DSA5105 Homework 2 Wang Jiangyi

Q1. [Multiple choices question]

(a) False

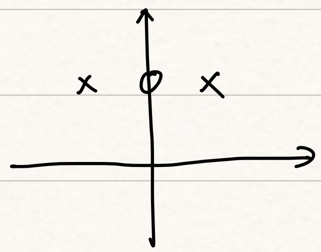
(b) False

Reason: $\nabla J(w) = 0 \Leftrightarrow \underline{\Phi^T \Phi w = \Phi^T y}$

$\longrightarrow$ when $\Phi^T\Phi$ is not invertible, then we may have infinitely many $\hat{w}$

(c) False

Reason:



(d) False, should use validation set to tune the hyper-params and use test set to evaluate the model

(e) True

Q2. [linear basis regression]

$\Phi := \begin{bmatrix} \phi(x_1)^T \\ \vdots \\ \phi(x_N)^T \end{bmatrix} \in \mathbb{R}^{N\times M}$ $\quad w := \begin{pmatrix} w_0 \\ \vdots \\ w_{M-1} \end{pmatrix} \in \mathbb{R}^M$

(a) $\hat{W} = \underset{w\in\mathbb{R}^M}{\operatorname{argmin}} \ R_{emp}(w)$

$= \underset{w\in\mathbb{R}^M}{\operatorname{argmin}} \ \frac{1}{2N}(\Phi w - y)^T(\Phi w - y)$

$= \underset{w\in\mathbb{R}^M}{\operatorname{argmin}} \ \underline{w^T\Phi^T\Phi w - 2w^T\Phi^T y := J(w)}$

$\Rightarrow \hat{w} = \underset{w\in\mathbb{R}^M}{\operatorname{argmin}} \ J(w)$, which is an unconstrained optimization problem

① since $\nabla_w J(w) = 2\Phi^T\Phi w - 2\Phi^T y$

$\nabla_w^2 J(w) = 2\Phi^T\Phi \succeq 0$ (PSD) then $J(w)$ is convex (function) w.r.t $\underline{w}$

② for convex & differentiable function $J(w)$,

$$\hat{w} = \underset{w \in \mathbb{R}^M}{\operatorname{argmin}} J(w) \Leftrightarrow \nabla_w J(\hat{w}) = 0$$

$$\Leftrightarrow 2\Phi^T\Phi\hat{w} - 2\Phi^T y = 0$$

$$\Leftrightarrow \hat{w} = (\Phi^T\Phi)^{-1}\Phi^T y$$

(given that $\Phi^T\Phi$ is invertible)

(b) [Weighted Least Square]

$$\hat{w}_{WLS} = \underset{w \in \mathbb{R}^M}{\operatorname{argmin}} \, R_{emp}(w, a)$$

$$= \underset{w \in \mathbb{R}^M}{\operatorname{argmin}} \sum a_i (w^T\phi(x_i) - y_i)^2 \qquad D(\Phi w - y) \cdot D$$

$$= \underset{w \in \mathbb{R}^M}{\operatorname{argmin}} [D(\Phi w - y)]^T [D(\Phi w - y)] \qquad \underline{D = \operatorname{diag}\{\sqrt{a_1}, \ldots, \sqrt{a_N}\}}$$

$$= \underset{w \in \mathbb{R}^M}{\operatorname{argmin}} (\Phi w - y)^T W (\Phi w - y) \qquad \underline{W = \operatorname{diag}\{a_1, \ldots, a_N\}}$$

$$= \underset{w \in \mathbb{R}^M}{\operatorname{argmin}} \underline{w^T\Phi^T W \Phi w - 2w^T\Phi^T W y} := J(w)$$

① since $\nabla_w J(w) = 2\Phi^T W \Phi w - 2\Phi^T W y$

$$\nabla_w^2 J(w) = 2\Phi^T W \Phi \succeq 0 \qquad \underline{(\text{since } a_i > 0 \text{ for } i = 1, 2, \ldots, N)}$$

then $J(w)$ is convex (function) w.r.t $\underline{w}$

② for convex & differentiable function $J(w)$,

$$\hat{w}_{WLS} = \underset{w \in \mathbb{R}^M}{\operatorname{argmin}} J(w) \Leftrightarrow \nabla_w J(\hat{w}_{WLS}) = 0$$

$$\Leftrightarrow 2\Phi^T W \Phi \hat{w}_{WLS} - 2\Phi^T W y = 0$$

$$\Leftrightarrow \hat{w}_{WLS} = (\Phi^T W \Phi)^{-1}\Phi^T W y$$

(given that $\Phi^T\Phi$ is invertible)

(c) [Application Scenario for WLS]

① When we have prior that some data points in dataset $\mathcal{D}$ are outliers, we can assign low weights (small a_i) to them

② Linear regression can be viewed as the Gaussian Model that

$y \sim \text{Gaussian}(w^T\phi(x), \Sigma)$ where $\Sigma = \sigma^2 I$. $\boxed{+\ \text{MLE}}$

↳ isotropic Gaussian

Weighted Least Square can be viewed as the Gaussian Model that

$y \sim \text{Gaussian}(w^T\phi(x), \Sigma)$ where $\Sigma = \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_N^2 \end{pmatrix}$

$\left(a_i = \frac{1}{\sigma_i^2}\right)$ $\boxed{+\text{MLE}}$

→ from this perspective, we can take deviance of each data point into consideration:

↳ our confidence of each data

○ for those data points we have more confidence (due to the observation error), we can assign small σ_i^2 (large a_i) to them;

○ for those data points are more likely to be the noise, we can assign large σ_i^2 (small a_i) to them.

