Designer:陳麒宇( R07943034 )

Date:2019/12/05

# 機器學習基石 HW2

1. Go register for the Coursera version of the first part of the class and solve its homework 2.


2. Consider the “positive rectangle" hypothesis set, which includes any hypothesis that returns +1 when x is within an axis-parallel rectangle and -1 elsewhere. Show that the VC Dimension of the hypothesis set is no less than 4. Please provide proof of your answer.

   可以將題意理解為證明平面上存在 4 個點，這 4 個點可以被 positive rectangle shatter，那麼 VC Dimension 必定>=4。

   首先假設平面上存在某四點可構成一正方形，窮舉各點可能情況：

   A. 四點皆為正：將正方形各邊向外擴張即可得到一矩形涵蓋四點
   B. 三正一負：取負號點的兩個鄰點連線，向負號點的反方向拉伸出矩形，直到三個點皆在矩形上時，各邊向外擴張即可得到一矩形涵蓋三點
   C. 二正二負：取兩正號點連線，將線段的長寬擴張後即可得到一矩形涵蓋兩點
   D. 一正三負：取正號點並用一矩形將其涵蓋即可
   E. 四點皆為負：矩形未涵蓋任何點

   可知平面存在 4 個點可以被 positive rectangle shatter，因此 VC Dimension 不會低於 4

3. Consider the “triangle waves" hypothesis set on R. What the VC-Dimension of such an H? Please prove your answer.

   首先分析函數 h(x)在數線上的含意，h(x)函數以 0 為基準將實數線以 4 為單位切割成無限個區間，若某點與該區間中心距離大於 1，則該點的函數值為正，反之則為負。

   因此對於 N 個點來說，我們必存在 2N 個區間能夠 shatter 這 N 個點，因此 VC Dimension 等於無限大

4. For any two hypothesis sets H1 and H2 that come with non-empty intersection, prove that $d_{vc}(H1 \cap H2) <= d_{vc}(H1)$.

   若 hypothesis 的數量越大則越有機會 shatter 更多點，反之亦然。因此若將 H1 與 H2 做交集後的大小勢必會小於等於原本的大小，因此能夠 shatter 的點也會小於等於原本的 hypothesis，所以 VC Dimension 也會小於等於原本的 VC Dimension

5. Consider H1 as the positive-ray hypothesis set (as discussed in class), and H2 as the negative-ray hypothesis set (which contains the negation of each hypothesis in H1). We showed that mH1 (N) = N + 1 in class. Write down mH1UH2 (N) and use that to calculate $d_{vc}$(H1 U H2). (Hint: This may partially help you solve Question 15 on Coursera.)

   $m_{H1UH2}(N) = 2N$，negative ray 的成長函數其實也是 N+1，但 positive 跟 negative 重疊的部分為全正與全負，因此扣掉這兩種狀況後結果為 2N。$d_{vc}$(H1 U H2) = 2，可透過不等式 2N<=N^2 求解出最大且能夠被 shatter 的 N

6. In the next problem, you are asked to implement such an algorithm and run your program on an artificial data set. We shall start by generating a one-dimensional data by the procedure below:

   (a) Generate x by a uniform distribution in [-1, 1].

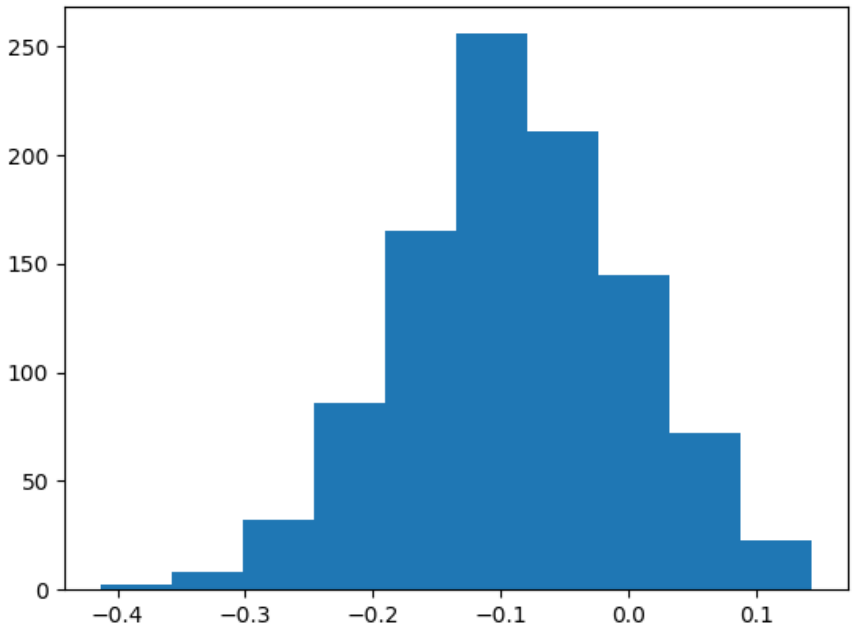
   (b) Generate y by f(x) = s(x) + noise where s(x) = sign(x) and the noise flips the result with 20% probability.

   For any decision stump $h_{s,\theta}$ with $\theta \epsilon$ [-1, 1], express $E_{out}(h_{s,\theta})$ as a function of θ and s. Write down your derivation steps.

