Homework 2 - Theory Part

Reza Bayat and Arian Khorasani

November 2021

1 Question 1

These two rules are applied in the following sections:

$$\mathbb{E}[A] = \mathbb{E}[A] + \mathbb{E}[B]$$

$$\mathbb{E}[(A+B)^2] = \mathbb{E}[A^2 + B^2 + 2AB] = \mathbb{E}[A^2] + \mathbb{E}[B^2] + 2\mathbb{E}[AB]$$

We can define the expected prediction error on (x', y') as a following term:

$$x' = \mathbb{E}_{\text{train },y'}\left[\left(y' - h_D\left(x'\right)\right)^2\right]$$

$$\begin{aligned}
\mathbb{E}_{\text{train },y'}\left[\left(y' - h_D\left(x'\right)\right)^2\right] &= \mathbb{E}_{\text{train},y'}\left[\left(f(x') + \epsilon - h_D\left(x'\right)\right)^2\right] \\
&= \mathbb{E}_{\text{train},y'}\left[\left(\left(f(x') - h_D\left(x'\right)\right) + \epsilon\right)^2\right] \\
&= \mathbb{E}\left[\left(f(x') - h_D\left(x'\right)\right)^2\right] + \mathbb{E}\left[\epsilon^2\right] + 2\mathbb{E}\left[\left(f(x') - h_D\left(x'\right)\right)\right]\mathbb{E}\left[\epsilon\right] \\
&= \mathbb{E}\left[\left(f(x') - h_D\left(x'\right)\right)^2\right] + \sigma_\epsilon^2
\end{aligned}$$

Note that $\mathbb{E}[\epsilon] = 0$

We can now derive the first term from above result:

$$\begin{aligned}
\mathbb{E}\left[\left(f(x') - h_D\left(x'\right)\right)^2\right] &= \mathbb{E}\left[\left(f\left(x'\right) - h_D\left(x'\right)\right)^2\right] \\
&= \mathbb{E}\left[\left(\left(f\left(x'\right) - \mathbb{E}\left[h_D\left(x'\right)\right]\right) + \left(\mathbb{E}\left[h_D\left(x'\right)\right] - h_D\left(x'\right)\right)\right)^2\right] \\
&= \mathbb{E}\left[\left(f(x') - \mathbb{E}\left[h_D\left(x'\right)\right]\right)^2\right] + 2\mathbb{E}\left[\left(f(x') - \mathbb{E}\left[h_D\left(x'\right)\right]\right)\left(\mathbb{E}\left[h_D\left(x'\right)\right] - h_D\left(x'\right)\right)\right] \\
&+ \mathbb{E}\left[\left(\mathbb{E}\left[h_D\left(x'\right)\right] - h_D\left(x'\right)\right)^2\right]
\end{aligned}$$

We have the following facts:

$$\mathbb{E}\left[\left(f(x') - \mathbb{E}\left[h_D\left(x'\right)\right]\right)^2\right] = bias^2$$

$$\mathbb{E}\left[\left(\mathbb{E}\left[h_D\left(x'\right)\right]-h_D\left(x'\right)\right)^2\right]=variance$$

Note: the bias term is independent of the expectation operator, so by continuing on above expression, we have:

$$=bias^2+2\,bais\,\mathbb{E}\left[\left(\mathbb{E}\left[h_D\left(x'\right)\right]-h_D\left(x'\right)\right)\right]+variance$$

Since $\mathbb{E}\left[\left(\mathbb{E}\left[h_D\left(x'\right)\right]-h_D\left(x'\right)\right)\right]=0$:

$$\mathbb{E}\left[\left(f(x')-h_D\left(x'\right)\right)^2\right]=\sigma_\epsilon^2+bias^2+variance$$

2 Question 2

2.1 a

Yes

$$\phi(x) = \begin{cases} x^2 & \text{if} \quad 2k \leq x < 2k+1 \\ -x^2 & \text{if} \quad 2k+1 \leq x < 2k+2 \end{cases}$$

All Xs will have positive values and all Os will have negative values, so they are linearly separable.

