

# ECE 595: Homework 6

Yi Qiao, Class ID 187

(Spring 2019)

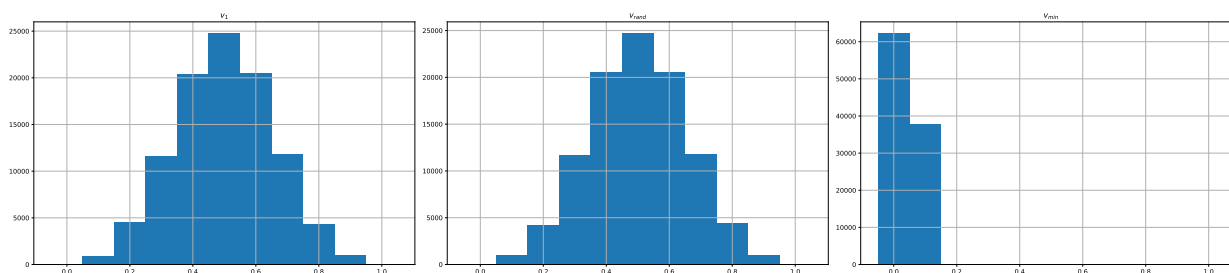
## Exercise 1: Hoeffding Inequality

(a) probability of getting a head for coins  $c_1$ ,  $c_{rand}$  and  $c_{min}$

Since they are all fair coins,

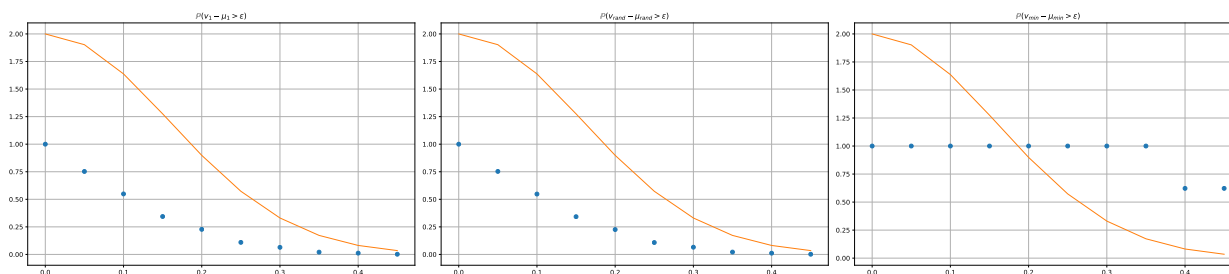
$$\mathbb{P}(c_1 = head) = \mathbb{P}(c_{rand} = head) = \mathbb{P}(c_{min} = head) = 0.5$$

(b) python experiment



The first two are binominal, while the third one is not.

(c) plots



(d)

By observation, we can see that the first two obviously obey the Hoeffding's bound while the third one does not. This is because of its mean is different from its expectation, the expectation is around 0.38 and the mean is 0.5.

(e)

This experiments states that, if we want to bound the error by Hoeffding, we have to make sure our data is i.i.d. and the expectation and mean are the same.

## Exercise 2: VC Dimension

### (a) Compute the VC dimension

(i)

$$\mathcal{H} = \{h : \mathbb{R} \rightarrow \{-1, +1\} | h(x) = +1, \forall x \in [a, \infty), a \in \mathbb{R}\} \cup \{h : \mathbb{R} \rightarrow \{-1, +1\} | h(x) = +1, \forall x \in (-\infty, a], a \in \mathbb{R}\} \quad (1)$$

By inspection, the VC dimension of the above hypothesis set is 2. Since the hypothesis set is composed by two step functions that are symmetric to each other, they can shatter 2 points in 4 ways while obviously with one of the functions can get 3 combinations,  $\{[+1, +1], [+1, -1], [-1, -1]\}$  by changing  $a$ , the other one will cover  $\{[-1, -1], [-1, +1], [+1, +1]\}$ , thus the union will be  $\{[+1, +1], [+1, -1], [-1, +1], [-1, -1]\}$ . While it is linear in a 1D space, it cannot shatter more than 2 points.

(ii)

$$\mathcal{H} = \{h : \mathbb{R} \rightarrow \{-1, +1\} | h(x) = +1, \forall x \in [a, b], a, b \in \mathbb{R}\} \cup \{h : \mathbb{R} \rightarrow \{-1, +1\} | h(x) = -1, \forall x \in [a, b], a, b \in \mathbb{R}\} \quad (2)$$

By inspection, the VC dimension of the above hypothesis set is 3. Similarly to the one above, The first hypothesis will cover  $\{[-1, -1, -1], [-1, -1, +1], [-1, +1, +1], [+1, +1, +1], [+1, +1, -1], [+1, -1, -1]\}$ , the second one will cover  $\{[+1, +1, +1], [+1, +1, -1], [+1, -1, -1], [-1, -1, -1], [-1, -1, +1], [-1, +1, +1]\}$ , the Union will cover all eight of them  $\{[-1, -1, -1], \dots, [+1, +1, +1]\}$ . While it is quadratic, it cannot shatter more than 3 points.

(iii)

$$\mathcal{H} = \left\{ h : \mathbb{R}^d \rightarrow \{-1, +1\} | h(x) = +1, \forall x \text{ where } \sqrt{\sum_{j=1}^d x_j^2} \leq b, b \in \mathbb{R} \right\} \quad (3)$$

By inspection, hypothesis function is a hyper ball. Thus, the VC dimension of the above hypothesis set is 1.

(b)

$$\mathcal{H} = \left\{ h_\alpha : \mathbb{R} \rightarrow \mathbb{R} \mid h_\alpha(x) = (-1)^{\lfloor \alpha x \rfloor}, \alpha \in \mathbb{R} \right\} \quad (4)$$

Let  $N$  training samples be  $\{x_i \in \mathbb{R} \mid i = 1, 2, \dots, N\}$  and the corresponding class be  $\{c_i \in \{-1, +1\} \mid i = 1, \dots, N\}$ .

For the base case  $N = 1$ , obviously,  $\alpha = 0$  and  $\alpha = 1$  can generate the two dichotomies. Since that, essentially, we take the floor of  $\alpha x$ , we do not care the decimal places. Or, if we choose  $x_i = 10^i$ , we do not care about what is left after the  $(N - 1)$ th decimal place of  $\alpha$ , since it will be floored out eventually.

Thus, if we have a hypothesis set  $\mathcal{H}_\alpha(x)$  who can shatter  $(N - 1)$  samples, and  $A$  be the set of all  $\alpha$  used in  $\mathcal{H}_\alpha(x)$ , we can always find the new set,

$$A' = \{\alpha' = \alpha + k10^{(1-N)} \mid \alpha \in A, k \in \{0, 1\}\}$$

where when  $k = 0$ , the  $N$ th sample is classified as  $-1$ , when  $k = 1$ , the  $N$ th sample is classified as  $+1$ . Thus,

$$\mathcal{H}_{\alpha'} = \left\{ h_{\alpha'} : \mathbb{R} \rightarrow \mathbb{R} \mid h_{\alpha'}(x) = (-1)^{\lfloor \alpha' x \rfloor}, \alpha' \in A' \right\}$$

can shatter  $N$  points.

By induction, we can see that  $N$  is not bounded at all,  $VC_{dimension} = \infty$ .

This is far worse than the perceptron due to infinite VC-dimension. The model is far too complicated thus have no ability to generalize at all.

### Exercise 3: Bias-Variance Trade-off

(a)

$$\begin{aligned}
\boldsymbol{\theta}_{\mathcal{D}} &= \underset{\boldsymbol{\theta}_h}{\operatorname{argmin}} E_{\text{aug}}(h) \\
&= \underset{\boldsymbol{\theta}_h}{\operatorname{argmin}} E_{\text{in}}(h) + \frac{\lambda}{N} \boldsymbol{\theta}_h^T \boldsymbol{\theta}_h \\
&= \underset{\boldsymbol{\theta}_h}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^N (\boldsymbol{\theta}_h^T \mathbf{x}_n - y_n)^2 + \frac{\lambda}{N} \boldsymbol{\theta}_h^T \boldsymbol{\theta}_h \\
&= \underset{\boldsymbol{\theta}_h}{\operatorname{argmin}} \left\| \begin{bmatrix} \mathbf{x}_1^T & 1 \\ \mathbf{x}_2^T & 1 \\ \vdots & \vdots \\ \mathbf{x}_N^T & 1 \end{bmatrix} \boldsymbol{\theta}_h - \mathbf{y} \right\|_2^2 + \lambda \boldsymbol{\theta}_h^T \boldsymbol{\theta}_h
\end{aligned} \tag{5}$$

substitute  $\begin{bmatrix} \mathbf{x}_1^T & 1 \\ \mathbf{x}_2^T & 1 \\ \vdots & \vdots \\ \mathbf{x}_n^T & 1 \end{bmatrix}$  with  $\mathbf{A}$ , we get

$$\boldsymbol{\theta}_{\mathcal{D}} = \underset{\boldsymbol{\theta}_h}{\operatorname{argmin}} \|\mathbf{A}\boldsymbol{\theta}_h - \mathbf{y}\|_2^2 + \lambda \boldsymbol{\theta}_h^T \boldsymbol{\theta}_h \tag{6}$$

Taking the derivative, we can see

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}_h} &= 2\mathbf{A}^T(\mathbf{A}\boldsymbol{\theta}_h - \mathbf{y}) + 2\lambda\boldsymbol{\theta}_h = 0 \\
(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})\boldsymbol{\theta}_h &= \mathbf{A}^T\mathbf{y} \\
\boldsymbol{\theta}_{\mathcal{D}} = \boldsymbol{\theta}_h^* &= (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T(\mathbf{A}\boldsymbol{\theta}_f + \boldsymbol{\epsilon})
\end{aligned} \tag{7}$$

(b)

Continue from the last problem, expand what we already got,

$$\begin{aligned}
\boldsymbol{\theta}_{\mathcal{D}} &= (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T(\mathbf{A}\boldsymbol{\theta}_f + \boldsymbol{\epsilon}) \\
&= (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\mathbf{A}\boldsymbol{\theta}_f + (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\boldsymbol{\epsilon} \\
&= (\mathbf{I} - \lambda(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1})\boldsymbol{\theta}_f + (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\boldsymbol{\epsilon} \\
&= \boldsymbol{\theta}_f - \lambda(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\boldsymbol{\theta}_f + (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\boldsymbol{\epsilon}
\end{aligned} \tag{8}$$

(c)

(i)

$$\begin{aligned}
\bar{g}(\mathbf{x}) &= \mathbb{E}_{\mathcal{D}} [h^{(\mathcal{D})}(\mathbf{x})] = \mathbb{E}_{\mathcal{D}} [\boldsymbol{\theta}_{\mathcal{D}}^T \mathbf{x}] \\
&= \mathbb{E}_{\mathcal{D}} \left[ (\boldsymbol{\theta}_f - \lambda(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\boldsymbol{\theta}_f + (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\boldsymbol{\epsilon})^T \mathbf{x} \right] \\
&= \boldsymbol{\theta}_f^T \mathbf{x} - \lambda \mathbf{x}^T \mathbb{E}_{\mathbf{A}} [(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}] \boldsymbol{\theta}_f + \mathbf{x}^T \mathbb{E}_{\mathbf{A}} [(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1} \mathbf{A}^T] \mathbb{E}[\boldsymbol{\epsilon}] \\
&= \boldsymbol{\theta}_f^T \mathbf{x} - \lambda \mathbf{x}^T \mathbb{E}_{\mathbf{A}} [(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}] \boldsymbol{\theta}_f
\end{aligned} \tag{9}$$

(ii)

$$\begin{aligned}
(\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2 &= (\boldsymbol{\theta}_f^T \mathbf{x} - \lambda \mathbf{x}^T \mathbb{E}_{\mathbf{A}}[(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1}] \boldsymbol{\theta}_f - \boldsymbol{\theta}_f^T \mathbf{x})^2 \\
&= \lambda^2 (\mathbf{x}^T \mathbb{E}_{\mathbf{A}}[(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1}] \boldsymbol{\theta}_f)^T (\mathbf{x}^T \mathbb{E}_{\mathbf{A}}[(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1}] \boldsymbol{\theta}_f) \\
&= \lambda^2 (\boldsymbol{\theta}_f^T \mathbb{E}_{\mathbf{A}}[(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1}]^T \mathbf{x} \mathbf{x}^T \mathbb{E}_{\mathbf{A}}[(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1}] \boldsymbol{\theta}_f) \\
&= \lambda^2 \text{trace}(\mathbf{x} \mathbf{x}^T (\mathbb{E}_{\mathbf{A}}[(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1}] \boldsymbol{\theta}_f \boldsymbol{\theta}_f^T \mathbb{E}_{\mathbf{A}}[(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1}]))
\end{aligned} \tag{10}$$

(iii)

Plug in  $\mathbf{A}^T \mathbf{A} \approx N \mathbf{I}$ , we got

$$\begin{aligned}
\text{bias} &= \mathbb{E}_{\mathcal{X}}[(\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2] \approx \lambda^2 \text{trace}(\mathbb{E}_{\mathcal{X}}[\mathbf{x} \mathbf{x}^T] ((N + \lambda) \mathbf{I})^{-1} \boldsymbol{\theta}_f \boldsymbol{\theta}_f^T ((N + \lambda) \mathbf{I})^{-1}) \\
&= \frac{\lambda^2}{(N + \lambda)^2} \text{trace}(\boldsymbol{\theta}_f \boldsymbol{\theta}_f^T) \\
&= \frac{\lambda^2}{(N + \lambda)^2} \boldsymbol{\theta}_f^T \boldsymbol{\theta}_f \\
&= \frac{\lambda^2}{(N + \lambda)^2} \|\boldsymbol{\theta}_f\|_2^2
\end{aligned} \tag{11}$$

(iv)

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[(h^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2] &= \mathbb{E}_{\mathcal{D}}[(\boldsymbol{\theta}_{\mathcal{D}}^T \mathbf{x} - \boldsymbol{\theta}_f^T \mathbf{x} + \lambda \mathbf{x}^T \mathbb{E}_{\mathbf{A}}[(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1}] \boldsymbol{\theta}_f)^2] \\
&= \mathbb{E}_{\mathcal{D}}[(\boldsymbol{\theta}_f - \lambda(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \boldsymbol{\theta}_f + (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \boldsymbol{\epsilon})^T \mathbf{x} - \boldsymbol{\theta}_f^T \mathbf{x} + \lambda \mathbf{x}^T \mathbb{E}_{\mathbf{A}}[(\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1}] \boldsymbol{\theta}_f)^2] \\
&\approx \mathbb{E}_{\mathcal{D}}[(-\lambda \mathbf{x}^T (N \mathbf{I} + \lambda \mathbf{I})^{-1} \boldsymbol{\theta}_f + \mathbf{x}^T (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \boldsymbol{\epsilon} + \lambda \mathbf{x}^T (N \mathbf{I} + \lambda \mathbf{I})^{-1} \boldsymbol{\theta}_f)^2] \\
&= \mathbb{E}_{\mathcal{D}}[(\mathbf{x}^T (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \boldsymbol{\epsilon})^2] \\
&= \mathbb{E}_{\mathbf{A}}[(\mathbf{x}^T (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T)^2] \mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^2] \\
&= \sigma^2 \mathbb{E}_{\mathbf{A}}[(\mathbf{x}^T (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T)^T (\mathbf{x}^T (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T)] \\
&= \sigma^2 \mathbb{E}_{\mathbf{A}}[\text{trace}(\mathbf{A} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{x} \mathbf{x}^T (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T)] \\
&= \sigma^2 \mathbb{E}_{\mathbf{A}}[\text{trace}(\mathbf{x} \mathbf{x}^T (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{A} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1})]
\end{aligned} \tag{12}$$

(v)

$$\begin{aligned}
\text{var} &= \mathbb{E}_{\mathcal{X}}[\mathbb{E}_{\mathcal{D}}[(h^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2]] \\
&= \sigma^2 \mathbb{E}_{\mathbf{A}}[\text{trace}(\mathbb{E}_{\mathcal{X}}[\mathbf{x} \mathbf{x}^T] (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{A} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1})] \\
&= \sigma^2 \mathbb{E}_{\mathbf{A}}[\text{trace}(\mathbf{I} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{A} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1})] \\
&= \sigma^2 \mathbb{E}_{\mathbf{A}}[\text{trace}((\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{A} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{A} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1})] \\
&\approx \frac{\sigma^2}{N} \mathbb{E}_{\mathbf{A}}[\text{trace}(\mathbf{A} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{A} (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T)] \\
&= \frac{\sigma^2}{N} \mathbb{E}_{\mathbf{A}}[\text{trace}(H^2(\lambda))]
\end{aligned} \tag{13}$$