcal{H}_1) = d_1$, $d_{vc}(\mathcal{H}_2) = d_2$. The number of ways N particular points can be classified using $\mathcal{H}_1 \cap \mathcal{H}_2$ is at most the number of classifications using $\mathcal{H}_1$ plus the number of classifications using $\mathcal{H}_2$, which means:

$$\begin{aligned} m_{\mathcal{H}_1 \cap \mathcal{H}_2}(N) &\leq m_{\mathcal{H}_1}(N) + m_{\mathcal{H}_2}(N) \\ &\leq \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=0}^{d_2} \binom{N}{i} \end{aligned}$$

Cause $\binom{N}{i} = \binom{N}{N-i}$, this can be rewritten as:

$$m_{\mathcal{H}_1 \cap \mathcal{H}_2}(N) \leq \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=N-d_2}^{N} \binom{N}{i}$$

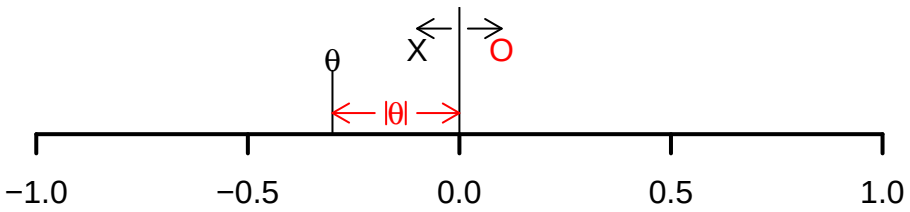
When $N \geq d_1 + d_2 + 2$,

$$m_{\mathcal{H}_1 \cap \mathcal{H}_2}(N) \leq \sum_{i=0}^{N} \binom{N}{i} - \binom{N}{d_1+1} = 2^N - \binom{N}{d_1+1} < 2^N$$

The upper bound of $d_{vc}(m_{\mathcal{H}_1 \cup \mathcal{H}_2}(N))$ is $d_1 + d_2 + 1$. We can use this to get the upper bound of $d_{vc}(\cup_{k=1}^{K} \mathcal{H}_k)$:

$$\begin{aligned} d_{vc}(\cup_{k=1}^{K} \mathcal{H}_k) &\leq d_{vc}(\mathcal{H}_1) + d_{vc}(\cup_{k=2}^{K} \mathcal{H}_k) + 1 \\ &\leq d_{vc}(\mathcal{H}_1) + d_{vc}(\mathcal{H}_2) + d_{vc}(\cup_{k=3}^{K} \mathcal{H}_k) + 2 \\ &\leq \sum_{k=1}^{K} d_{vc}(\mathcal{H}_k) + K - 1 \end{aligned}$$

Question 16



As the figure above :

- when $s = +1$, $h_{s,\theta}(x)$ output $+1$ for points right off θ and -1 otherwise, then $E_{in} = |\theta|/2$.
- when $s = -1$, $h_{s,\theta}(x)$ output -1 for points right off θ and $+1$ otherwise, then $E_{in} = 1 - |\theta|/2$.

Combine the two equations into one, we get:

$$E_{in} = 0.5 + 0.5s(|\theta| - 1)$$

Refer to Question 1, we can easily compute E_{out} using E_{in}:

$$\lambda\mu + (1-\lambda)(1-\mu)$$

Here $\mu = 0.5 + 0.5s(|\theta| - 1)$ and $\lambda = 0.8$, then:

$$E_{out}(h_{s,\theta}) = 0.5 + 0.3s(|\theta| - 1)$$

Reference

[1] Machine Learning Fall 2011: Homework 2 [http://www.cs.cmu.edu/~epxing/Class/10701-11f/HW/HW2_solution.pdf]

[2] Foundations of Machine Learning: assignment 2 [http://www.cs.nyu.edu/~mohri/ml/ml10/sol2.pdf]