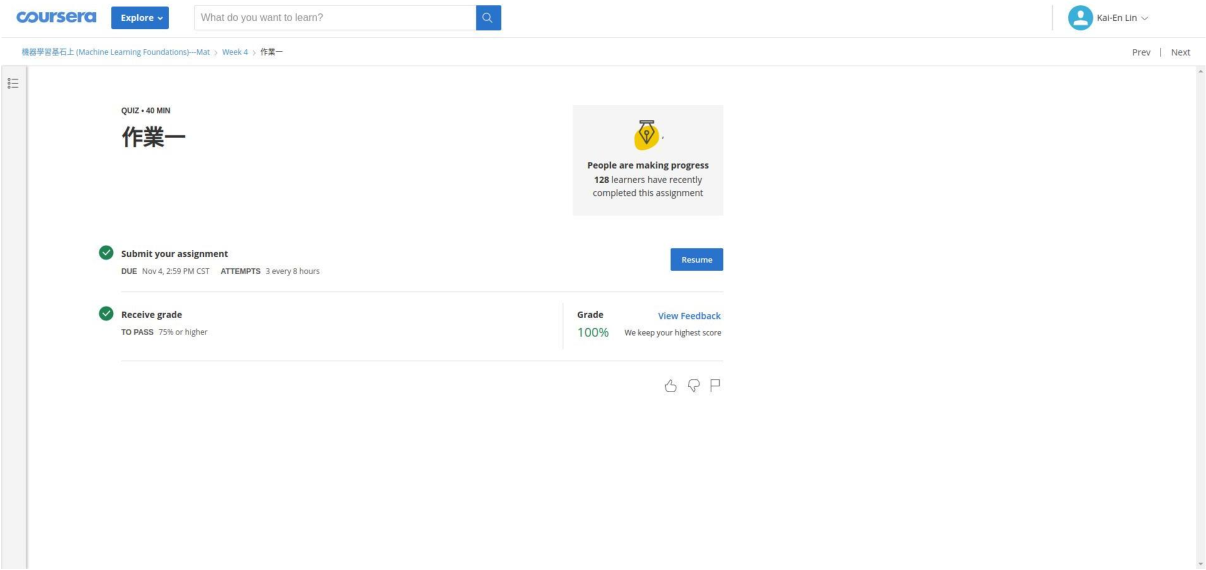
Homework #1

Student Name: 林楷恩
Student ID: b07902075

1


coursera
Explore
What do you want to learn?
Kai-En Lin
機器學習基石上 (Machine Learning Foundations)---Mat > Week 4 > 作業一
Prev | Next
Resume
View Feedback

QUIZ • 40 MIN

作業一

People are making progress
128 learners have recently completed this assignment

Submit your assignment
DUE Nov 4, 2:59 PM CST ATTEMPTS 3 every 8 hours

Receive grade
TO PASS 75% or higher

Grade
100%
We keep your highest score

2

Semi-supervised learning can be used on verifying whether an article on Internet is fake or not. Because the Internet is full of complicated information which is hard to classify even for most of educated people. Typically, to correctly label this kind of data requires a sophisticated human agent to spend a lot of time collecting information from different sources and carefully judge which one is more reliable. Thus, the cost of labeling will be quite expensive for supervised learning, and semi-supervised learning can help.

3

Let the set of all different f be $F = \{f_1, f_2, \ldots, f_m\}$, $m = 2^L$, because for each $x_{N+\ell}$ where $1 \le \ell \le L$, $f(x_{N+\ell})$ has two possible outputs (1 & -1). And since the set enumerates all possible output sequence of $f(x_{N+1}), f(x_{N+2}), \ldots, f(x_{N+L})$, for any $\mathcal{A}$, there exists a f_i such that $E_{OTS}(\mathcal{A}(\mathcal{D}), f_k) = E_{OTS}(f_i, f_k)$, $\forall\ 1 \le k \le m$, which means the f_i that can be seem as equal to $\mathcal{A}(\mathcal{D})$ considering only the outputs of the test inputs. Hence, the problem become:

$$\textit{prove} \sum_{k=1}^{2^L} P(f_k) \times E_{OTS}(f_i, f_k) = \text{constant},\ \forall\ 1 \le i \le 2^L$$

Because all the f are equally likely in probability, $P(f_k) = \dfrac{1}{2^L}$, $\forall\ 1 \le k \le 2^L$

$$\Longrightarrow \sum_{k=1}^{2^L} P(f_k) \times E_{OTS}(f_i, f_k) = \frac{1}{2^L} \sum_{k=1}^{2^L} E_{OTS}(f_i, f_k)$$

Then I count how many f_k *s.t.* $E_{OTS}(f_i, f_k) = t$ for $0 \le t \le L$ separately and add them together at the end.

$E_{OTS}(f_i, f_k) = t$ means there are t of $x_{N+\ell}$ *s.t.* $f_i(x_{N+\ell}) \ne f_k(x_\ell)$ where $1 \le \ell \le L$. So the number of f_k is the number of way to choose t positions from 1 to L in $f_i(x_{N+\ell})$ to invert, which is C_t^L. And the OTS they contribute is $\frac{1}{L} t\, C_t^L$. Thus,

$$\begin{aligned}
\sum_{k=1}^{2^L} E_{OTS}(f_i, f_k) &= \sum_{t=1}^{L} \frac{1}{L} t\, C_t^L \\
&= \frac{1}{L} \sum_{t=1}^{L} t \frac{L!}{t!(L-t)!} = \frac{1}{L} \sum_{t=1}^{L} \frac{L!}{(t-1)!(L-t)!} \\
&= \frac{1}{L} \sum_{t=1}^{L} L \frac{(L-1)!}{(t-1)!(L-t)!} = \sum_{t=1}^{L} C_{t-1}^{L-1} \\
&= \sum_{t=0}^{L-1} C_t^{L-1} \\
&= 2^{L-1} \text{ (By \textit{Binomial Theorem}, } \sum_{k=0}^{n} C_k^n x^k = (1+x)^n)
\end{aligned}$$

Finally, combine the results:

$$\begin{aligned}
&\mathbb{E}_f \Big\{ E_{OTS}(\mathcal{A}(\mathcal{D}), f) \Big\} \\
= &\sum_{k=1}^{2^L} P(f_k) \times E_{OTS}(f_i, f_k) \\
= &\frac{1}{2^L} \sum_{k=1}^{2^L} E_{OTS}(f_i, f_k) \\
= &\frac{1}{2^L} \sum_{t=1}^{L} t\, C_t^L \\
= &\frac{1}{2^L} 2^{L-1} \\
= &\frac{1}{2} = \textit{constant}
\end{aligned}$$

Thus, the statement is proved.

4

If we get five green 1's, then all the dice we pick are of kind A and D. Given that the quantity of dice is super large, each time we have $\frac{1}{4}+\frac{1}{4}=\frac{1}{2}$ of probability to pick an A or D. Then the probability of pick 5 dice of A or D is:

$$(\frac{1}{2})^5=\frac{1}{2^5}=\frac{1}{32}$$

5

For number 1 to 6, the conditions of making them all green is:

- 1: A or D
- 2: B or D
- 3: A or D
- 4: B or C
- 5: A or C
- 6: B or C

I find that the conditions of 1 and 3 are the same, and so are 4 and 6. Thus, the 6 conditions are reduced to 4 conditions: we get some number that is purely green if and only if one of the following conditions satisfied. Using the same method as previous problem, the probabilities of them (considered independently) can be easily calculated.

① all dice are of kind A or D → $P(\textcircled{1})=\frac{1}{32}$

② all dice are of kind A or C → $P(\textcircled{2})=\frac{1}{32}$

③ all dice are of kind B or D → $P(\textcircled{3})=\frac{1}{32}$

④ all dice are of kind B or C → $P(\textcircled{4})=\frac{1}{32}$

Then I consider the overlapping conditions:

- **exactly 2 conditions overlap**:
  - $\textcircled{1}\wedge\textcircled{2}$: $\{all\ dice\ are\ A\ or\ D\}\cap\{all\ dice\ are\ A\ or\ C\}=\{all\ dice\ are\ A\}$, $P(\textcircled{1}\wedge\textcircled{2})=\frac{1}{4^5}$
  - $\textcircled{1}\wedge\textcircled{3}$: $\{all\ dice\ are\ A\ or\ D\}\cap\{all\ dice\ are\ B\ or\ D\}=\{all\ dice\ are\ D\}$, $P(\textcircled{1}\wedge\textcircled{3})=\frac{1}{4^5}$
  - $\textcircled{1}\wedge\textcircled{4}$: $\{all\ dice\ are\ A\ or\ D\}\cap\{all\ dice\ are\ B\ or\ C\}=\emptyset$, $P(\textcircled{1}\wedge\textcircled{4})=0$
  - $\textcircled{2}\wedge\textcircled{3}$: $\{all\ dice\ are\ A\ or\ C\}\cap\{all\ dice\ are\ B\ or\ D\}=\emptyset$, $P(\textcircled{2}\wedge\textcircled{3})=0$
  - $\textcircled{2}\wedge\textcircled{4}$: $\{all\ dice\ are\ A\ or\ C\}\cap\{all\ dice\ are\ B\ or\ C\}=\{all\ dice\ are\ C\}$, $P(\textcircled{2}\wedge\textcircled{4})=\frac{1}{4^5}$
  - $\textcircled{3}\wedge\textcircled{4}$: $\{all\ dice\ are\ B\ or\ D\}\cap\{all\ dice\ are\ B\ or\ C\}=\{all\ dice\ are\ B\}$, $P(\textcircled{3}\wedge\textcircled{4})=\frac{1}{4^5}$
- **exactly 3 or 4 conditions overlap**: Because all of A, B, C, D appears only twice in the conditions, it is impossible that 3 or 4 conditions being satisfied at the same time. So their probabilities are 0.