   對於加入 noise 的 target function，可以透過 λμ+(1-λ)(1-μ)來表示 $E_{out}$，其中 λ 表示受到 noise 影響的正確率，在本題中加入 20% noise 因此 λ 為 0.8。
   而 μ 表示無 noise 情況下 hypothesis 的錯誤率，可以比對兩函數 f(x) = sign(x)與 h(x) = s*sign(x-θ)來計算，由於 x 為範圍在-1~1 之間的 uniform distribution，因此可直接用區間所占比例表示機率，對於 s = 1 比對 sign(x)與 sign(x-θ)會發現有大小為|θ|的區間沒有重疊，亦即在範圍為 2 的區間中有範圍|θ|的區間答案錯誤，錯誤率為|θ|/2，同理對於 s = -1 將兩者的 sign 函數比對未重疊部分為 2-|θ|，錯誤率為(2-|θ|)/2，使用線性合併求得 μ = (s+1)/2*(|θ|/2) - (s-1)/2*((2-|θ|)/2)
   最後將 λ = 0.8 與 μ = (s+1)/2*(|θ|/2) - (s-1)/2*((2-|θ|)/2)代入 $E_{out}$ = λμ+(1-λ)(1-μ)，可得 $E_{out}$ = 0.5+0.3*s*(|θ|-1)

7. Generate a data set of size 20 by the procedure above and run the one-dimensional decision stump algorithm on the data set. Record $E_{in}$ and compute $E_{out}$ with the formula above. Repeat the experiment 1000 times and plot a histogram of $E_{in}$ - $E_{out}$. Describe your findings.

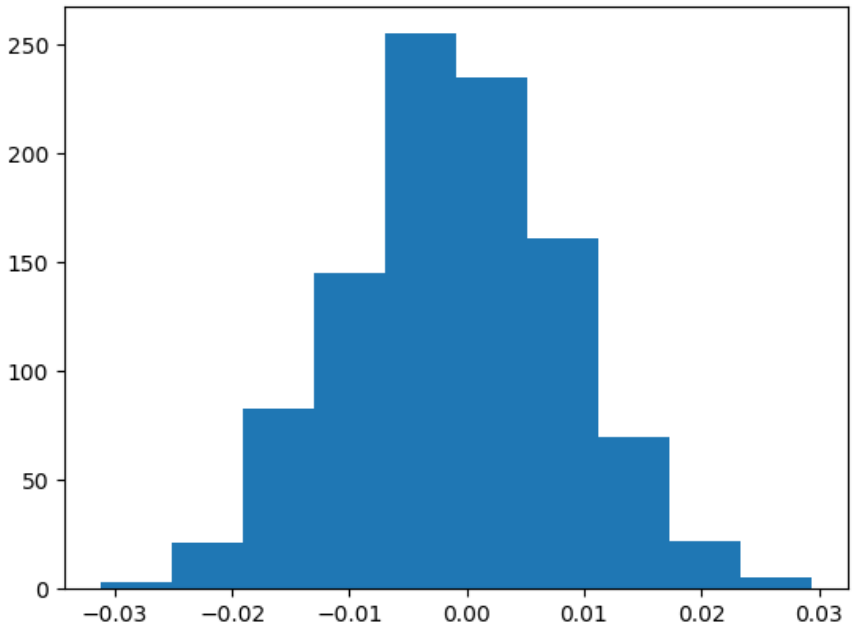
   $E_{in} = 0.171$, $E_{out} = 0.261$




   在實現過程中我們嘗試找出能產生最小 $E_{in}$ 的 hypothesis，在本題中我們的 hypothesis 就是 s 跟 θ，用它們來計算出 $E_{out}$ 的數值，比較 tricky 的部分是 20%的 noise 使得 $E_{in}$ 在尋找 hypothesis 時會去找出能夠使自己最小的 hypothesis，但找到的 hypothesis 可能會使得 $E_{out}$ 增大，所以才會有一種 $E_{in}$ 略小於 0.2 而 $E_{out}$ 略大於 0.2 的感覺

8. Generate a data set of size 2000 by the procedure above and run the one-dimensional decision stump algorithm on the data set. Record $E_{in}$ and compute $E_{out}$ with the formula above. Repeat the experiment 1000 times and plot a histogram of $E_{in}$ - $E_{out}$. Describe your findings and compare the findings with those in the previous problem.

   $E_{in} = 0.199$, $E_{out} = 0.200$




本題比較直觀的解釋就是 data size 越大，我們訓練出來的模型就更能處理沒看過的資料，因此 $E_{in}$ 會更接近 $E_{out}$。另一種解釋就是資料量變大使得 θ 的分布也相對均勻了許多，因此對錯誤率的影響會轉移到 noise 上，對 $E_{in}$ 來說可以嘗試找到一個抵銷 noise 干擾的最佳數值，但 noise 影響在資料上，因此將 hypothesis 傳給 $E_{out}$ 計算時只會看到 θ 的影響，因此 $E_{out}$ 的數值是非常穩定的，而 $E_{in}$ 是類似常態的分布相減後就是上圖

9. Prove what the VC-Dimension of the hypothesis set in Question 10 on Coursera is.

   從本題的敘述中可得知 hypothesis 要將 x 歸類至大小為 2^d 的 hyper-rectangular region，而每個 region 可為 1 or -1，因此總共會有 2^(2^d)種組合，因此不論我們的 N 為何，組合數都只會與 d 有關，從而得知 $2^{(d_{vc})} = 2^{(2^d)} \Rightarrow d_{vc} = 2^d$