
Machine Learning

Answer Sheet for Homework 2

Da-Min HUANG

R04942045

Graduate Institute of Communication Engineering, National Taiwan University

November 2, 2015

Problem 1

Since h makes an error ($y \neq f(\mathbf{x})$) with probability μ , consider

1. $h = f(\mathbf{x}) \neq y$: $(1 - \mu)(1 - \lambda)$

2. $h \neq f(\mathbf{x}) = y$: $\mu\lambda$

So the probability of error that h makes in approximating the noisy target y is

$$(1 - \mu)(1 - \lambda) + \mu\lambda \tag{1}$$

□

Problem 2

Consider

$$(1 - \mu)(1 - \lambda) + \mu\lambda = 1 - \mu - \lambda + 2\mu\lambda = (1 - \lambda) + \mu(2\lambda - 1) \tag{2}$$

If h is independent of μ , then $2\lambda - 1 = 0 \Rightarrow \lambda = 0.5$.

□

Problem 3

Let

$$4(2N)^{d_{vc}} \exp\left(-\frac{1}{8}\epsilon^2 N\right) := \delta \quad (3)$$

we have

$$\exp\left(-\frac{1}{8}\epsilon^2 N\right) = \frac{\delta}{2^{d_{vc}+2} \cdot N^{d_{vc}}} \quad (4)$$

$$-\frac{1}{8}\epsilon^2 N = \ln\left(\frac{\delta}{2^{d_{vc}+2}}\right) - d_{vc} \ln(N) \quad (5)$$

Now take $d_{vc} = 10$, $\delta = 0.95$ and $\epsilon \leq 0.05$, we have

$$\epsilon = \sqrt{-\frac{8}{N} \left[\ln\left(\frac{0.95}{2^{12}}\right) - 10 \ln(N) \right]} \leq 0.05 \Rightarrow \epsilon \geq 442810 \quad (6)$$

So the closest numerical approximation of the sample size is 443000.

□

Problem 4

Let $m_{\mathcal{H}}(N) = (N)^{d_{vc}}$, so

1. Original VC Bound:

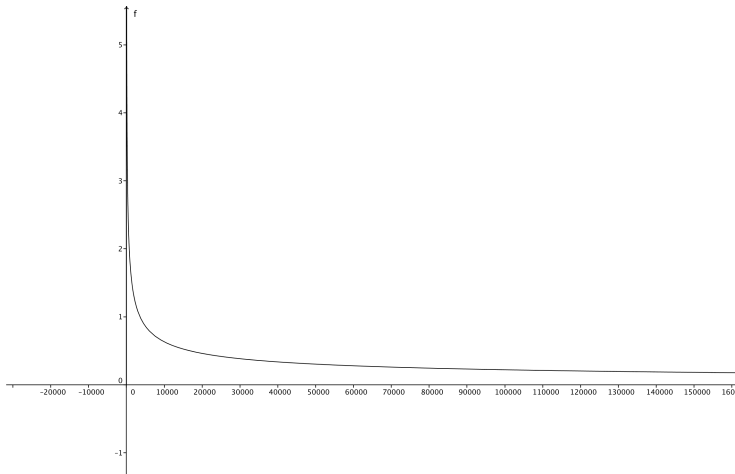


Figure 1: Original VC Bound

$$\sqrt{\frac{8}{N} \ln\left(\frac{4(2N)^{d_{vc}}}{\delta}\right)} = \sqrt{\frac{8}{10000} \ln\left(\frac{4(20000)^{50}}{0.05}\right)} \approx 0.63217 \quad (7)$$

