ECE 595: Homework 6

Yi Qiao, Class ID 187

(Spring 2019)

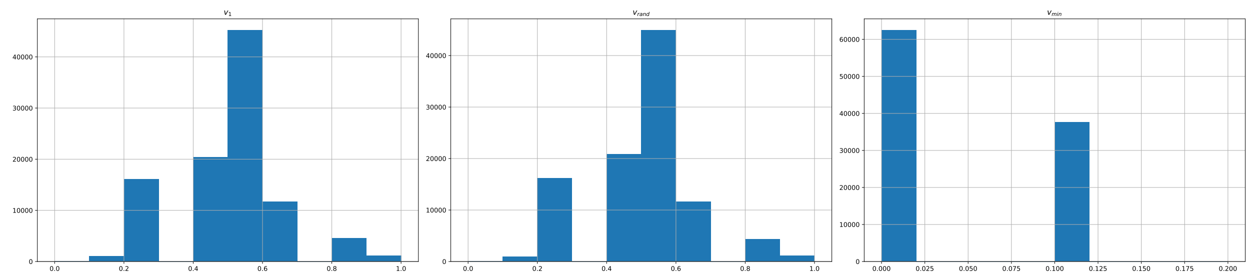
Exercise 1: Hoeffding Inequality

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

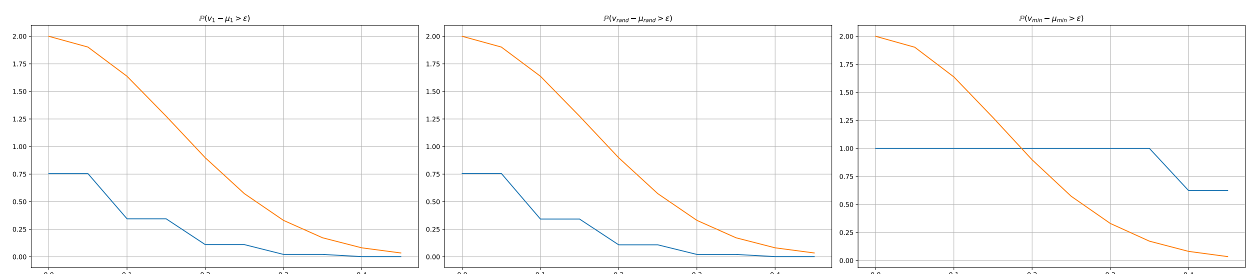
Since they are all fair coins,

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

(b) python experiment



(c) plots



(d)

By observation, we can see that the first two obviously obey the Hoeffding's bound while the third one does not.

(e)

something something...

Exercise 2: VC Dimension

(a) Compute the VC dimension

(i)

$$\begin{aligned}\mathcal{H} =& \{h : \mathbb{R} \to \{-1,+1\} | h(x) = +1, \forall x \in [a,\infty), a \in \mathbb{R}\} \cup \\ & \{h : \mathbb{R} \to \{-1,+1\} | h(x) = +1, \forall x \in (-\infty, a], a \in \mathbb{R}\}\end{aligned} \tag{1}$$

By inspection, the VC dimension of the above hypothesis set is 2.

(ii)

$$\begin{aligned}\mathcal{H} =& \{h : \mathbb{R} \to \{-1,+1\} | h(x) = +1, \forall x \in [a,b], a, b \in \mathbb{R}\} \cup \\ & \{h : \mathbb{R} \to \{-1,+1\} | h(x) = -1, \forall x \in [a,b], a, b \in \mathbb{R}\}\end{aligned} \tag{2}$$

By inspection, the VC dimension of the above hypothesis set is 3.

(iii)

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

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

(b)

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

Even though the above hypothesis set has only one parameter, it is periodic, thus by tuning the period/frequency, you can match any number of data points you want by finding their GCD. This hypothesis set has simply too large VC dimension, which is far beyond the model complexity. Thus, this will perform far worse than perceptron due to over-fitting.

Exercise 3: Bias-Variance Trade-off

(a)

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

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

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

Taking the derivative, we can see

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

(b)

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

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

(c)

(i)

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

(ii)

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

(iii)

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

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

(iv)

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

(v)

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