Q3. [nearest-neighbour] → kNN with $k=1$

a) $$i(x) = \underset{j\in[N]}{\operatorname{argmin}} \|x - x_j\|_2^2 = \underset{j\in[N]}{\operatorname{argmin}} \langle x - x_j, x - x_j\rangle$$
$$= \underset{j\in[N]}{\operatorname{argmin}} \langle x_j, x_j\rangle - 2\langle x, x_j\rangle$$
$$= \underset{j\in[N]}{\operatorname{argmin}} \langle x_j - 2x, x_j\rangle$$

b) feature map $\phi: x\in\mathbb{R}^d \longmapsto \phi(x)\in\mathbb{R}^M$, denote $K(x,y) = \langle\phi(x), \phi(y)\rangle$

$f(x) = y_{i(\phi(x))}$ such that $i(\phi(x)) = \underset{j\in[N]}{\operatorname{argmin}} \|\phi(x) - \phi(x_j)\|_2^2 = \underset{j\in[N]}{\operatorname{argmin}} K(x_j - 2x, x_j)$

Q4. [LR]

(a) we have $\underline{\mathbb{E}[\varepsilon_i] = 0}$ $\underline{\mathbb{E}[\varepsilon_i^2] = \mathrm{Var}[\varepsilon_i] + \mathbb{E}[\varepsilon_i]^2 = \sigma^2}$

$$\min_{w_0, w_1} \mathbb{E}\left[\frac{1}{2N} \sum_{i=1}^{N} [w_0 + w_1 x_i - y_i + w_1 \varepsilon_i]^2 \right]$$

$$\Leftrightarrow \min_{w_0, w_1} \mathbb{E}\left[\frac{1}{2N} \sum_{i=1}^{N} [(w_0 + w_1 x_i - y_i)^2 + w_1^2 \varepsilon_i^2 + w_1 \varepsilon_i \cdot C_i] \right]$$

where $\underline{C_i = 2(w_0 + w_1 x_i - y_i)}$

$$\Leftrightarrow \min_{w_0, w_1} \frac{1}{2N} \sum_{i=1}^{N} \left[(w_0 + w_1 x_i - y_i)^2 + w_1^2 \underbrace{\mathbb{E}[\varepsilon_i^2]}_{\sigma^2} + \underbrace{\mathbb{E}[\varepsilon_i]}_{0} w_1 C_i \right]$$

$$\Leftrightarrow \min_{w_0, w_1} \frac{1}{2N} \sum_{i=1}^{N} (w_0 + w_1 x_i - y_i)^2 + \frac{w_1^2}{2} \sigma^2$$

$$\Leftrightarrow \min_{w_0, w_1} \underbrace{\frac{1}{2} \| Xw - y \|_2^2}_{\text{OLS}} + \underbrace{\frac{N\sigma^2}{2} w_1^2}_{\text{regularization}} \text{ on } w_1$$

$$w = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

$$\longrightarrow \min_{w} \frac{1}{2} \left(w^T X^T X w - 2 w^T X^T y \right) + \frac{1}{2} w^T K w \qquad K = \begin{pmatrix} 0 & \\ & N\sigma^2 \end{pmatrix}$$

$$\Leftrightarrow \min_{w} \frac{1}{2} w^T (X^T X + K) w - w^T X^T y := J(w)$$

$$\nabla J(w) = (X^T X + K) w - X^T y$$

$$\nabla^2 J(w) = X^T X + K \succeq 0$$

$$\Longrightarrow \hat{w} \in \underset{w}{\mathrm{argmin}}\, J(w) \Leftrightarrow \nabla J(\hat{w}) = 0$$

$$\Leftrightarrow \underline{\hat{w} = (X^T X + K)^{-1} X^T y}$$

$\uparrow$

if $(X^T X + K)$ is invertible

(b) If $\mathbb{E}[\varepsilon_i] = b \neq 0$, then $\mathbb{E}[\varepsilon_i^2] = \sigma^2 + b^2$

$$\min_{w_0, w_1} \frac{1}{2N} \sum_{i=1}^{N} \left[(w_0 + w_1 x_i - y_i)^2 + w_1^2 \mathbb{E}[\varepsilon_i^2] + \mathbb{E}[\varepsilon_i] w_1 C_i \right]$$

where $C_i = 2(w_0 + w_1 x_i - y_i)$

$$\Leftrightarrow \min_{w} \frac{1}{2N}(Xw-y)^T(Xw-y) + \frac{1}{2}w_1^2(\sigma^2+b^2) + \frac{1}{N}\cdot b w_1 \sum_{i=1}^{N}(w_0+w_1x_i-y_i)$$

$$\frac{1}{2N}(w^TX^TXw - 2w^TX^Ty) + \frac{1}{2N}w^TKw + \frac{2b}{2N}\left[w^T T_1 w - w^T T_2^T y\right]$$

$$\boxed{K = \begin{pmatrix} 0 & \\ & N(\sigma^2+b^2)\end{pmatrix}} \qquad \boxed{w^T\begin{pmatrix}0\\1\end{pmatrix}\mathbb{1}^T(Xw-y)}$$

$$= w^T\begin{pmatrix}0 & \cdots & 0\\ 1 & \cdots & 1\end{pmatrix}Xw - w^T\begin{pmatrix}0 & \cdots & 0\\ 1 & \cdots & 1\end{pmatrix}y$$

$\mathbb{R}^{2\times N}$

$$\Leftrightarrow \min_{w} \frac{1}{2}w^TX^TXw + \frac{1}{2}w^TKw + \frac{2b}{2}w^TT_1w - w^TX^Ty - \frac{2b}{2}w^TT_2^Ty := J(w)$$

$$\boxed{\begin{aligned} T_1 &= \begin{pmatrix} 0 & 0\\ N & \sum x_i\end{pmatrix} \in \mathbb{R}^{2\times 2}\\ T_2 &= \begin{pmatrix} 0 & 1\\ \vdots & \vdots\\ 0 & 1\end{pmatrix} \in \mathbb{R}^{N\times 2}\end{aligned}}$$

$$\begin{cases} \nabla J(w) = (X^TX + K + 2bT_1)w - (X+bT_2)^Ty \\ \nabla^2 J(w) = X^TX + K + 2bT_1 \end{cases}$$

Suppose $\nabla^2 J(w) \succeq 0 \Rightarrow J(w)$ is convex on w

$\Rightarrow \hat{w} = \operatorname{argmin} J(w) \Leftrightarrow \nabla J(\hat{w}) = 0$

$$\Leftrightarrow \hat{w} = (X^TX + K + 2bT_1)^{-1}(X+bT_2)^Ty$$

$\boxed{\text{Conclusion}}$: Minimization Problem changed. But we can still achieve the closed-form solution under some assumption.

Q5. [RBF kernel]

$$k(x,y) = \exp\left(-\frac{1}{2s^2}\|x-y\|_2^2\right) \qquad s>0$$

Prove that $k(\cdot,\cdot)$ is a valid kernel function.

$\boxed{\text{Pf}}$: $\|x-y\|_2^2 = \|x\|_2^2 - 2x^Ty + \|y\|_2^2$

$$\Rightarrow k(x,y) = \exp\left(-\frac{1}{2s^2}\|x\|_2^2\right)\exp\left(\frac{1}{s^2}x^Ty\right)\exp\left(-\frac{1}{2s^2}\|y\|_2^2\right)$$

From Taylor Expansion,

$$\Rightarrow \exp\left(\frac{1}{s^2}x^Ty\right) = \sum_{n=0}^{\infty} \frac{1}{n!}\left(\frac{1}{s^2}x^Ty\right)^n$$

Then, we can show that, RBF kernel $K(\cdot,\cdot)$ is valid.

Reason can be shown as follows:

① $K(x,y) = x^Ty \rightarrow$ linear kernel, which is valid kernel

② use scaling property $\Rightarrow k(x,y) = \frac{1}{s^2}x^Ty$ is valid kernel

③ use product property $\Rightarrow K_n(x,y) = \left(\frac{1}{s^2}x^Ty\right)^n$ is valid kernels

④ use scaling property $\Rightarrow k_n(x,y) = \frac{1}{n!}\left(\frac{1}{s^2}x^Ty\right)^n$ is valid kernels

⑤ use addition property $\Rightarrow K_n(x,y) = \sum_{k=0}^{n} \frac{1}{n!}\left(\frac{1}{s^2}x^Ty\right)^n$ is valid kernels

⑥ use limit property $\Rightarrow K(x,y) = \lim_{n\to\infty} K_n(x,y)$

$$= \lim_{n\to\infty} \sum_{k=0}^{\infty} \frac{1}{n!}\left(\frac{1}{s^2}x^Ty\right)^n$$

$$= \exp\left(\frac{1}{s^2}x^Ty\right)$$ is valid kernel

⑦ use normalization property

$$\Rightarrow k(x,y) = \exp\left(-\frac{1}{2s^2}\|x\|_2^2\right)\underline{\exp\left(\frac{1}{s^2}x^Ty\right)}\exp\left(-\frac{1}{2s^2}\|y\|_2^2\right)$$

$$= \exp\left[-\frac{1}{2s^2}\left(\|x\|_2^2 - 2x^Ty + \|y\|_2^2\right)\right]$$

$$= \exp\left(-\frac{1}{2s^2}\|x-y\|_2^2\right)$$ is valid kernel

That is, RBF kernel is a valid kernel!

#