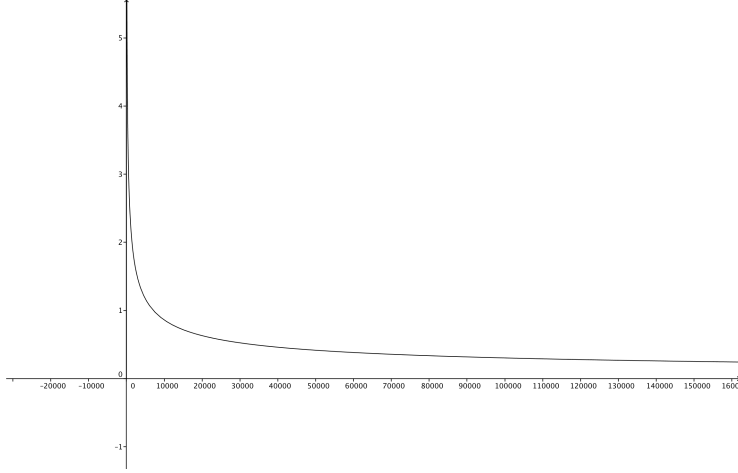


Figure 2: Variant VC bound

2. Variant VC bound:

$$\sqrt{\frac{16}{N} \ln \left(\frac{2(N)^{d_{vc}}}{\sqrt{\delta}} \right)} = \sqrt{\frac{16}{10000} \ln \left(\frac{2(10000)^{50}}{\sqrt{0.05}} \right)} \approx 0.86043 \quad (8)$$

3. Rademacher Penalty Bound:

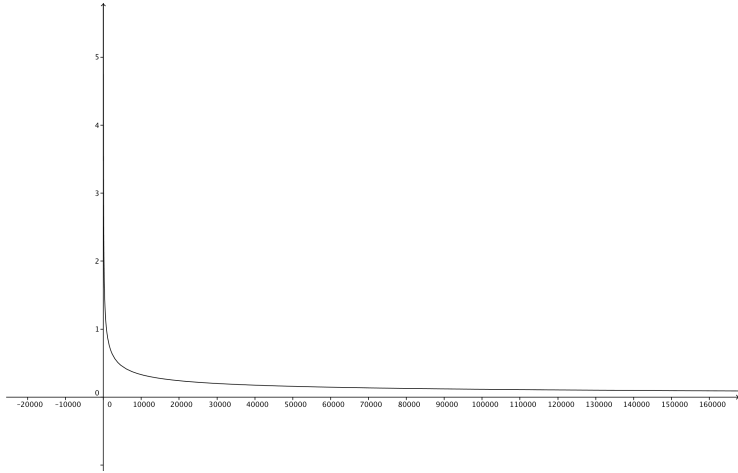


Figure 3: Rademacher Penalty Bound

$$\sqrt{\frac{2 \ln \left(2N (N)^{d_{vc}} \right)}{N}} + \sqrt{\frac{2}{N} \ln \left(\frac{1}{\delta} \right)} + \frac{1}{N} \quad (9)$$

$$= \sqrt{\frac{2 \ln \left(20000 (10000)^{50} \right)}{10000}} + \sqrt{\frac{2}{10000} \ln \left(\frac{1}{0.05} \right)} + \frac{1}{10000} \quad (10)$$

$$\approx 0.33131 \quad (11)$$

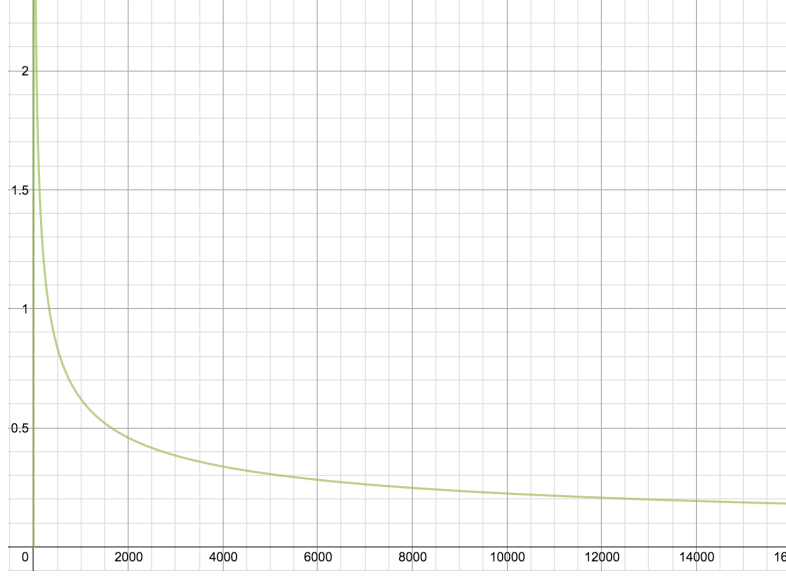


Figure 4: Parrondo and Van den Broek

4. Parrondo and Van den Broek:

$$\epsilon \leq \sqrt{\frac{1}{N} \left(2\epsilon + \ln \left(\frac{6(2N)^{d_{vc}}}{\delta} \right) \right)} = \sqrt{\frac{1}{10000} \left(2 \times \epsilon + \ln \left(\frac{6(20000)^{50}}{0.05} \right) \right)} \quad (12)$$

