机器学习导论
作业一

141110091, 吴璐欢, lhwunju@outlook.com

2018 年 3 月 27 日

1 [25pts] Basic Probability and Statistics

随机变量 X 的概率密度函数如下，

$$f_X(x) = \begin{cases} \frac{1}{4} & 0 < x < 1; \\ \frac{3}{8} & 3 < x < 5; \\ 0 & \text{otherwise}. \end{cases} \tag{1.1}$$

(1) **[5pts]** 请计算随机变量 X 的累积分布函数 $F_X(x)$;

(2) **[10pts]** 随机变量 Y 定义为 $Y = 1/X$，求随机变量 Y 对应的概率密度函数 $f_Y(y)$；

(3) **[10pts]** 试证明，对于非负随机变量 Z，如下两种计算期望的公式是等价的。

$$\mathbb{E}[Z] = \int_{z=0}^{\infty} z f(z)\, \mathrm{d}z. \tag{1.2}$$

$$\mathbb{E}[Z] = \int_{z=0}^{\infty} \Pr[Z \geq z]\, \mathrm{d}z. \tag{1.3}$$

同时，请分别利用上述两种期望公式计算随机变量 X 和 Y 的期望，验证你的结论

Solution. *(1) By definition of cumulative density function, $F_X(x) = \int_{-\infty}^{x} f_X(x)$. After calculation we have :*

$$F_X(x) = \begin{cases} 0 & x \leq 0; \\ \frac{1}{4}x & 0 < x < 1; \\ \frac{1}{4} & 1 \leq x \leq 3; \\ \frac{3}{8}x - \frac{7}{8} & 3 < x \leq 5; \\ 1 & x \geq 5. \end{cases} \tag{1.4}$$

(2) Denote the CDF of Y by $F_Y(y)$. Then we have:

$$F_Y(y) = P(Y < y) = P(\frac{1}{X} < y) = P(X > \frac{1}{y}) = 1 - F_X(\frac{1}{y}) \quad \textit{if } y > 0 \tag{1.5}$$

$$F_Y(y) = 0 \quad \textit{if } y \le 0 \tag{1.6}$$

For $y > 0$, we can substitute (1.4) into (1.5). After calculation we obatin:

$$F_Y(y) = \begin{cases} 0 & y \le \frac{1}{5}; \\ \frac{15}{8} - \frac{3}{8y} & \frac{1}{5} < y < \frac{1}{3}; \\ \frac{3}{4} & \frac{1}{3} \le y < 1; \\ 1 - \frac{1}{4y} & y > 1. \end{cases} \tag{1.7}$$

Then, by differencing (1.7), we obatin the probability density function of Y:

$$f_Y(y) = \begin{cases} 0 & y \le \frac{1}{5}; \\ \frac{3}{8y^2} & \frac{1}{5} < y < \frac{1}{3}; \\ 0 & \frac{1}{3} \le y < 1; \\ \frac{1}{4y^2} & y > 1. \end{cases}$$

(3) *(i) Proof*

$$\begin{aligned} \mathbb{E}[Z] &= \int_{z=0}^{\infty} z f(z)\,\mathrm{d}z \\ &= \int_0^{\infty} \big[\int_0^{\infty} \mathbb{1}_{(0,z)}(t)\,\mathrm{d}t\big] f(z)\,\mathrm{d}z \\ &= \int_0^{\infty} \big[\int_0^{\infty} \mathbb{1}_{(0,z)}(t) f(z)\,\mathrm{d}\mathrm{d}t\big]\,\mathrm{d}z \\ &= \int_0^{\infty} P_r\big[X \ge t\big]\,\mathrm{d}t \\ &= \int_0^{\infty} Pr\big[X \ge x\big]\,\mathrm{d}x \end{aligned}$$

which gives us, (1.2) $\iff$ (1.3). □

(ii) Validation

- *Calculating the expectations according to (1.2):*

$$\begin{aligned} \mathbb{E}[X] &= \int_0^1 \frac{1}{4} x\,\mathrm{d}x + \int_3^5 \frac{3}{8} x\,\mathrm{d}x \\ &= \frac{1}{8} x^2 \Big|_0^1 + \frac{3}{16} x^2 \Big|_3^5 \\ &= \frac{25}{8} \end{aligned}$$

$$\begin{aligned} \mathbb{E}[Y] &= \int_{\frac{1}{5}}^{\frac{1}{3}} \frac{3}{8y^2} y\,\mathrm{d}y + \int_1^{\infty} \frac{1}{4y^2} y\,\mathrm{d}y \\ &= \frac{3}{8} \ln y \Big|_{\frac{1}{5}}^{\frac{1}{3}} + \frac{1}{4} \ln y \Big|_1^{\infty} \\ &= \infty \end{aligned}$$