2.2 b

Yes

$$\phi(x) = x_1 x_2$$

All yellow points will have positive values and all blue points will have negative values, so they are linearly separable; however, some blue(yellow) points in first(second) region may overlap with another points in third(fourth) region. For example, $(2, 2)$ which is a yellow point in the first region, will overlap with $(-2, -2)$ in the third region, since $2 \times 2 = (-2) \times (-2) = 4$

2.3 c

Yes, we can define the radius of a circle by r, so one way to make the dataset linearly separable is presented in the following:

$$r^2 = x_1^2 + x_2^2$$

$$\phi(x) = \begin{cases} (x_1, x_2, x_1^2 + x_2^2) & \text{if} \quad 2k \leq x < 21+1 \\ (x_1, x_2, -(x_1^2 + x_2^2)) & \text{if} \quad 2k+1 \leq x < 21+2 \end{cases}$$

The kernel of the above feature map is:

$$K(x, x') = \phi(x)^T \phi(x') = \begin{cases} (x_1 x_1', x_2 x_2', (x_1^2 + x_2^2)((x_1')^2 + (x_2')^2))) & \text{if} \quad 2k \leq x < 2k+1 \\ (x_1 x_1', x_2 x_2', -(x_1^2 + x_2^2)((x_1')^2 + (x_2')^2)) & \text{if} \quad 2k+1 \leq x < 2k+2 \end{cases}$$

3 Question 3

3.1 a

$$
\begin{aligned}
\frac{\mathrm{d}f(x)}{\mathrm{d}x} &= \frac{\mathrm{d}\left(\log(x^4)sin(x^3)\right)}{\mathrm{d}x} \\
&= \frac{\mathrm{d}\left(\log(x^4)\right)}{\mathrm{d}x}sin(x^3) + \frac{\mathrm{d}\left(sin(x^3)\right)}{\mathrm{d}x}\log(x^4) \\
&= \frac{4sin(x^3)}{x} + 3x^2cos(x^3)\log(x^4)
\end{aligned}
$$

3.2 b

$$
\frac{\mathrm{d}\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)}{\mathrm{d}x} = -\frac{1}{2\sigma^2}2(x-\mu) = \frac{-(x-\mu)}{\sigma^2}
$$

$$
\begin{aligned}
\frac{\mathrm{d}f(x)}{\mathrm{d}x} &= \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)\frac{\mathrm{d}\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)}{\mathrm{d}x} \\
&= \frac{-(x-\mu)}{\sigma^2}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)
\end{aligned}
$$

3.3 c

i) Dimensions

The dimension of $\frac{\partial f_1}{\partial x}$ is 1×2.
The dimension of $\frac{\partial f_2}{\partial x}$ is $1 \times n$.
The dimension of $\frac{\partial f_3}{\partial x}$ is $n^2 \times n$.

ii) Jacobians

The jacobian of f_1:

$$
\frac{\partial f_1}{\partial x} = \left[\frac{\partial(\sin(x_1)\cos(x_2))}{\partial x_1} \quad \frac{\partial(\sin(x_1)\cos(x_2))}{\partial x_2}\right]
$$

$$
\frac{\partial f_1}{\partial x} = \left[\begin{array}{cc} \cos(x_1)\cos(x_2) & -\sin(x_1)\sin(x_2) \end{array}\right]
$$

The jacobian of f_2:

$$
f_2(x,y) = x^\top y = x_1y_1 + \cdots + x_ny_n
$$

$$
\frac{\partial f_1}{\partial x} = y^\top
$$

The jacobian of f_3:

$$x^\top x = \begin{bmatrix} x_1^2 & x_1x_2 & \cdots & x_1x_n \\ x_2x_1 & x_2^2 & \cdots & x_2x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_nx_1 & x_nx_2 & \cdots & x_n^2 \end{bmatrix}$$

The derivative of the above matrix will be in the higher order.
The derivative of $x^\top x$ with respect to x_1:

$$x^\top x = \begin{bmatrix} 2x_1 & x_2 & \cdots & x_n \\ x_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_1 & 0 & \cdots & 0 \end{bmatrix}$$

The derivative of $x^\top x$ with respect to x_2:

$$x^\top x = \begin{bmatrix} 0 & x_1 & \cdots & 0 \\ x_1 & 2x_2 & \cdots & x_n \\ \vdots & \vdots & \ddots & \vdots \\ 0 & x_n & \cdots & 0 \end{bmatrix}$$

$$\vdots$$

And the derivative of $x^\top x$ with respect to x_n:

$$x^\top x = \begin{bmatrix} 0 & 0 & \cdots & x_1 \\ 0 & 0 & \cdots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1 & x_2 & \cdots & 2x_n \end{bmatrix}$$