$$\Rightarrow \epsilon \leq 0.22370 \quad (13)$$

5. Devroye:

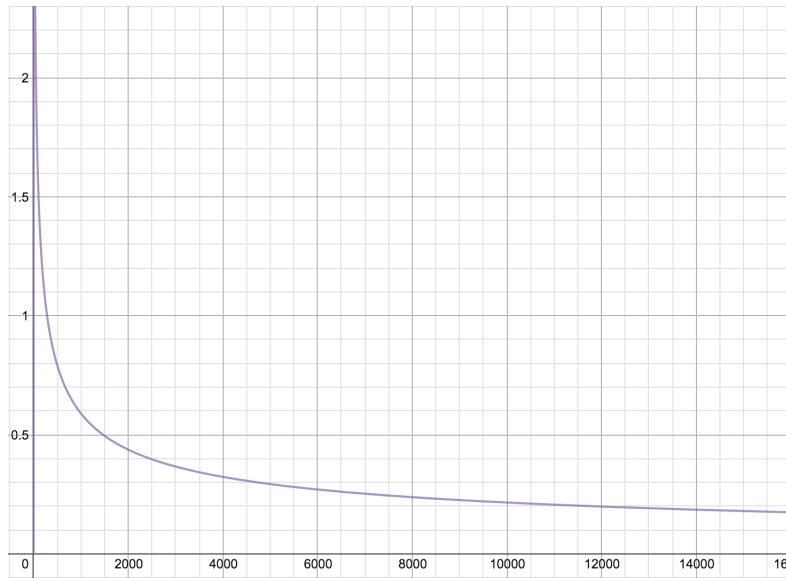


Figure 5: Devroye

$$\epsilon \leq \sqrt{\frac{1}{2N} \left(4\epsilon(1+\epsilon) + \ln \left(\frac{4(N^2)^{d_{vc}}}{\delta} \right) \right)} \quad (14)$$

$$= \sqrt{\frac{1}{20000} \left(4\epsilon(1+\epsilon) + \ln \left(\frac{4(10000^2)^{50}}{0.05} \right) \right)} \quad (15)$$

$$\Rightarrow \epsilon \leq 0.21523 \quad (16)$$

So the tightest bound is Devroye.

□

Problem 5

Let $m_{\mathcal{H}}(N) = (N)^{d_{vc}}$, so

1. Original VC Bound:

$$\sqrt{\frac{8}{N} \ln \left(\frac{4(2N)^{d_{vc}}}{\delta} \right)} = \sqrt{\frac{8}{5} \ln \left(\frac{4(10)^{50}}{0.05} \right)} \approx 13.828 \quad (17)$$

2. Variant VC bound:

$$\sqrt{\frac{16}{N} \ln \left(\frac{2(N)^{d_{vc}}}{\sqrt{\delta}} \right)} = \sqrt{\frac{16}{5} \ln \left(\frac{2(5)^{50}}{\sqrt{0.05}} \right)} \approx 16.264 \quad (18)$$

3. Rademacher Penalty Bound:

$$\sqrt{\frac{2 \ln \left(2N(N)^{d_{vc}} \right)}{N}} + \sqrt{\frac{2}{N} \ln \left(\frac{1}{\delta} \right)} + \frac{1}{N} \quad (19)$$

$$= \sqrt{\frac{2 \ln \left(10(5)^{50} \right)}{5}} + \sqrt{\frac{2}{5} \ln \left(\frac{1}{0.05} \right)} + \frac{1}{5} \quad (20)$$

$$\approx 7.0488 \quad (21)$$

4. Parrondo and Van den Broek:

$$\epsilon \leq \sqrt{\frac{1}{N} \left(2\epsilon + \ln \left(\frac{6(2N)^{d_{vc}}}{\delta} \right) \right)} = \sqrt{\frac{1}{5} \left(2\epsilon + \ln \left(\frac{6(10)^{50}}{0.05} \right) \right)} \quad (22)$$

$$\Rightarrow \epsilon \leq 5.1014 \quad (23)$$

5. Devroye:

$$\epsilon \leq \sqrt{\frac{1}{2N} \left(4\epsilon(1 + \epsilon) + \ln \left(\frac{4(N^2)^{d_{vc}}}{\delta} \right) \right)} \quad (24)$$

$$= \sqrt{\frac{1}{10} \left(4\epsilon(1 + \epsilon) + \ln \left(\frac{4(5^2)^{50}}{0.05} \right) \right)} \quad (25)$$

$$\Rightarrow \epsilon \leq 5.5931 \quad (26)$$

So the tightest bound is Parrondo and Van den Broek.