- *Calculating the expectations according to (1.3), we have:*

$$
\begin{aligned}
\mathbb{E}[X] &= \int_0^1 (1-\frac{1}{4}x)\,\mathrm{d}x + \int_1^3 (1-\frac{1}{4})\,\mathrm{d}x + \int_3^5 (1-\frac{3}{8}x+\frac{7}{8})\,\mathrm{d}x \\
&= (x-\frac{1}{8}x^2)\Big|_0^1 + \frac{3}{4}x\Big|_1^3 + (\frac{15}{8}x-\frac{3}{16}x^2)\Big|_3^5 \\
&= \frac{25}{8}
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[Y] &= \int_{\frac{1}{5}}^{\frac{1}{3}} (1-\frac{15}{8}+\frac{3}{8y})\,\mathrm{d}y + \int_{\frac{1}{3}}^{1} (1-\frac{3}{4})\,\mathrm{d}y + \int_1^{\infty} (1-1+\frac{1}{4y})\,\mathrm{d}y \\
&= \infty
\end{aligned}
$$

From the calculations above, the equivalence of (1.2) and (1.3) is validated for cases of X and Y.

2 [20pts] Strong Convexity

通过课本附录章节的学习，我们了解到凸性 (convexity) 对于机器学习的优化问题来说是非常良好的性质。下面，我们将引入比凸性还要好的性质——强凸性 (strong convexity)。

定义 1 (强凸性). 记函数 $f:\mathcal{K}\to\mathbb{R}$，如果对于任意 $\mathbf{x},\mathbf{y}\in\mathcal{K}$ 及任意 $\alpha\in[0,1]$，有以下命题成立

$$f((1-\alpha)\mathbf{x}+\alpha\mathbf{y})\le(1-\alpha)f(\mathbf{x})+\alpha f(\mathbf{y})-\frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{x}-\mathbf{y}\|^2. \tag{2.1}$$

则我们称函数 f 为关于范数 $\|\cdot\|$ 的 λ-强凸函数。

请证明，在函数 f 可微的情况下，式 (2.1)与下式 (2.2)等价，

$$f(\mathbf{y})\ge f(\mathbf{x})+\nabla f(\mathbf{x})^{\mathrm{T}}(\mathbf{y}-\mathbf{x})+\frac{\lambda}{2}\|\mathbf{y}-\mathbf{x}\|^2. \tag{2.2}$$

Proof. 此处用于写证明 *(中英文均可)*

(1) (2.1) ⇒ (2.2):

$$\begin{aligned}
(2.1) &\xRightarrow{\textit{move } f(x) \textit{ to the left}} & f((1-\alpha)\mathbf{x}+\alpha\mathbf{y})\le\alpha\big[f(\mathbf{x})-f(\mathbf{y})\big]-\frac{\lambda}{2}\alpha(1-\alpha)\left\|\mathbf{x}-\mathbf{y}\right\|^2\\
&\xRightarrow{\textit{f is differentiable}} & \nabla f(\mathbf{x})^{\mathrm{T}}\alpha(\mathbf{y}-\mathbf{x})+o(\alpha\left\|x-y\right\|)\le\alpha\big[f(x)-f(y)\big]-\frac{\lambda}{2}\alpha(1-\alpha)\left\|x-y\right\|^2\\
&\xRightarrow{\textit{both sides divided by } \alpha} & \nabla f(\mathbf{x})^{\mathrm{T}}(\mathbf{y}-\mathbf{x})+o(\left\|x-y\right\|)\le f(x)-f(y)-\frac{\lambda}{2}(1-\alpha)\left\|x-y\right\|^2\\
&\xRightarrow{\textit{let } \alpha \textit{ goes to } 0} & \nabla f(\mathbf{x})^{\mathrm{T}}(\mathbf{y}-\mathbf{x})\le f(x)-f(y)-\frac{\lambda}{2}\left\|x-y\right\|^2\\
&\xRightarrow{\textit{move items}} (2.2)
\end{aligned}$$

(2) (2.2) ⇒ (2.1):

Let $\mathbf{t}=1-\alpha)\mathbf{x}+\alpha\mathbf{y}$. *Note that:* $y-t=(1-\alpha)(\mathbf{y}-\mathbf{x})$ $t-x=\alpha(\mathbf{y}-\mathbf{x})$.