3.4 d

i)
The dimension of $\frac{df}{dz}$ is 1×1, and it is: $-\frac{1}{2}\exp\left(-\frac{1}{2}z\right)$
The dimension of $\frac{dz}{dy}$ is $1 \times D$, and it is: $y^\top\left(S^{-1} + \left(S^{-1}\right)^\top\right)$
The dimension of $\frac{dy}{dx}$ is $D \times D$, and it is a identity matrix.
In the chain rule, multiply all derivatives to get the final derivative.

ii)
The dimension of $\frac{df}{dx}$ is $1 \times D$:

$$f(x) = \operatorname{tr}\left(xx^T + \sigma I\right) = x_1^2 + \cdots + x_n^2 + n\sigma^2$$

So:

$$\frac{df}{dx} = 2x^\top$$

iii)

The dimension of $\frac{df}{dz}$ is $M \times M$

$$\frac{df}{dz} = \begin{bmatrix} 1 - \tanh^2(z_1) & 0 & \cdots & 0 \\ 0 & 1 - \tanh^2(z_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 - \tanh^2(z_M) \end{bmatrix}$$

Or:

$$\frac{df}{dz} = \begin{bmatrix} \frac{1}{\cosh^2 z_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\cosh^2 z_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\cosh^2 z_M} \end{bmatrix}$$

The dimension of $\frac{dz}{dx}$ is $M \times N$

$$\frac{dz}{dx} = A$$

The final derivative is the product of each component.

4 Question 4

4.1 a

$$\begin{aligned} R(f) &= \mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(f(x), y)] \\ &= \mathbb{E}_{(x,y)\sim\mathcal{P}}[\mathbb{1}_{f(x)\neq y}] \\ &= P_{(x,y)\sim\mathcal{P}}(f(x) \neq y) \end{aligned}$$

4.2 b

$$\begin{aligned} P(g(x) \neq Y \mid X = x) &= 1 - P(Y = g(x) \mid X = x) \\ &= 1 - [P(Y = 0, g(x) = 0 \mid X = x) + P(Y = 1, g(x) = 1 \mid X = x)] \\ &= 1 - \left[\mathbb{E}[\mathbb{1}_{Y=1}\mathbb{1}_{g(x)=1} \mid X = x] + \mathbb{E}[\mathbb{1}_{Y=0}\mathbb{1}_{g(x)=0} \mid X = x]\right] \\ &= 1 - \left[\mathbb{1}_{g(x)=1}\mathbb{E}[\mathbb{1}_{Y=1} \mid X = x] + \mathbb{1}_{Y=0}\mathbb{E}[\mathbb{1}_{g(x)=0} \mid X = x]\right] \\ &= 1 - \left[\mathbb{1}_{g(x)=1}P(Y = 1 \mid X = x) + \mathbb{1}_{g(x)=0}P(Y = 0 \mid X = x)\right] \\ &= 1 - \left[\mathbb{1}_{g(x)=1}\eta(x) + \mathbb{1}_{g(x)=0}(1 - \eta(x))\right] \end{aligned}$$

4.3 c

$$\begin{aligned} P(g(x) \neq Y \mid X = x) - P\left(f^*(x) \neq Y \mid X = x\right) &= 1 - \left[\mathbb{1}_{g(x)=1}\eta(x) + \mathbb{1}_{g(x)=0}(1 - \eta(x))\right] \\ &\quad - \left[1 - \left[\mathbb{1}_{f^*(x)=1}\eta(x) + \mathbb{1}_{f^*(x)=0}(1 - \eta(x))\right]\right] \\ \\ &= -\left[\mathbb{1}_{g(x)=1}\eta(x) + \mathbb{1}_{g(x)=0}(1 - \eta(x))\right] \\ &\quad + \left[\mathbb{1}_{f^*(x)=1}\eta(x) + \mathbb{1}_{f^*(x)=0}(1 - \eta(x))\right] \\ \\ &= \eta(x)[\mathbb{1}_{f^*(x)=1} - \mathbb{1}_{g(x)=1}] + (1 - \eta(x))[\mathbb{1}_{f^*(x)=0} - \mathbb{1}_{g(x)=0}] \\ &= \eta(x)[\mathbb{1}_{f^*(x)=1} - \mathbb{1}_{g(x)=1}] + (1 - \eta(x))[\mathbb{1}_{g(x)=1} - \mathbb{1}_{f^*(x)=1}] \\ &= 2\eta(x)\mathbb{1}_{f^*(x)=1} - 2\eta(x)\mathbb{1}_{g(x)=1} + \mathbb{1}_{g(x)=1} - \mathbb{1}_{f^*(x)=1} \\ &= (2\eta(x) - 1)\left(\mathbb{1}_{\{f^*(x)=1\}} - \mathbb{1}_{\{g(x)=1\}}\right) \end{aligned}$$