□

Problem 6

First, choose two point as the begin and end points, we have $\binom{N}{2} + 1$ choices, where $+1$ means the situation the begin and the end are the same.

Then, inside the interval should be positive or negative, there are 2 choices. Hence, we have

$$m_{\mathcal{H}}(N) = 2 \left(\binom{N}{2} + 1 \right) = N^2 - N + 2 \quad (27)$$

□

Problem 7

For $N = 4$, we have

$$4^2 - 4 + 2 = 14 < 16 = 2^4 \quad (28)$$

so the VC dimension is $4 - 1 = 3$.

□

Problem 8

It is like choose two radius in $(0, N]$, so we have $\binom{N+1}{2}$ choices in polar coordinate.

But we need to add the case when a and b make all hypothesis -1 , which means $b^2 - a^2 < r_i^2$, $1 \leq i \leq N$, where r_i is the distance from origin to hypothesis point.

Hence,

$$m_{\mathcal{H}}(N) = \binom{N+1}{2} + 1 \quad (29)$$

□

Problem 9

For $\sum_{i=0}^D c_i x^i$, we have D different roots at most, separated \mathbb{R} (x -axis) into $D+1$ sections; or no root in \mathbb{R} , which makes h_c to be $+1$ or -1 , $\forall x$.

The number of roots (from 0 to D) forms different result in $\{+1, -1\}^{D+1}$, like all $+1$ or $[+1, -1, +1, \dots], \dots$, which has at most 2^{D+1} choices. So the VC dimension is $D+1$.

□

Problem 10

The VC dimension of simplified decision tree is equal to the number of hyper-rectangular regions, where each region returns the same \mathbf{v} .

d -dimension space has 2^d independent hyper-rectangular regions separated by d lines since each dimension in \mathbb{R}^d has two choices 0 or 1. This can shatter at most 2^d vectors. So the VC dimension is 2^d .

For example, we can have 2^d points $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{2^d}$ are different combination of $\{-1, +1\}^d$. Choose $\mathbf{t} = \{0, 0, \dots, 0\}$, we can put each \mathbf{x}_i in different hyper-rectangular. If $\mathbf{S}_i \in \mathbf{S}$, then $h_{\mathbf{t}, \mathbf{S}} = 1$, else -1 . Hence the 2^d points can be shattered by choosing different collection \mathbf{S} .

If there are $2^d + 1$ points, then there are at least two points in the same hyper-rectangular region. These two points have same label no matter what \mathbf{S} is, so the hypothesis cannot shatter $2^d + 1$ points.

□

Problem 11

If there are N point on \mathbb{R} , from 1st point to N^{th} point, put the i^{th} point on 4^i . For $1 \leq k \leq 2^N$, using

$$1 + \frac{1}{2} \left(\frac{2k-2}{2^{N+1}} \right) < \alpha_k < 1 + \frac{1}{2} \left(\frac{2k-1}{2^{N+1}} \right) \quad (30)$$

those α_k can build all $\{+1, -1\}^N$ combinations, so this triangle wave hypothesis can shatter any N . Hence the VC dimension is ∞ .

□

Problem 12

Suppose $\min_{1 \leq i \leq N-1} 2^i m_{\mathcal{H}}(N-i)$ is not an upper bound.

Since $m_{\mathcal{H}}(N) > m_{\mathcal{H}}(m)$, $\forall m < N$. So we must have $m_{\mathcal{H}}(N) > \min_{1 \leq i \leq N-1} 2^i m_{\mathcal{H}}(N-i)$.

Let $i = k$ satisfies the minimum condition. Consider following cases.

1. $N - k < d_{vc}$.

Then we have

$$\min_{1 \leq i \leq N-1} 2^i m_{\mathcal{H}}(N-i) = 2^k \times 2^{N-k} = 2^N > m_{\mathcal{H}}(N) \quad (31)$$

which is a contradiction.

2. $N - k \geq d_{vc}$.