Then using (2.2), we have:

$$\begin{aligned}
f(\mathbf{y})&\ge f(\mathbf{t})+\nabla f(\mathbf{t})^{\mathrm{T}}(\mathbf{y}-\mathbf{t})+\frac{\lambda}{2}\|\mathbf{y}-\mathbf{t}\|^2\\
&\ge f(\mathbf{t})+\nabla f(\mathbf{t})^{\mathrm{T}}(1-\alpha)(\mathbf{y}-\mathbf{t})+\frac{\lambda}{2}(1-\alpha)^2\|\mathbf{y}-\mathbf{t}\|^2
\end{aligned}$$

$$\Rightarrow f(\mathbf{t})\le f(\mathbf{y})-\nabla f(\mathbf{t})^{\mathrm{T}}(1-\alpha)(\mathbf{y}-\mathbf{t})-\frac{\lambda}{2}(1-\alpha)^2\|\mathbf{y}-\mathbf{t}\|^2 \tag{2.3}$$

Similarily, it could be derived that:

$$f(\mathbf{t})\le f(\mathbf{x})+\nabla f(\mathbf{t})^{\mathrm{T}}\alpha(\mathbf{y}-\mathbf{t})-\frac{\lambda}{2}\alpha^2\|\mathbf{y}-\mathbf{t}\|^2 \tag{2.4}$$

Then,

$$\alpha*(2.3)+(1-\alpha)*(2.4)\Rightarrow(2.1)$$

□

3 [20pts] Doubly Stochastic Matrix

随机矩阵 (stochastic matrix) 和双随机矩阵 (doubly stochastic matrix) 在机器学习中经常出现，尤其是在有限马尔科夫过程理论中，也经常出现在于运筹学、经济学、交通运输等不同领域的建模中。下面给出定义，

定义 2 (随机矩阵)**.** 设矩阵 $\mathbf{X} = [x_{ij}] \in \mathbb{R}^{d \times d}$ 是非负矩阵，如果 $\mathbf{X}$ 满足

$$\sum_{j=1}^{d} x_{ij} = 1, \quad i = 1, 2, \cdots, d. \tag{3.1}$$

则称矩阵 $\mathbf{X}$ 为随机矩阵 *(stochastic matrix)*。如果 $\mathbf{X}$ 还满足

$$\sum_{i=1}^{d} x_{ij} = 1, \quad j = 1, 2, \cdots, d. \tag{3.2}$$

则称矩阵 $\mathbf{X}$ 为双随机矩阵 *(doubly stochastic matrix)*。

对于双随机矩阵 $\mathbf{X} \in \mathbb{R}^{d \times d}$，试证明

(1) **[10pts]** 矩阵 $\mathbf{X}$ 的信息熵 (entropy)满足 $H(\mathbf{X}) \leq d \log d$.

(2) **[10pts]** 矩阵 $\mathbf{X}$ 的谱半径 (spectral radius)$\rho(\mathbf{X})$ 等于 1，且是 $\mathbf{X}$ 的特征值;(提示：你可能会需要Perron–Frobenius 定理，可以基于此进行证明。)

Proof. *(1) Consider* $\mathbf{X}$ *as a Markov process, then according to Section "Data as a Markov process" in Entropy-wikipedia page, the entropy of* $\mathbf{X}$ *is defined by*

$$H(\mathbf{X}) = -\sum_{i=1}^{d} x_i \sum_{j=1}^{d} x_{ij} log(x_{ij}),$$

where $x_i \triangleq \sum_{j=1}^{d} x_{ij}$.