4.4 d

From previous section we have:

$$P(g(x) \neq Y \mid X = x) - P\left(f^*(x) \neq Y \mid X = x\right) = (2\eta(x) - 1)\left(\mathbb{1}_{\{f^*(x)=1\}} - \mathbb{1}_{\{g(x)=1\}}\right)$$

So $(2\eta(x) - 1)\left(\mathbb{1}_{\{f^*(x)=1\}} - \mathbb{1}_{\{g(x)=1\}}\right) \geq 0$ is equivalent to $P(g(x) \neq Y \mid X = x) \geq P\left(f^*(x) \neq Y \mid X = x\right)$, wh

We also know:

$$f^*(x) = \begin{cases} 1, & \text{if } \eta(x) \geq 1/2 \\ 0, & \text{otherwise} \end{cases}$$

So:
1) If $\eta(x) \geq 1/2$ then, $(2\eta(x) - 1) \geq 0$ and $\left(\mathbb{1}_{\{f^*(x)=1\}} - \mathbb{1}_{\{g(x)=1\}}\right) \geq 0$ (since $f^*(x) = 1$), therefore the final value is non-negative.

2) If $\eta(x) \leq 1/2$ then, $(2\eta(x) - 1) \leq 0$ and $\left(\mathbb{1}_{\{f^*(x)=1\}} - \mathbb{1}_{\{g(x)=1\}}\right) \leq 0$ (since $f^*(x) = 0$), therefore the final value is non-negative.

From above two conditions: $(2\eta(x) - 1)\left(\mathbb{1}_{\{f^*(x)=1\}} - \mathbb{1}_{\{g(x)=1\}}\right) \geq 0$

4.5 e

Since we don't know the true distribution of P, we can't construct the η, and $f^*(x)$ isn't realizable.

5 Question 5

5.1 a

$$risk = \sum_D (y - h(x))^2$$

5.2 b

$$\begin{aligned}
\mathop{\mathbb{E}}_{D\sim p}\left[\text{error}_{LOO}\right] &= \mathop{\mathbb{E}}_{D\sim p}\left[\frac{1}{n}\sum_{i=1}^{n}\ell\left(h_{D\setminus i}\left(x_i\right), y_i\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{D\sim p}\left[\ell\left(h_{D\setminus i}\left(x_i\right), y_i\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{\substack{D'\sim p,\\ (x,y)\sim p}}\left[\ell\left(h_{D'}(x), y\right)\right] \\
&= \mathop{\mathbb{E}}_{\substack{D'\sim p,\\ (x,y)\sim p}}\left[\left(y - h_{D'}(x)\right)^2\right]
\end{aligned}$$

With $n-1$ data points, the LOO error is an unbiased estimator risk of h'_D, therefore, when n is large we have unbiased estimator of risk of h_D.

5.3 c

From Linear Regression:

$$\theta^\star = (X^\top X)^{-1}X^\top y$$

The complexity of multiplying $d \times n$ matrix by $n \times d$ matrix is $O(dnd)$, so we have the following total complexity:

$$O(dnd + d^3 + ddn + dn) = O(d^3 + d^2 n)$$

5.4 d

$$\text{error}_{LOO} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left[\left(\mathbf{X}_{-i}^\top \mathbf{X}_{-i}\right)^{-1}\mathbf{X}_{-i}^\top \mathbf{y}_{-i}\right]^\top x_i\right)^2$$