Then we have

$$\min_{1 \leq i \leq N-1} 2^i m_{\mathcal{H}}(N-i) = 2^k m_{\mathcal{H}}(N-k) < m_{\mathcal{H}}(N) < 2^N \quad (32)$$

This is impossible since $m_{\mathcal{H}}(d_{vc}) = 2^{d_{vc}} \Rightarrow 2m_{\mathcal{H}}(d_{vc}) = 2^{d_{vc}+1} > m_{\mathcal{H}}(d_{vc} + 1)$. So $2^k m_{\mathcal{H}}(N-k) > m_{\mathcal{H}}(N)$, which leads to a contradiction.

Thus, $\min_{1 \leq i \leq N-1} 2^i m_{\mathcal{H}}(N-i)$ is an upper bound.

□

Problem 13

$m_{\mathcal{H}}(N) = 2^{\lfloor \sqrt{N} \rfloor}$ cannot be a growth function.

If there is no break point, the growth function should be 2^N ; if there is break point k , then the growth function is bounded by $\sum_{i=0}^{k-1} \binom{N}{i}$ if $k \geq 2$.

The break point is 2 since $2^{\lfloor \sqrt{2} \rfloor} = 2^1 < 2^2$. Consider $N = 25$, we have

$$2^{\lfloor \sqrt{25} \rfloor} = 2^5 = 32 > \binom{25}{0} + \binom{25}{1} = 26 \quad (33)$$

Hence this is not a growth function.

□

Problem 14

The smallest case of $\bigcap_{k=1}^K \mathcal{H}_k = \{0\}$, so $d_{vc}(\{0\}) = 0$.

The biggest intersection is the smallest set of \mathcal{H}_i , $1 \leq i \leq k$, which is

$$\{0\} \subseteq \bigcap_{k=1}^K \mathcal{H}_k \subseteq \min_{1 \leq k \leq K} \{\mathcal{H}_k\} \quad (34)$$

Suppose $A \subseteq B$, then we have $d_{vc}(A) \leq d_{vc}(B)$ since if hypothesis set is greater than or equal, the VC dimension cannot be smaller.

So the upper bound of $d_{vc}\left(\bigcap_{k=1}^K \mathcal{H}_k\right)$ is $\min_{1 \leq k \leq K} \{d_{vc}(\mathcal{H}_k)\}$. Hence

$$0 \leq d_{vc}\left(\bigcap_{k=1}^K \mathcal{H}_k\right) \leq \min_{1 \leq k \leq K} \{d_{vc}(\mathcal{H}_k)\} \quad (35)$$

□

Problem 15

The smallest union is the biggest set of \mathcal{H}_i , $1 \leq i \leq k$. So the lower bound of $d_{vc}\left(\bigcup_{k=1}^K \mathcal{H}_k\right)$ is $\max_{1 \leq k \leq K} \{d_{vc}(\mathcal{H}_k)\}$.

Claim: $d_{vc}(\mathcal{H}_1 \cup \mathcal{H}_2) \leq d_{vc}(\mathcal{H}_1) + d_{vc}(\mathcal{H}_2) + 1$.

Proof of claim:

Define $d_{vc}(\mathcal{H}_1) = d_1$, $d_{vc}(\mathcal{H}_2) = d_2$. The number can be classified using $\mathcal{H}_1 \cup \mathcal{H}_2$ is at most the number of classifications using \mathcal{H}_1 plus the number of classifications using \mathcal{H}_2 . So

$$m_{\mathcal{H}_1 \cup \mathcal{H}_2}(N) \leq m_{\mathcal{H}_1}(N) + m_{\mathcal{H}_2}(N) \quad (36)$$

Since $B(N, K) \leq \sum_{i=0}^d \binom{N}{i}$ ($d = K - 1$), we have

$$m_{\mathcal{H}_1 \cup \mathcal{H}_2}(N) \leq \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=0}^{d_2} \binom{N}{i} \quad (37)$$

Then

$$m_{\mathcal{H}_1 \cup \mathcal{H}_2}(N) \leq \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=0}^{d_2} \binom{N}{N-i} = \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=N-d_2}^N \binom{N}{i} \quad (38)$$

Now, if $N - d_2 > d_1 + 1$, that is $N \geq d_1 + d_2 + 2$.

$$m_{\mathcal{H}_1 \cup \mathcal{H}_2}(N) \leq \sum_{i=0}^N \binom{N}{i} - \binom{N}{d_1+1} = 2^N - \binom{N}{d_1+1} < 2^N \quad (39)$$

$$\Rightarrow m_{\mathcal{H}_1 \cup \mathcal{H}_2}(N) < 2^{d_1+d_2+2} \quad (40)$$

$$\Rightarrow m_{\mathcal{H}_1 \cup \mathcal{H}_2}(N) \leq 2^{d_1+d_2+1} \quad (41)$$

If $N - d_2 \leq d_1 + 1$, then we have $N \leq d_1 + d_2 + 1$ and

$$\underbrace{m_{\mathcal{H}_1 \cup \mathcal{H}_2}(N) \leq 2^N}_{\text{holds for all } m_{\mathcal{H}}(N)} \leq 2^{d_1 + d_2 + 1} \quad (42)$$

Hence, $d_{vc}(\mathcal{H}_1 \cup \mathcal{H}_2) \leq d_1 + d_2 + 1$.