Therefore, substituting (3.2),

$$H(\mathbf{X}) = -\sum_{i=1}^{d} \sum_{j=1}^{d} x_{ij} \log(x_{ij})$$

Let

$$f(x) \triangleq -x \log(x),$$

then

$$f'(x) = -\log(x) - 1,$$
$$f''(x) = -\frac{1}{x} < 0 \quad for\ 0 \leq x \leq 1$$
$$f"(x) = -\infty < 0 \quad for\ x = 0.$$

Therefore, $f(x)$ *is a convex-upward function. So we could apply Jensen's inequality:*

$$\begin{aligned}
\frac{\sum_{j=1}^{d} f(x_{ij})}{d} &\leq f(\frac{\sum_{j=1}^{d} x_{ij}}{d}) = f(\frac{1}{d}) & \forall i = 1, \cdots, d \\
\Rightarrow \qquad \frac{\sum_{j=1}^{d} -x_{ij}\log(x_{ij})}{d} &\leq -\frac{1}{d}\log(\frac{1}{d}) & \forall i = 1, \cdots, d \qquad (3.3) \\
\Rightarrow \qquad \sum_{j=1}^{d} -x_{ij}\log(x_{ij}) &\leq \log(d) & \forall i = 1, \cdots, d
\end{aligned}$$

The equality holds if and only if all x_{ij} are equal.
Or, according to Principle of Maximum Entropy, the maximum entropy discrete probability distribution is the uniform distribution, we could also obtain the inequality (3.3).
Therefore,

$$H(\mathbf{X}) = \sum_{i=1}^{d}\sum_{j=1}^{d} -x_{ij}\log(x_{ij}) \leq \sum_{i=1}^{d}\log(d) = d\log(d)$$

□

(2) Since $\mathbf{X}$ is a positive matrix, then by Perron-Frobenius Theorem, we have

$$1 = \min_{i}\sum_{j=1}^{d} x_{ij} \leq r \leq \max_{j}\sum_{i=1}^{d} x_{ij} = 1,$$

where r is the Perron–Frobenius eigenvalue, and also the spectral radius of $\mathbf{X}$. □

4 [15pts] Hypothesis Testing

在数据集 D_1, D_2, D_3, D_4, D_5 运行了 A, B, C, D, E 五种算法，算法比较序值表如表1所示:

表 1: 算法比较序值表

数据集	算法 A	算法 B	算法 C	算法 D	算法 E
D_1	4	3	5	2	1
D_2	3	5	2	1	4
D_3	4	5	3	1	2
D_4	5	2	4	1	3
D_5	3	5	2	1	4

使用 Friedman 检验 ($\alpha = 0.05$) 判断这些算法是否性能都相同。若不相同，进行 Nemenyi 后续检验 ($\alpha = 0.05$)，并说明性能最好的算法与哪些算法有显著差别。

Solution. *We can turn the table into a corresponding matrix* $R = (r_{ij})_{5\times5}$.
According to the method of Friedman test, we can compute:

$$\bar{r}_{\cdot j} = \sum_{i=1}^{5} r_{ij}$$

$$\Rightarrow \quad \bar{r}_{\cdot 1} = 3.8, \bar{r}_{\cdot 2} = 4, \bar{r}_{\cdot 3} = 3.2, \bar{r}_{\cdot 4} = 1.2, \bar{r}_{\cdot 5} = 2.8$$

$$\bar{r} = \frac{1}{5\times5}\sum_{i=1}^{5}\sum_{j=1}^{5} r_{ij} = \frac{1}{5}\sum_{j=1}^{5}\bar{r}_{\cdot j} = 3$$

$$SS_t = n\sum_{j=0}^{5}(\bar{r}_{\cdot j} - \bar{r})^2 = 24.8$$

$$SS_e = \frac{1}{5\times(5-1)}\sum_{i=1}^{5}\sum_{j=1}^{5}(r_{ij} - \bar{r})^2 = 2.5$$

$$Q = \frac{SS_t}{SS_e} = 9.92$$

Let $\tau_{\chi^2} = Q = 9.92$,

$$\tau_F = \frac{(n-1)\tau_{\chi^2}}{n(k-1) - \tau_{\chi^2}} = 3.9365.$$

For $\alpha = 0.05$, $n = k = 5$, *F's critical value is 3.007. Since* $3.9365 > 3.007$, *the null hypothesis is rejected, which states that the performance of given algorithms are similar.*
Then, we perform the Nemenyi post-hoc test.

$$CD = q_\alpha\sqrt{\frac{k(k+1)}{6n}} = 2.728 \times \sqrt{\frac{5\times6}{6\times5}} = 2.728.$$