So the complexity of the above formula will be:

$$O(n(dnd + d^3 + ddn + dn)) = O(n(d^3 + d^2 n)) = O(nd^3 + d^2 n^2)$$

5.5 e

From section d we have:

$$\begin{aligned}
\text{error}_{LOO} &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left[\left(\mathbf{X}_{-i}^\top \mathbf{X}_{-i}\right)^{-1}\mathbf{X}_{-i}^\top \mathbf{y}_{-i}\right]^\top x_i\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{y}_{-i}^\top \mathbf{X}_{-i}\left(\mathbf{X}_{-i}^\top \mathbf{X}_{-i}\right)^{-1}\mathbf{x}_i\right)^2
\end{aligned}$$

From the question and the above formula, we should show that the following statement is true:

$$y_i - \mathbf{y}_{-i}^\top \mathbf{X}_{-i} \left(\mathbf{X}_{-i}^\top \mathbf{X}_{-i}\right)^{-1} \mathbf{x}_i = \frac{y_i - \mathbf{w}^{*\top} \mathbf{x}_i}{1 - \mathbf{x}_i^\top \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{x}_i}$$

So:

$$\begin{aligned}\left(y_i - \mathbf{y}_{-i}^\top \mathbf{X}_{-i} \left(\mathbf{X}_{-i}^\top \mathbf{X}_{-i}\right)^{-1} \mathbf{x}_i\right) \left(1 - \mathbf{x}_i^\top \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{x}_i\right) &= y_i - \mathbf{y}_{-i}^\top \mathbf{X}_{-i} \left(\mathbf{X}_{-i}^\top \mathbf{X}_{-i}\right)^{-1} \mathbf{x}_i \\ &- y_i \mathbf{x}_i^\top \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{x}_i + \mathbf{y}_{-i}^\top \mathbf{X}_{-i} \left(\mathbf{X}_{-i}^\top \mathbf{X}_{-i}\right)^{-1} \mathbf{x}_i \mathbf{x}_i^\top \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{x}_i\end{aligned}$$

Note, we know that $(\mathbf{X}^\top \mathbf{X}) = \mathbf{X}_{-i}^\top \mathbf{X}_{-i} + \mathbf{x}_i \mathbf{x}_i^\top$, so the last term of the above equation will be:

$$\begin{aligned}\mathbf{y}_{-i}^\top \mathbf{X}_{-i} \left(\mathbf{X}_{-i}^\top \mathbf{X}_{-i}\right)^{-1} \mathbf{x}_i \mathbf{x}_i^\top \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{x}_i &= \mathbf{y}_{-i}^\top \mathbf{X}_{-i} \left(\mathbf{X}_{-i}^\top \mathbf{X}_{-i}\right)^{-1} \left(\mathbf{X}^\top \mathbf{X} - \mathbf{X}_{-i}^\top \mathbf{X}_{-i}\right) \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{x}_i \\ &= \mathbf{y}_{-i}^\top \mathbf{X}_{-i} \left(\mathbf{X}_{-i}^\top \mathbf{X}_{-i}\right)^{-1} \mathbf{x}_i - \mathbf{y}_{-i}^\top \mathbf{X}_{-i} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{x}_i\end{aligned}$$

We can now apply this expression, to continue the previous derivation:
Note, we also know that $\mathbf{X} = (y_i \mathbf{x}_i^\top + \mathbf{y}_{-i}^\top \mathbf{X}_{-i})$

$$\begin{aligned}&= y_i - \mathbf{y}_{-i}^\top \mathbf{X}_{-i} \left(\mathbf{X}_{-i}^\top \mathbf{X}_{-i}\right)^{-1} \mathbf{x}_i \\ &- y_i \mathbf{x}_i^\top \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{x}_i + \mathbf{y}_{-i}^\top \mathbf{X}_{-i} \left(\mathbf{X}_{-i}^\top \mathbf{X}_{-i}\right)^{-1} \mathbf{x}_i \mathbf{x}_i^\top \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{x}_i \\ \\ &= y_i - y_i \mathbf{x}_i^\top \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{x}_i - \mathbf{y}_{-i}^\top \mathbf{X}_{-i} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{x}_i \\ &= y_i - \left(y_i \mathbf{x}_i^\top + \mathbf{y}_{-i}^\top \mathbf{X}_{-i}\right) \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{x}_i \\ &= y_i - \left(\mathbf{y}^\top \mathbf{X}\right) \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{x}_i \\ &= y_i - \mathbf{w}^{*\top} \mathbf{x}_i\end{aligned}$$

The complexity of the above expression is $O(d^3 + d^2 n)$, which is efficient than the expression in section d.