Then, by the *Principle of Inclusion and Exclusion*, the answer is:

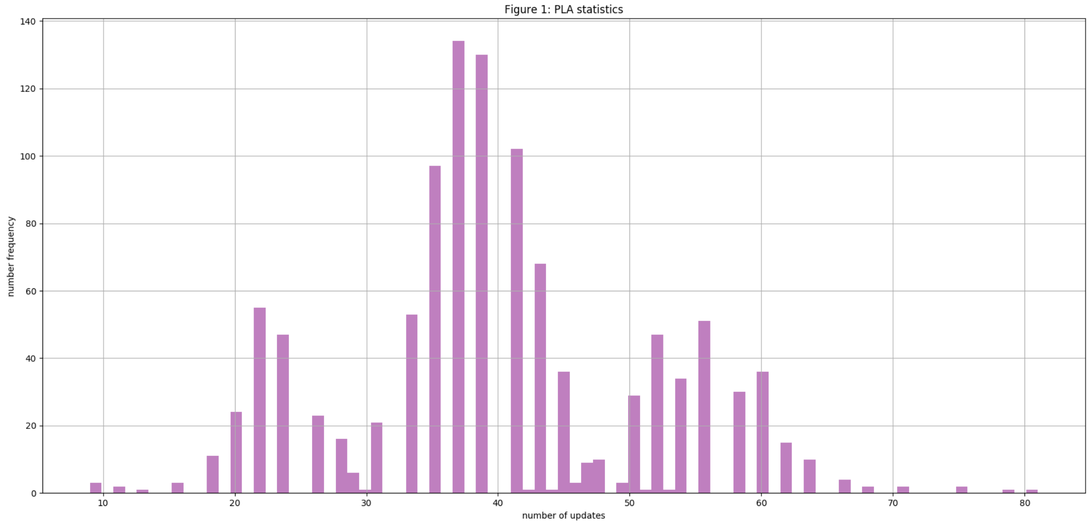
$$P(\textcircled{1}\vee\textcircled{2}\vee\textcircled{3}\vee\textcircled{4})=4(\frac{1}{32})-4(\frac{1}{4^5})+0-0=\frac{32}{256}-\frac{1}{256}=\frac{31}{256}$$

I find the two problems are similar to the **BAD Data** problem described in the lecture 4. Getting five green 1's is like a **BAD** data for a hypothesis (say it h_1) that make $E_{out}(h_1)$ and $E_{in}(h_1)$ far away, and the same for the other numbers. On the other hand, getting some numbers that is purely green is like a **BAD** data for "some hypotheses". Thus, I can try to analyze the two results from this perspective.

Comparing this result with the previous problem, $\frac{31}{256}$ is far larger than $\frac{1}{32}=\frac{8}{256}$, but still bounded by the *union bound* $=4$(assume the number of hypotheses is 4) $\times\frac{8}{256}=\frac{32}{256}$, which indicates the probability of **BAD** data can be seem as bounded.

