Machine Learning

Answer Sheet for Homework 7

Da-Min HUANG
R04942045
Graduate Institute of Communication Engineering, National Taiwan University

January 6, 2016

Problem 1

Set $\mu_- = 1 - \mu_+$, we have

$$
\begin{aligned}
1 - \mu_+^2 - \mu_-^2 &= 1 - \mu_+^2 - (1 - \mu_+)^2 = (1 - \mu_+)(1 + \mu_+) - (1 - \mu_+)^2 &(1)\\
&= 2\mu_+(1 - \mu_+) = -2\mu_+^2 + 2\mu_+ = -2\left(\mu_+ - \frac{1}{2}\right)^2 + \frac{1}{2} &(2)\\
&\leq \frac{1}{2} &(3)
\end{aligned}
$$

Hence, if $\mu_+ = 1/2 \in [0, 1]$, then the maximum value of Gini index is $1/2$.

□

Problem 2

The normalized Gini index is

$$
\frac{\left(1 - \mu_+^2 - \mu_-^2\right)}{\left(\frac{1}{2}\right)} = 2\left(1 - \mu_+^2 - \mu_-^2\right) \tag{4}
$$

The squared error can be rewritten as

$$
\begin{aligned}
\mu_+(1 - (\mu_+ - \mu_-))^2 + \mu_-(-1 - (\mu_+ - \mu_-))^2 &= 4\mu_+(1 - \mu_+)^2 + 4\mu_+^2(1 - \mu_+) &(5)\\
&= 4\mu_+(1 - \mu_+) \leq 4 \times \frac{1}{4} = 1 &(6)
\end{aligned}
$$

Hence the normalized squared error is

$$4\mu_+ (1 - \mu_+) = 2 (2\mu_+ (1 - \mu_+)) = 2 \left((1 - \mu_+) (1 + \mu_+) - (1 - \mu_+)^2\right) \tag{7}$$
$$= 2 \left(1 - \mu_+^2 - \mu_-^2\right) \tag{8}$$

which is equal to normalized Gini index.

□

Problem 3

The probability of one example not sampled is

$$\left(1 - \frac{1}{N}\right)^{pN} = \frac{1}{\left(\frac{N}{N-1}\right)^{pN}} = \frac{1}{\left(1 + \frac{1}{N-1}\right)^{pN}} = \left(\frac{1}{\left(1 + \frac{1}{N-1}\right)^{N}}\right)^{p} \tag{9}$$

As $N \to \infty$, we have

$$\lim_{N\to\infty} \left(\frac{1}{\left(1 + \frac{1}{N-1}\right)^{N}}\right)^{p} = \left(\lim_{N\to\infty} \frac{1}{\left(1 + \frac{1}{N-1}\right)^{N}}\right)^{p} = \left(\frac{1}{e}\right)^{p} = e^{-p} \tag{10}$$

So there approximately $e^{-p} \cdot N$ of the examples not sampled.

□

Problem 4

Since $G = \text{Uniform}\left(\{g_k\}_{k=1}^3\right)$, so if at least two terms of $\{g_k\}_{k=1}^3$ ouput wrong result, then G outputs wrong result. Let $\{E_k\}_{k=1}^3$ be the set of examples that $\{g_k\}_{k=1}^K$ got wrong results. Apparently $|E_3| > |E_2| > |E_1|$ and $|E_1| + |E_2| > |E_3|$. So

1. Maximum of $E_{\text{out}}(G)$ happens at $E_3 \subset (E_1 \cup E_2)$. Then G outputs wrong result in the region of E_3 with $E_{\text{out}}(G) = 0.35$.

2. Minimum of $E_{\text{out}}(G)$ happens at $E_i \cap E_j = \phi$, $i \neq j$ and $1 \leq i, j \leq 3$ with $i, j \in \mathbb{N}$. Then G always ouputs the correct result since $(E_1 \cup E_2 \cup E_3) \subset \{\text{all examples}\}$.

Hence, $0 \leq E_{\text{out}}(G) \leq 0.35$.

□

Problem 5

Since $G = \text{Uniform}\left(\{g_k\}_{k=1}^{K}\right)$, so if at least $(K+1)/2$ terms of $\{g_k\}_{k=1}^{K}$ ouput wrong result, then G outputs wrong result. Let $\{E_k\}_{k=1}^{K}$ be the set of examples that $\{g_k\}_{k=1}^{K}$ got wrong results.

If G outputs wrong result on some example $\mathbf{x}$, then we have

$$\mathbf{x} \in \bigcap_{i=1}^{((K+1)/2)+m} E_{\alpha_i} \tag{11}$$

where α_i is some index satisfies $1 \leq \alpha_i \leq K$ and $m \in (\mathbb{N} \cup \{0\})$ with $0 \leq m < (K+1)/2$.
And

$$\left| \bigcap_{i=1}^{((K+1)/2)+m} E_{\alpha_i} \right| \leq \frac{2}{K+1+2m} \sum_{k=1}^{K} e_k \leq \frac{2}{K+1} \sum_{k=1}^{K} e_k \tag{12}$$

(12) holds due to

$$\bigcap_{i=1}^{((K+1)/2)+m} E_{\alpha_i} \subseteq E_\beta \tag{13}$$

where β is some index such that $|E_\beta| = \min_{\alpha_i} |E_{\alpha_i}|$. So

$$\left(\frac{K+1}{2} + m\right) \left| \bigcap_{i=1}^{((K+1)/2)+m} E_{\alpha_i} \right| \leq \left(\frac{K+1}{2} + m\right) |E_\beta| \leq \sum_{k=1}^{K} |E_k| \tag{14}$$

(14) holds since size of E_β is the samllest among $(((K+1)/2)+m)$ terms and $\sum_{k=1}^{K} |E_k|$ must contains the $(((K+1)/2)+m)$ terms.
Hence, we have

$$E_{\text{out}}(G) \leq \frac{2}{K+1} \sum_{k=1}^{K} e_k \tag{15}$$

$\square$

Problem 6

$\square$

Problem 7

$\square$

Problem 8

$\square$

Problem 9

$\square$

Problem 10

$\square$

Problem 11

$\square$

Problem 12

$\square$

Problem 13

$\square$

Problem 14

$\square$

Problem 15

$\square$

Problem 16

□

Problem 17

□

Problem 18

□

Problem 19

□

Problem 20

□

Reference

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