Machine Learning

Answer Sheet for Homework 1

Da-Min HUANG

R04942045

Graduate Institute of Communication Engineering, National Taiwan University

October 14, 2015

Problem 1

(i) Prime number has its math and programmable definition.

(ii)
- Pattern: how the credit card charged.
- Definition: not easily programmable.
- Data: history of bank operation.

(iii) It has programmable definition.

(iv)
- Pattern: cycle for traffic lights.
- Definition: not easily programmable.
- Data: history of traffic condition.

(v)
- Pattern: age of people.
- Definition: not enough data.
- Data: medical record.

Hence, the answer is (ii), (iv) and (v).

□

Problem 2

It learns with implicit information sequentially so it is type of reinforcement learning.

$\square$

Problem 3

It learns without labels so it is type of unsupervised learning.

$\square$

Problem 4

Every picture has its label (face or non-face) so it is type of supervised learning.

$\square$

Problem 5

It schedule experiments strategically so it is type of active learning.

$\square$

Problem 6

Now we have

$$E_{OTS}\left(g,\ f\right) = \frac{1}{L}\sum_{\ell=1}^{L}\left[g\left(\mathbf{x_{N+\ell}}\right) \neq f\left(\mathbf{x}_{N+\ell}\right)\right] \tag{1}$$

It is easily to find that $g\left(\mathbf{x_{N+\ell}}\right) \neq f\left(\mathbf{x}_{N+\ell}\right)$ when $N+\ell$ is even. So

$$\sum_{\ell=1}^{L}\left[g\left(\mathbf{x_{N+\ell}}\right) \neq f\left(\mathbf{x}_{N+\ell}\right)\right] = \left\lfloor\frac{N+L}{2}\right\rfloor - \left\lfloor\frac{L}{2}\right\rfloor \tag{2}$$

since there are $\lfloor z/2 \rfloor$ even numbers between 1 and $z \in \mathbb{Z}$. Hence,

$$E_{OTS}\left(g,\ f\right) = \frac{1}{L}\left(\left\lfloor\frac{N+L}{2}\right\rfloor - \left\lfloor\frac{L}{2}\right\rfloor\right) \tag{3}$$

$\square$

Problem 7

Since f generate $\mathcal{D}$ so the output of f is fixed for $1 \le n \le N$.
There are still L terms need to be determined, each with two choices (-1 or $+1$). So the answer is 2^L.

$\square$

Problem 8

Since $\mathcal{A}_1$ and $\mathcal{A}_2$ generate $\mathcal{D}$ in a noiseless setting, so

$$\{E_{OTS}(\mathcal{A}_1,\ f)\}_{n=1}^{N} = \{E_{OTS}(\mathcal{A}_2,\ f)\}_{n=1}^{N} = \{0\} \tag{4}$$

But for $N < n \le N+L$,

$$\mathbb{P}_f\left\{\{E_{OTS}(\mathcal{A}_1,\ f)\}_{n=N+1}^{N+L} = \{E_{OTS}(\mathcal{A}_2,\ f)\}_{n=N+1}^{N+L}\right\} = \frac{1}{2^L} \tag{5}$$

since $\mathcal{A}_1$ and $\mathcal{A}_2$ have 2^L choices. But

$$\begin{aligned}
\mathbb{E}_f\{E_{OTS}(\mathcal{A}_1,\ f)\} &= \mathbb{E}_f\left(\frac{1}{L}\sum_{\ell=1}^{L}[\mathcal{A}_1(\mathbf{x_{N+\ell}}) \neq f(\mathbf{x}_{N+\ell})]\right) &(6)\\
&= \frac{1}{L}\mathbb{E}_f\left(\sum_{\ell=1}^{L}[\mathcal{A}_1(\mathbf{x_{N+\ell}}) \neq f(\mathbf{x}_{N+\ell})]\right) &(7)\\
&= \frac{1}{L}\left(\underbrace{\frac{L}{2}}_{\text{Expected error number}}\right) = \frac{1}{2} &(8)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_f\{E_{OTS}(\mathcal{A}_2,\ f)\} &= \mathbb{E}_f\left(\frac{1}{L}\sum_{\ell=1}^{L}[\mathcal{A}_2(\mathbf{x_{N+\ell}}) \neq f(\mathbf{x}_{N+\ell})]\right) &(9)\\
&= \frac{1}{L}\mathbb{E}_f\left(\sum_{\ell=1}^{L}[\mathcal{A}_2(\mathbf{x_{N+\ell}}) \neq f(\mathbf{x}_{N+\ell})]\right) &(10)\\
&= \frac{1}{L}\left(\frac{L}{2}\right) = \frac{1}{2} &(11)
\end{aligned}$$

because there are only 2 output choices, so the expectation value of error rate should be $1/2$ in a noiseless setting.

$\square$

Problem 9

If $\nu = 0.5$, then there are 5 orange marbles.

$$\mathbb{P}\left(5 \text{ orange marbles}\right) = \underbrace{(0.5)^5}_{\text{5 orange}} \times \underbrace{(0.5)^5}_{\text{5 green}} \times \binom{10}{5} = \frac{63}{256} \approx 0.2461 \tag{12}$$

□

Problem 10

If $\nu = 0.9$, then there are 9 orange marbles.

$$\mathbb{P}\left(9 \text{ orange marbles}\right) = \underbrace{(0.9)^9}_{\text{9 orange}} \times \underbrace{(0.1)^1}_{\text{1 green}} \times \binom{10}{9} = \frac{3^{18}}{2^9 \times 5^9} \approx 0.3874 \tag{13}$$

□

Problem 11

If $\nu \leq 0.1$, then there are 1 orange marbles or 0 orange marbles,

$$\mathbb{P}\left(1 \text{ orange marbles}\right) = \underbrace{(0.9)^1}_{\text{1 orange}} \times \underbrace{(0.1)^9}_{\text{9 green}} \times \binom{10}{1} = 9.0 \times 10^{-9} \tag{14}$$

$$\mathbb{P}\left(0 \text{ orange marbles}\right) = \underbrace{(0.1)^{10}}_{\text{10 green}} = 0.1 \times 10^{-9} \tag{15}$$

$$\Rightarrow \mathbb{P}\left(\nu \leq 0.1\right) = 9.0 \times 10^{-9} + 0.1 \times 10^{-9} = 9.1 \times 10^{-9} \tag{16}$$

□

Problem 12

By Hoeffding's Inequality: $\mathbb{P}\left[|\nu - \mu| > \epsilon\right] \leq 2\exp\left(-2\epsilon^2 N\right)$, we have

$$\text{Bound} = 2\exp\left(-2 \times (0.9 - 0.1)^2 \times 10\right) = 5.52 \times 10^{-6} \tag{17}$$

□

Problem 13

To get all orange 1, we can only pick B or C kind. Since each kind is with same quantity, then we have

$$\mathbb{P}\left(\text{pick B or C}\right) = \frac{1}{2} \tag{18}$$

so

$$\mathbb{P}\left(\text{all orange 1}\right) = \frac{1}{2}^5 = \frac{1}{32} = \frac{8}{256} \tag{19}$$

$\square$

Problem 14

Consider the situations:

1. Only one number purely orange.

   The only possible number are 2 and 5. So

$$\mathbb{P}\left(\text{only 2}\right) = \frac{1}{4^5}\left(\underbrace{\binom{5}{1}}_{\text{1A4C}} + \underbrace{\binom{5}{2}}_{\text{2A3C}} + \underbrace{\binom{5}{3}}_{\text{3A2C}} + \underbrace{\binom{5}{4}}_{\text{4A1C}}\right) = \frac{30}{1024} = \mathbb{P}\left(\text{only 5}\right) \tag{20}$$

2. Two numbers purely orange.

   The possible numbers pair are $(1,\ 3)$ and $(4,\ 6)$. So

$$\mathbb{P}\left((1,\ 3)\right) = \frac{1}{4^5}\left(\underbrace{\binom{5}{1}}_{\text{1B4C}} + \underbrace{\binom{5}{2}}_{\text{2B3C}} + \underbrace{\binom{5}{3}}_{\text{3B2C}} + \underbrace{\binom{5}{4}}_{\text{4B1C}}\right) = \frac{30}{1024} = \mathbb{P}\left((4,\ 6)\right) \tag{21}$$

3. Three numbers purely orange.

   The possible numbers pair are $(1,\ 2,\ 3)$, $(4,\ 5,\ 6)$, $(1,\ 3,\ 5)$ and $(2,\ 4,\ 6)$. So

$$\mathbb{P}\left((1,\ 2,\ 3)\right) = \frac{1}{4^5} = \frac{1}{1024} = \mathbb{P}\left((4,\ 5,\ 6)\right) = \mathbb{P}\left((1,\ 3,\ 5)\right) = \mathbb{P}\left((2,\ 4,\ 6)\right) \tag{22}$$

So

$$\mathbb{P}\left(\text{some number purely orange}\right) = 2 \times \frac{30}{1024} + 2 \times \frac{30}{1024} + 4 \times \frac{1}{1024} = \frac{31}{256} \tag{23}$$

$\square$

Problem 15

The number of updates before the algorithm halts is 45 times update, the index of the example that results in the last mistake is 135.

□

Problem 16

The average number of updates before the algorithm halts is 40.477. And the histogram is

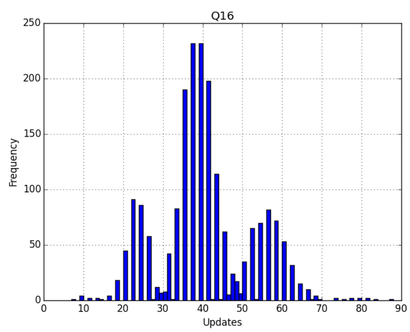

Figure 1: Q16 histogram

□

Problem 17

The average number of updates before the algorithm halts 40.219. Compare with the

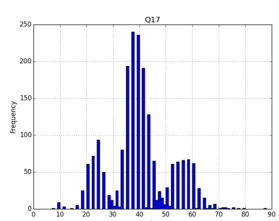

Figure 2: Q17 histogram

previous problem, we can see that they are similar. But the peak of $\eta = 0.5$ moves a little left. Test for $\eta = 0.1$ and $\eta = 0.01$, we have

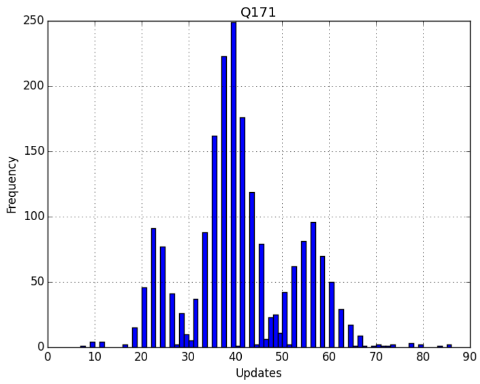

Figure 3: Q17 with $\eta = 0.1$

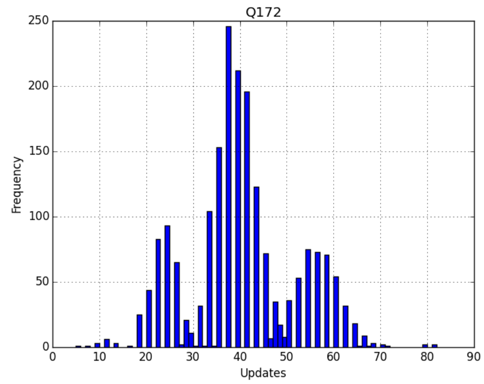

Figure 4: Q17 with $\eta = 0.01$

Seems like have no trend of moving left and the average is still around 40 (39.97 and 39.982, respectively). So the value of η affects the number of updates little.
In fact, this because initial value of $\mathbf{w}$ is $\mathbf{0}$. So even the update term $\eta y_{n(t)}\mathbf{x}_{n(t)}$ is small, we still have

$$\frac{\|\mathbf{w}_{\eta=0.5}\|}{\|\mathbf{w}\|} = \eta, \ \frac{\mathbf{w}_{\eta=0.5}}{\|\mathbf{w}_{\eta=0.5}\|} \cdot \frac{\mathbf{w}}{\|\mathbf{w}\|} = 1 \tag{24}$$

This implies η only affects the absolute value of $\mathbf{w}_\eta$. So the number of update will not change.

$\square$

Problem 18

The average error rate on the test set is 0.130997.

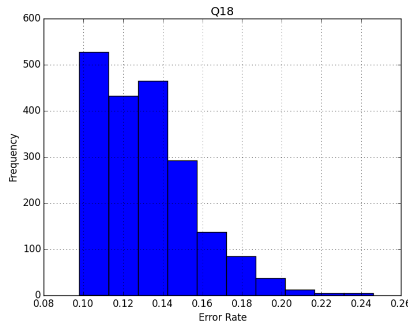

Figure 5: Q18 histogram

□

Problem 19

The average error rate on the test set is 0.364533. Compare with previous problem, we

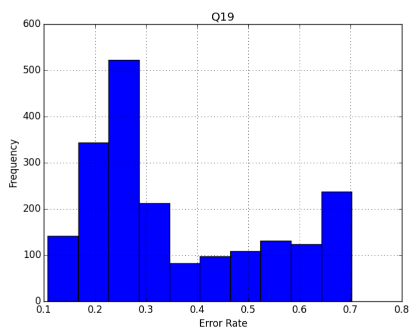

Figure 6: Q19 histogram

see that the error rate is not monotonic decreasing, with two local maximum around 0.3 and 0.7. Distribution of error rate is irregular.

□

Problem 20

The average error rate on the test set is 0.11408.

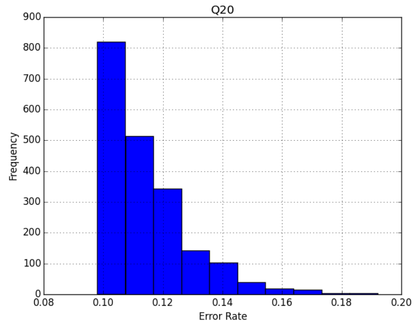

Figure 7: Q20 histogram

Compare with Problem 18, the figure is still monotonic decreasing. But the number between error rate= 0.10 and 0.11 increases, about $1.1 \sim 1.5$ times greater than Q18's. So the increase number of updates lower the average of error rate.

□

Problem 21

Use python to calculate the time factor,

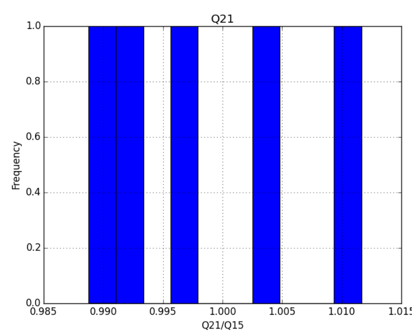

Figure 8: Q21 histogram

where

$$\text{Q21/Q15} = \frac{\text{Time of PLA as all } \mathbf{x}_n \text{(train set of Q15) scale down by a factor of 20}}{\text{Time of normal PLA}} \tag{25}$$

The histogram record Q21/Q15 in repeated 20 times. We can find that two methods costs almost same time. Since we scale down all $\mathbf{x}_n$, so during every update of $\mathbf{w}_t$,

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_{n(t)} \left(\mathbf{x}_{n(t)}/20\right) = \mathbf{w}_t + \frac{1}{20} y_{n(t)} \mathbf{x}_{n(t)} \tag{26}$$

From the conclusion of Problem 17, we know that the factor acts just like η, so it does not make PLA algorithm run faster if the initial value of $\mathbf{w}$ is $\mathbf{0}$.

Also, if the initial value of $\mathbf{w}$ is not $\mathbf{0}$, it should cost more time to update the angle of $\mathbf{w}$ to final result since the update term is smaller.

□

Reference

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