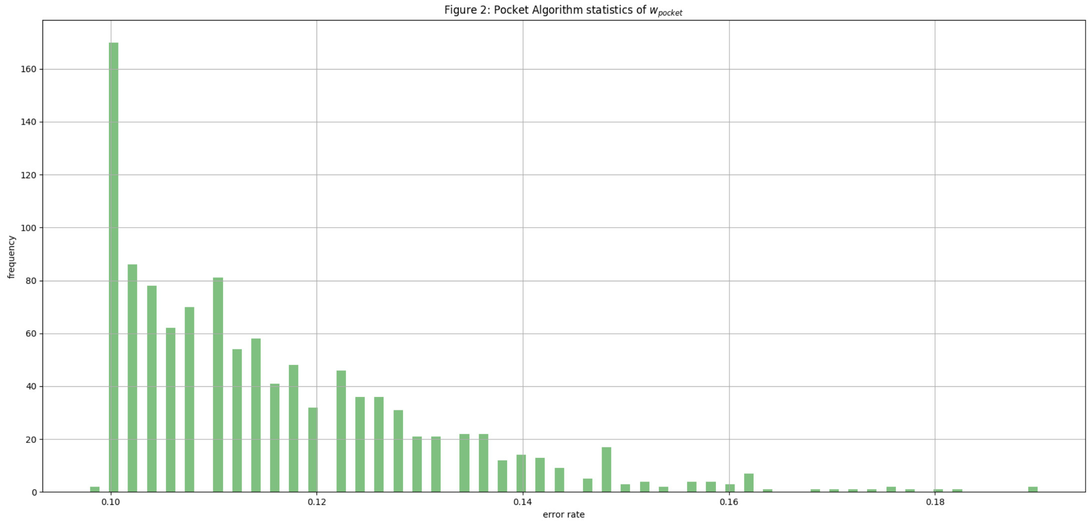
6

The average number of updates is **40.152** times.



7

The average error rate on the test set is **0.116**.



8

The average error rate of w_{100} on the test set is **0.320**, which is much worse than the preformance of w_{pocket}. I think it means w_t does not keep decreasing for every updates. That is, the algorithm may correct one error but generate many more error, though it ends up getting the best w if t becomes large enough. However, if it halts at a t that is not that large, the current w_t may not be the best among w_0 to w_t.

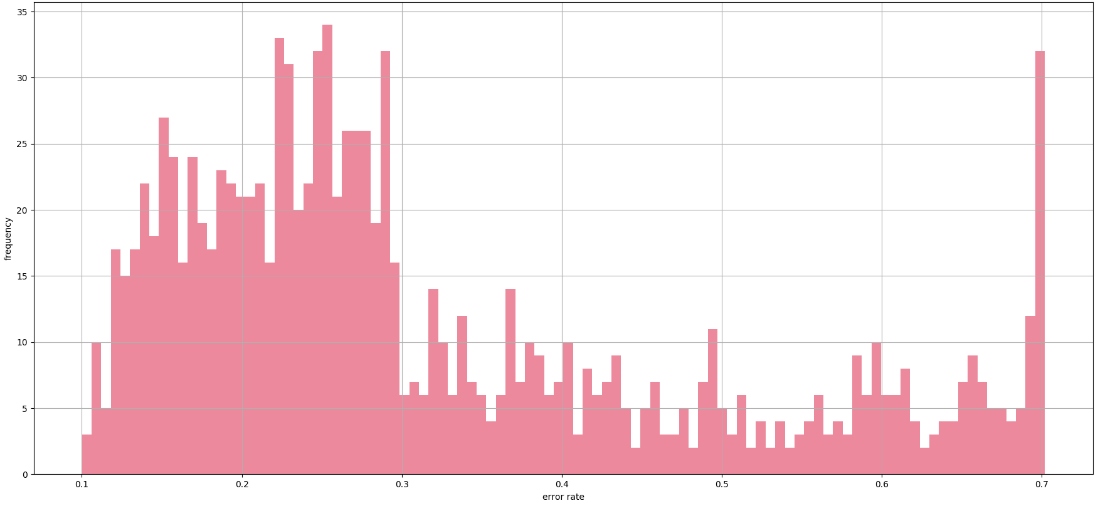


Figure 3: Pocket Algorithm statistics of w_{100}

9

If all x_n are scaled down linearly be a factor of 10, then the upper bound $\frac{R^2}{\rho^2}$ becomes:

$$\begin{aligned}\frac{R'^2}{\rho'^2} &= \frac{\max\limits_n ||\frac{x_n}{10}||^2}{(\min\limits_n y_n \frac{w_f^T}{||w_f||}\frac{x_n}{10})^2} \\ &= \frac{\frac{1}{10^2}\max\limits_n ||x_n||^2}{\frac{1}{10^2}(\min\limits_n y_n \frac{w_f^T}{||w_f||}x_n)^2} \\ &= \frac{\max\limits_n ||x_n||^2}{(\min\limits_n y_n \frac{w_f^T}{||w_f||}x_n)^2} \\ &= \frac{R^2}{\rho^2}\end{aligned}$$

The upper bound of T does not change at all, so his plan will not work.