From the equation above, we have

$$\max_{1 \leq k \leq K} \{d_{vc}(\mathcal{H}_k)\} \leq d_{vc}\left(\bigcup_{k=1}^K \mathcal{H}_k\right) \leq (K-1) + \sum_{i=1}^K d_{vc}(\mathcal{H}_i) \quad (43)$$

□

Problem 16

If $s = +1$, consider following cases.

1. $h_{s,\theta}(x) = -1$ as $\text{sign}(x) = +1 \Rightarrow 0 < x \leq \theta \leq +1$.
2. $h_{s,\theta}(x) = +1$ as $\text{sign}(x) = -1 \Rightarrow -1 \leq \theta \leq x \leq 0$.

So we have $\mathbb{P}(h_{+1,\theta}(x) \neq \text{sign}(x)) = |\theta|/2$, then E_{out} is

$$E_{out} = \underbrace{0.2 \times \left(1 - \frac{|\theta|}{2}\right)}_{\text{flipped}} + \underbrace{0.8 \times \frac{|\theta|}{2}}_{\text{no flipped}} = 0.2 + 0.3|\theta| \quad (44)$$

Similarly, if $s = -1$, then E_{out} is

$$E_{out} = \underbrace{0.2 \times \frac{|\theta|}{2}}_{\text{flipped}} + \underbrace{0.8 \times \left(1 - \frac{|\theta|}{2}\right)}_{\text{no flipped}} = 0.8 - 0.3|\theta| \quad (45)$$

Combine two cases, we have

$$E_{out} = 0.5 + 0.3s(|\theta| - 1) \quad (46)$$

□

Problem 17

The average $E_{in} = 0.17123$.

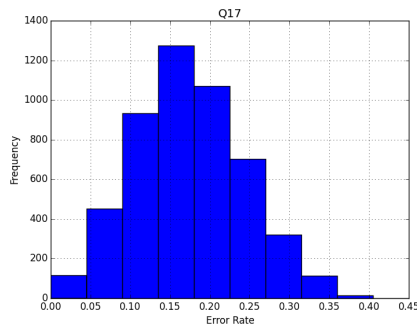


Figure 6: Q17 histogram

□

Problem 18

The average $E_{out} = 0.26586$.

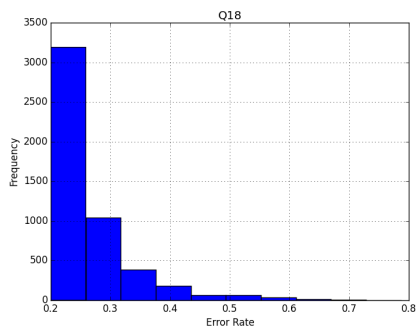


Figure 7: Q18 histogram

□

Problem 19

The optimal decision stump is the fourth dimension. The $E_{in} = 0.25000$. The optimal decision stump is the 4th column.

□

Problem 20

The $E_{out} = E_{test} = 0.36000$.

□

Problem 21

Consider any $\mathcal{H} = \{-1, +1\}^N$, with $N \geq k \geq 1$. If there are at most $k - 1$ variables be -1 . We have $m_{\mathcal{H}}(N) = \sum_{i=0}^{k-1} \binom{N}{i}$ dichotomies with no subset of k variables can be shattered.

Since $B(N, k)$ bounds $m_{\mathcal{H}}(N)$, we have

$$B(N, k) \geq \sum_{i=0}^{k-1} \binom{N}{i} \quad (47)$$

Combine the conclusion $B(N, k) \leq \sum_{i=0}^{k-1} \binom{N}{i}$, which had been proved in class, we have

$$B(N, k) = \sum_{i=0}^{k-1} \binom{N}{i} \quad (48)$$

□

Reference

- [1] Lecture Notes by Hsuan-Tien LIN, Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan.