IFT-6390 Fundamentals of Machine Learning
Professor: Ioannis Mitliagkas
Students: Abhay Puri (20209505), Saurabh Bodhe (20208545)

Homework 2 - Theoretical part

Solutions

1. **Solution to Q1**

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

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

Using $\mathbb{E}[\epsilon] = 0$
Using $\mathbb{E}[x - \mathbb{E}[x]] = 0$

$$\mathbb{E}[(h_D(x') - y')^2] = \mathbb{E}[(h_D(x') - \mathbb{E}[h_D(x')])^2] + \mathbb{E}[(\mathbb{E}[h_D(x') - (f(x'))^2] + \mathbb{E}[\epsilon^2]$$

$$= Var(h_D(x')) + Bias(h_D(x'))^2 + \mathbb{E}[\epsilon^2]$$
$$= Variance + Bias^2 + \mathbb{E}[\epsilon^2]$$

2. **Solution to Q2**

   (a) The answer is Yes.
   $f(x) = \lfloor x \rfloor \% 2$

   (b) The answer is yes.
   $f(x_1, x_2) = x_1 * x_2$

   (c) Yes, it can be possible.
   $f(x_1, x_2) = \sqrt{x_1^2 + x_2^2}$, $f(y_1, y_2) = \sqrt{y_1^2 + y_2^2}$
   $\phi(x) = \lfloor(\sqrt{x_1^2 + x_2^2})\rfloor \% 2$
   The kernel is :
   K(x, y) = $\phi(x) * \phi(y) = \lfloor(\sqrt{x_1^2 + x_2^2})\rfloor \% 2 * \lfloor(\sqrt{y_1^2 + y_2^2})\rfloor \% 2$

3. **Solutions to Q3**

(a)

$$f\left(x\right)=log\left(x^{4}\right)sin\left(x^{3}\right)$$

$$\text{Applying the Product rule we get}$$

$$\frac{d}{dx}[f(x)g(x)]=f(x)g'(x)+f'(x)g(x)$$

$$\frac{d}{dx}\left(\log(x^{4})sin(x^{3})\right)=log(x^{4})cos(x^{3})(3x^{2})+\frac{1}{x^{4}}(sin(x^{3}))(4x^{3})$$

$$f'(x)=(12x^{2})log(x)cos(x^{3})+\frac{4}{x}sin(x^{3})$$

(b)

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

$$\text{Applying the Product rule we get}$$

$$\frac{d}{dx}[f(x)g(x)]=f(x)g'(x)+f'(x)g(x)$$

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

$$f'(x)=f(x)\left(\frac{-1}{\sigma}(x-\mu)\right)$$

(c)

$$f_{1}\left(x\right)=sin\left(x_{1}\right)cos\left(x_{2}\right),\ x\ \in\ \mathbb{R}^{2}$$

$$f_{2}\left(x,y\right)=x^{T}y\ ,\ x,y\ \in\ \mathbb{R}^{n}$$

$$f_{3}\left(x\right)=xx^{T}\ ,\ x\ \in\ \mathbb{R}^{n}$$

i. f_1

$$\frac{\partial f_{1}}{\partial x_{1}}=cos(x_{1})cos(x_{2})$$

$$\frac{\partial f_{1}}{\partial x_{2}}=-sin(x_{1})sin(x_{2})$$

$$J=\begin{bmatrix}\frac{\partial f_{1}}{\partial x_{1}} & \frac{\partial f_{2}}{\partial x_{2}}\end{bmatrix}=[cos(x_{1})cos(x_{2})-sin(x_{1})sin(x_{2})]\in\ \mathbb{R}^{1\times 2}$$

ii. f_2

$$\frac{\partial f_{2}}{\partial x}=\begin{bmatrix}\frac{\partial f_{2}}{\partial x_{1}} & .....\frac{\partial f_{2}}{\partial x_{n}}\end{bmatrix}=[y_{1}....y_{n}]=y^{T}\in\ \mathbb{R}^{n}$$

$$\frac{\partial f_{2}}{\partial y}=\begin{bmatrix}\frac{\partial f_{2}}{\partial y_{1}} & .....\frac{\partial f_{2}}{\partial y_{n}}\end{bmatrix}=[x_{1}....x_{n}]=x^{T}\in\ \mathbb{R}^{n}$$

$$J=\begin{bmatrix}\frac{\partial f_{2}}{\partial x} & \frac{\partial f_{2}}{\partial y}\end{bmatrix}=[y^{T}x^{T}]\in\ \mathbb{R}^{1\times 2n}$$

iii. f_3

$$xx^T = \begin{bmatrix} x_1x^T \\ x_2x^T \\ \vdots \\ x_nx^T \end{bmatrix} = \begin{bmatrix} xx_1 & xx_2 & \dots xx_n \end{bmatrix} \in \mathbb{R}^{n\times n}$$

$$\frac{\partial f_3}{\partial x_1} = \begin{bmatrix} x^T \\ 0_n^T \\ \vdots \\ 0_n^T \end{bmatrix} + \begin{bmatrix} x & 0_n & \dots 0_n \end{bmatrix} \in \mathbb{R}^{n\times n}$$

$$\frac{\partial f_3}{\partial x_i} = \begin{bmatrix} 0_{(i-1)} \times n \\ x^T \\ \vdots \\ 0_{(n-1+1)} \times n \end{bmatrix} \begin{bmatrix} 0_{(i-1)} \times n & x^T & \dots 0_{(n-i+1)} \times n \end{bmatrix} \in \mathbb{R}^{n\times n}$$

We need to concatenate the results to get the final results.

$$J = \begin{bmatrix} \frac{\partial f_3}{\partial x_1} & \dots \frac{\partial f_3}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{(n\times n)\times n}$$

(d) i.

$$f(z) = exp\left(-\frac{1}{2}z\right)$$
$$z = g(y) = y^T S^{-1} y$$
$$y = h(x) = x - \mu$$

$$\frac{df}{dx} = \frac{\partial f}{\partial z}\frac{\partial g}{\partial y}\frac{\partial h}{\partial x} \in \mathbb{R}^{1\times D}$$

$$\frac{\partial f}{\partial z} = -\frac{1}{2}exp(-\frac{1}{2}z) \in \mathbb{R}^{1\times 1}$$
$$\frac{\partial g}{\partial y} = 2y^T S^{-1} \in \mathbb{R}^{1\times D}$$
$$\frac{\partial h}{\partial x} = I_D \in \mathbb{R}^{D\times D}$$

so that

$$\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = -\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{S}^{-1}$$

ii.

$$f(x) = tr\left(xx^T + \sigma I\right)$$

Here $x \in \mathbb{R}^D$ and $tr\,(A)$ is the trace of A.

We define $X = xx^T$
Trace sums up all the diagonal elements, such that

$$\frac{\partial}{\partial x_j}(tr(X + \sigma^2 I) = \sum_{i=1}^{D} \frac{\partial X_{ii} + \sigma^2}{\partial x_j} = 2x_j$$

for j = 1,.....,D. We can conclude that

$$\frac{\partial}{\partial x_j}(tr(xx^T + \sigma^2 I) = 2x^T \in \mathbb{R}^{1\times D}$$

iii.

$$f = tanh\,(z)$$

Here $f \in \mathbb{R}^M$.

$$z = Ax + b$$

Here $x \in \mathbb{R}^N$, $A \in \mathbb{R}^{M\times N}$, $b \in \mathbb{R}^M$.

Here, tanh is applied to every component of x.

$$\frac{\partial f}{\partial z} = diag(1 - tanh^2(z)) \in \mathbb{R}^{M\times M}$$

$$\frac{\partial z}{\partial x} = \frac{\partial Ax}{\partial x} = A \in \mathbb{R}^{M\times N}$$

$$\frac{\partial f}{\partial x} \in \mathbb{R}^{M\times N}$$

Overall derivative is an $M \times N$ matrix.

4. **Solutions to Q5**

(a) $R_{\text{emp}}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(h\,(x) - y^2\right)$

$R(f) = \mathbb{E}_{p(x,y)}\left(h\,(x) - y^2\right)$

(b)

$$\mathop{\mathbb{E}}_{D\sim p}[\text{error}_{LOO}] = \mathop{\mathbb{E}}_{\substack{D'\sim p\\(x,y)\sim p}}[(y - h_{D'}(x))^2]$$

$$\mathop{\mathbb{E}}_{D\sim p}[\text{error}_{LOO}] = \mathop{\mathbb{E}}_{D'\sim p}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(h_{D\backslash i}(x_i), y_i)\right]$$

$$\mathop{\mathbb{E}}_{D\sim p}[\text{error}_{LOO}] = \frac{1}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{D\sim p}\left[\ell\left(h_{D\backslash i}\,(x_i)\,, y_i\right)\right]$$

$$\mathop{\mathbb{E}}_{D\sim p}[\text{error}_{LOO}] = \frac{1}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{\substack{D'\sim p\\(x,y)\sim p}}[\ell\,(h_{D'}(x), y)]$$

$$\mathbb{E}_{D \sim p}[\text{error}_{LOO}] = \mathbb{E}_{\substack{D' \sim p, \\ (x,y) \sim p}} \left[(y - h_{D'}(x))^2 \right]$$

error_{LOO} is an (almost) unbiased estimator of the risk of $h_{D'}$ because D' contains (n-1) samples. D contains n samples and n is very large.

(c) For n observations and d weights, complexity of Matrix Multiplication of $X^T X$ is $\mathcal{O}(d^2 n)$ and the matrix inverse complexity is $\mathcal{O}(d^3)$.
Total complexity $= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ is $= \mathcal{O}(d^2 n) + \mathcal{O}(d^3)$.

(d)

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

Total complexity $= \mathcal{O}(d^2 n^2) + \mathcal{O}(d^3 n)$.

(e) From the previous question we have the

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

We are putting this value in the given equation we get

$$(y_i - \mathbf{y}_{-i}^\top \mathbf{X}_{-i} \left(\mathbf{X}_{-i}^\top \mathbf{X}_{-i} \right)^{-1} \mathbf{x}_i)^2 = \left(\frac{y_i - \mathbf{w}^{*\top} \mathbf{x}_i}{1 - \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i} \right)^2$$

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

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

Now $\mathbf{X_{-i}}^\top \mathbf{X_{-i}} = \left(\mathbf{X_i}^\top \mathbf{X_i} \right) - \mathbf{x}_i \mathbf{x}_i^T$ or $\mathbf{X}^\top \mathbf{X} = \mathbf{X_{-i}}^\top \mathbf{X_{-i}} + \mathbf{x}_i \mathbf{x}_i^T$
By putting this value we get

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

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

Hence proved.
As we know that combining w^* and $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ the complexity is $\mathcal{O}(d^2n) + \mathcal{O}(d^3)$.

4) $f^*(x) = \begin{cases} 1 & \eta(x) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$ $\eta(x) = P(Y=1 \mid X=x)$

a)

$$R(f) = \mathbb{E}_{(x,y)\sim p}[\,l(f(x),y)\,]$$

$$= \mathbb{E}_{(x,y)\sim p}[\mathbb{1}(f(x) \neq y)]$$

$$= 0 * P_{(x,y)\sim p}(f(x) = y) + 1 * P_{(x,y)\sim p}(f(x) \neq y)$$

$$= P_{(x,y)\sim p}(f(x) \neq y)$$

---

b) $P(g(x) = y \mid X = x) = P(g(x) = 1, y = 1 \mid X = x) + P(g(x) = 0, y = 0 \mid X = x)$

$= P(g(x) = 1 \mid X = x) * P(y = 1 \mid X = x) + P(g(x) = 0 \mid X = x) * P(y = 0 \mid X = x)$ [Since y & g are independent]

$P(g(x) = 1 \mid X = x) = \mathbb{1}(g(x) = 1)$

$P(g(x) = 0 \mid X = x) = \mathbb{1}(g(x) = 0)$

[Since $g(x)$ is deterministic given x]

we know, $\eta(x) = P(Y = 1 \mid X = x)$

$\therefore P(Y = 0 \mid X = x) = 1 - \eta(x)$

$$P(g(x) = y \mid X = x)$$
$$= \mathbb{1}(g(x)=1)\,\eta(x) + \mathbb{1}(g(x)=0)(1-\eta(x))$$

$$P(g(x) \neq y \mid X = x) = 1 - P(g(x) = y \mid X = x)$$
$$= 1 - \left[\mathbb{1}(g(x)=1)\,\eta(x) + \mathbb{1}(g(x)=0)(1-\eta(x))\right]$$

---

c)

$$P(g(x) \neq Y \mid X = x) - P(f^*(x) \neq Y \mid X = x)$$
$$= \left[\mathbb{1}(f^*(x)=1)\,\eta(x) + \mathbb{1}(f^*(x)=0)(1-\eta(x))\right]$$
$$- \left[\mathbb{1}(g(x)=1)\,\eta(x) + \mathbb{1}(g(x)=0)(1-\eta(x))\right]$$

we know, $\mathbb{1}(f^*(x)=0) = 1 - \mathbb{1}(f^*(x)=1)$

$\mathbb{1}(g(x)=0) = 1 - \mathbb{1}(g(x)=1)$

$$= \left[\mathbb{1}(f^*(x)=1)\,\eta(x) + (1 - \mathbb{1}(f^*(x)=1))(1-\eta(x))\right]$$
$$- \left[\mathbb{1}(g(x)=1)\,\eta(x) + (1 - \mathbb{1}(g(x)=1))(1-\eta(x))\right]$$
$$= \eta(x)\left(\mathbb{1}(f^*(x)=1) - \mathbb{1}(g(x)=1)\right) +$$
$$(1-\eta(x))\left(\mathbb{1}(g(x)=1) - \mathbb{1}(f^*(x)=1)\right)$$
$$= (2\eta(x)-1)\left(\mathbb{1}(f^*(x)=1) - \mathbb{1}(g(x)=1)\right)$$

d) i) $\underline{\text{if } \eta(x) \geq 1/2} \Rightarrow (2\eta(x)-1) \geq 0$

$\quad \& \ \mathbb{1}(f^*(x)=1) = 1$

$\Rightarrow \mathbb{1}(f^*(x)=1) - \mathbb{1}(g(x)=1) = 1 - \mathbb{1}(g(x)=1)$

≥ 0

$\therefore (2\eta(x)-1)(\mathbb{1}(f^*(x)=1) - \mathbb{1}(g(x)=1)) \geq 0$

ii) $\underline{\text{if } \eta(x) < 1/2} \Rightarrow (2\eta(x)-1) < 0$

$\quad \& \ \mathbb{1}(f^*(x)=1) = 0$

$\Rightarrow \mathbb{1}(f^*(x)=1) - \mathbb{1}(g(x)=1) = 0 - \mathbb{1}(g(x)=1)$

≤ 0

$\therefore (2\eta(x)-1)(\mathbb{1}(f^*(x)=1) - \mathbb{1}(g(x)=1)) \geq 0$

Hence for both cases we get the desired result

e) T.P $R(g) \geq R(f^*)$

from (c) & (d),

$P(g(x) \neq Y | X=x) - P(f^*(x) \neq Y | X=x)$

≥ 0

$\Rightarrow P(g(x) \neq Y | X=x) \geq P(f^*(x) \neq Y | X=x)$

$$P(g(x) \neq Y) = \int_x P(g(x) \neq Y \mid X = x)\, p(X = x)\, dx$$

$$\& \; P(f^*(x) \neq Y) = \int_x P(f^*(x) \neq Y \mid X = x)\, p(X = x)\, dx$$

we have,

$$P(g(x) \neq Y \mid X = x) \geq P(f^*(x) \neq Y \mid X = x)$$

$$\Rightarrow \int_x P(g(x) \neq Y \mid X = x)\, p(X = x)\, dx \geq \int_x P(f^*(x) \neq Y \mid X = x)\, p(X = x)\, dx$$

$$\Rightarrow \quad P(g(x) \neq Y) \geq P(f^*(x) \neq Y)$$

$$\Rightarrow \quad R(g) \geq R(f^*)$$
