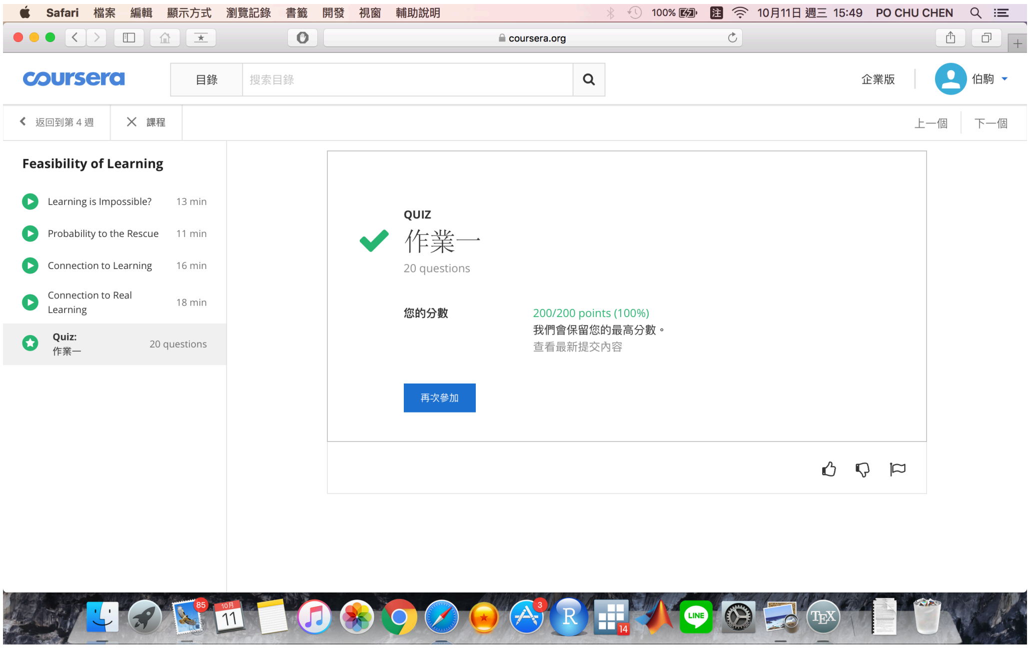
Home Work 1

Machine Learning Foundations

R04323050

經濟碩三　陳伯駒

1.



2.

Active Learning 中的主動是指讓機器 “主動” 去尋找使用者所需求的 label，資料結構上與半監督式學習有些相似，只有少部分是有經過標籤的，對於 unlabeled 的資料若要取得標籤將會浩費巨大的成本或是人力，此時將適用主動式學習。最常見的應用是在於生物與醫學實驗，像是蛋白質或是藥物的合成，都需要將化合物以某種特殊的的方式鍵結在一起。因此目標上我們想要找出某種函數形式 $f : \mathcal{X} \to \mathcal{Y}$ 是可以清楚地

分類出哪些化合物的合成方式是可以 bind to particular target(達到藥物成效)，所以 labeled Y_n 就可以分為 active(binds to target) & inactive，而其它 unlabeled Y_n 就是各種 description of chemical compounds。透過 Active Learning 我們可以讓機器透過學習來主動選擇可能的化學實驗，以改變化合物的鏈結方式 (或調配種類)，最後得到可能達到特殊需求的新化合物 (obtaining new labels)。

3.

$f(\mathbf{x_{N+\ell}}) = \mathbf{1}$, for $\ell = 1, 2, ..., L$

$$g(\mathbf{x_{N+\ell}}) = \begin{cases} 1 \text{ if } N+\ell \text{ is odd, } \ell = 1, 2, ..., L \\ -1, \text{ otherwise} \end{cases}$$

$\therefore \sum_{\ell=1}^{L} [\![g(\mathbf{x}_{N+\ell}) \neq f(\mathbf{x}_{N+\ell})]\!]$ 即 $N+1$ 至 $N+L$ 之間的偶數個數。

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

First we construct the proposition $\mathbf{P_n}$: the # of even number between $1 \sim n$ is $\lfloor \frac{n}{2} \rfloor$, where $n \geq 1$.

Case 1: Let n be even number

P_2: # of even number between $1 \sim 2$ is $1 = \lfloor \frac{2}{2} \rfloor$, holds.

P_4: # of even number between $1 \sim 4$ is $2 = \lfloor \frac{4}{2} \rfloor$, holds.

Suppose P_n holds, meaning # of even number between $1 \sim n$ is $\lfloor \frac{n}{2} \rfloor$

Consider P_{n+2}, then # of even number between $1 \sim n+2$ is $\lfloor \frac{n}{2} \rfloor + 1 =^{*} \lfloor \frac{n}{2} + 1 \rfloor = \lfloor \frac{n+2}{2} \rfloor$, holds.

Hence by mathematical induction, we know P_n holds for all n are even number.

Case 2: Let n be odd number

P_1: # of even number between $1 \sim 1$ is $1 = \lfloor \frac{1}{2} \rfloor$, holds.

---

*Let $\lfloor x \rfloor = m$, $\lfloor x+1 \rfloor = M$.

By the equivalence of floor function:

$\lfloor x \rfloor = m \Leftrightarrow m \leq x < m+1$

$\lfloor x+1 \rfloor = M \Leftrightarrow M \leq x+1 < M+1 \Leftrightarrow M-1 \leq x < M$

$\therefore M = m+1 \Leftrightarrow \lfloor x+1 \rfloor = \lfloor x \rfloor + 1$

P_3: # of even number between $1 \sim 3$ is $2 = \left\lfloor \frac{3}{2} \right\rfloor$, holds.

Suppose P_n holds, meaning # of even number between $1 \sim n$ is $\left\lfloor \frac{n}{2} \right\rfloor$

Consider P_{n+2}, similarly, P_{n+2} still holds.

Hence by mathematical induction, we know P_n holds for all n are odd number.

By case 1 & 2, we know P_n holds $\forall n$. Next, we will prove the **claim** by this proposition.

$$\begin{cases} A: N+L\text{為偶數，}N+1\text{為奇數} \\ B: N+L\text{為奇數，}N+1\text{為奇數} \end{cases} : \text{ \# of even number } = \left\lfloor \frac{N+L}{2} \right\rfloor - \left\lfloor \frac{N+1}{2} \right\rfloor$$

$$\begin{cases} C: N+L\text{為偶數，}N+1\text{為}\underline{\text{偶數}} \\ D: N+L\text{為奇數，}N+1\text{為}\underline{\text{偶數}} \end{cases} : \text{ \# of even number } = \left\lfloor \frac{N+L}{2} \right\rfloor - \left\lfloor \frac{N+1}{2} \right\rfloor + \underline{1}$$ ($\because$ 多扣 $N+1$ 這個偶數)

note that when:

甲. $N+1$ 為奇數時：$\left\lfloor \frac{N+L}{2} \right\rfloor = \left\lfloor \frac{N}{2} \right\rfloor$

乙. $N+1$ 為偶數時：$N-1$ 亦為偶數 ($\therefore \left\lfloor \frac{N-1}{2} \right\rfloor = \left\lfloor \frac{N}{2} \right\rfloor$)，另外 rewrite

$1 - \left\lfloor \frac{N+1}{2} \right\rfloor = -(\left\lfloor \frac{N+L}{2} \right\rfloor + (-1)) = -\left\lfloor \frac{N-1}{2} \right\rfloor$

By 甲、乙 & A、B、C、D：

we can rewrite the # of even number between $N+1 \sim N+L$ is : $\left\lfloor \frac{N+L}{2} \right\rfloor - \left\lfloor \frac{N}{2} \right\rfloor$, Q.E.D.

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

4.

We know $f(\mathbf{x}_n) = y_n$, $\forall(\mathbf{x_n}, y_n) \in \mathcal{D}$, where $\mathcal{D} = \{(\mathbf{x_n}, y_n)\}_{n=1}^{N}$。對於 training example 中的資料，f 組合已經固定，而 $\mathcal{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}, \mathbf{x_{N+1}}, ..., \mathbf{x_{N+L}}\}$，因此對剩下的 $\mathbf{x_{N+1}} \sim \mathbf{x_{N+L}}$ 共有 2^L 種取法 (L 個、每個有 ± 1 兩種)。

5.

Let the determisitc algorithm $\mathcal{A}$ defined by question, then

$$\begin{aligned}
\mathbb{E}_f\left\{E_{OTS}(\mathcal{A}(\mathcal{D}),f)\right\} &= \frac{1}{2^L}\sum_{i=1}^{2^L}\frac{1}{L}\sum_{\ell=1}^{L}[\![g(\mathbf{x}_{N+\ell})\neq f(\mathbf{x}_{N+\ell})]\!] \\
&= \frac{1}{2^L}\frac{1}{L}\sum_{\ell=1}^{L}\sum_{i=1}^{2^L}[\![g(\mathbf{x}_{N+\ell})\neq f(\mathbf{x}_{N+\ell})]\!] \quad \text{(since } f \text{ are equally likely in prob.)} \\
&= \frac{1}{2^L}\frac{1}{L}\sum_{\ell=1}^{L}(\frac{1}{2}\times\sum_{i=1}^{2^L}1) \\
&= \frac{1}{2^L}\frac{1}{L}\sum_{\ell=1}^{L}2^{L-1} = \frac{1}{2}
\end{aligned}$$

$\therefore$ 與所選演算法無關，故等式成立。

6.

若要選 5 次、所挑中的 1 都是綠色的，則一定要選中骰子 A 或 D，因此機率為 $(\frac{2}{4})^5 = \frac{1}{32}$。

7.

先 check 使數字全為綠色的骰子組合：

數字	骰子組合
1	A 或 D
2	B 或 D
3	A 或 D
4	B 或 C
5	A 或 C
6	B 或 C

因此選 5 次中，有 4 種骰子的組合可使"some number" 全為綠：{（A 或 D），（B 或 C），（B 或 D），（A 或 C）}，共有 $2^5 \times 4$ 種可能性。

但以上四種組合分別會重複計算 {(DDDDD), (AAAAA), (BBBBB), (CCCCC)} 這四

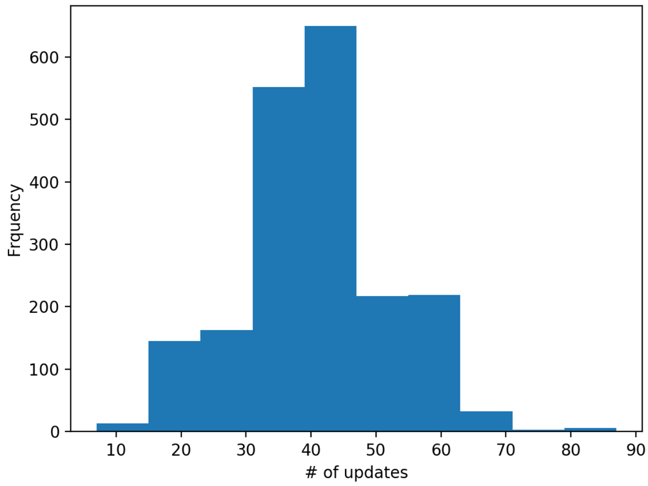
種組合，因此需扣除。

$\therefore$ 共有 $2^5 \times 4 - 4$ 種可能性，且 $N(\mathcal{S}) = 4^5$

$\therefore P = \frac{2^5 \times 4 - 4}{4^5} = \frac{31}{256}$

8.

Steps 的平均值接近 40 左右，且 histogram 近似於常態分佈。如下圖：



9.

In the slides, we know $T \leq \frac{R^2}{\rho^2}$, where $R^2 = \max_n \|\mathbf{x_n}\|^2$ and $\rho = \min_n y_n \frac{\mathbf{w_f^T}}{\|\mathbf{w_f}\|}\mathbf{x_n}$.

由上式可看出，T 的分子分母都有 $\|\mathbf{x_n}\|$ 的平方項，因此將所有的 $\mathbf{x_n}$ scale down linearly 對 T 的上界並無影響，無法讓演算法 overall 變快。