$|\bar{r}_{\cdot 1} - \bar{r}_{\cdot 2}| = 0.2, |\bar{r}_{\cdot 1} - \bar{r}_{\cdot 3}| = 0.6, |\bar{r}_{\cdot 1} - \bar{r}_{\cdot 4}| = 2.6, |\bar{r}_{\cdot 1} - \bar{r}_{\cdot 5}| = 1, |\bar{r}_{\cdot 2} - \bar{r}_{\cdot 3}| = 0.8, |\bar{r}_{\cdot 2} - \bar{r}_{\cdot 4}| = 2.8, |\bar{r}_{\cdot 2} - \bar{r}_{\cdot 5}| = 1.2, |\bar{r}_{\cdot 3} - \bar{r}_{\cdot 4}| = 2, |\bar{r}_{\cdot 3} - \bar{r}_{\cdot 5}| = 0.4. |\bar{r}_{\cdot 4} - \bar{r}_{\cdot 5}| = 1.6$,
We could see that only $|\bar{r}_{\cdot 2} - \bar{r}_{\cdot 4}| = 2.8 > 2.728 = CD$. *Therefore, we could draw the following conclusions:*

1. *D is the best-performing algorithm;*

2. *D and B have significant different performances;*

3. *There are no significant differences in performances of other algorithm pairs (D and A, D and C, and D and E).*

5 [20pts] ROC and AUC

现在有五个测试样例，其对应的真实标记和学习器的输出值如表2所示：

表 2: 测试样例表

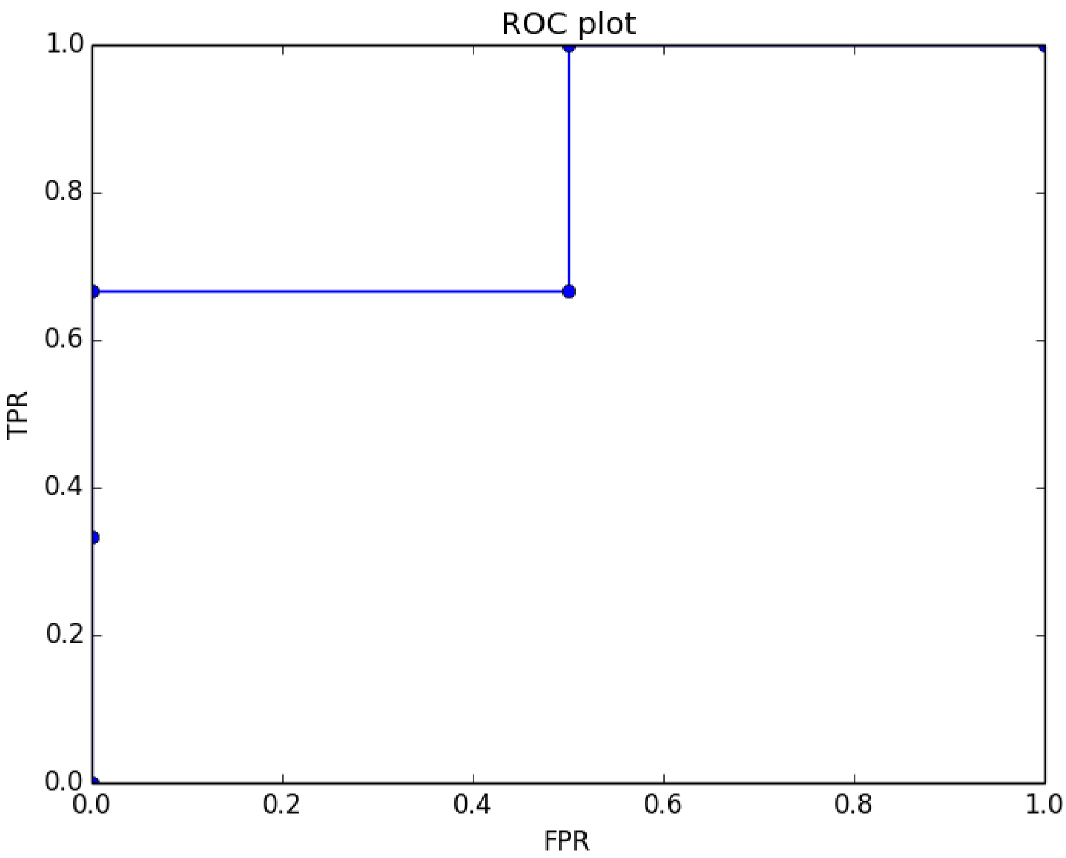
样本	x_1	x_2	x_3	x_4	x_5
标记	+	+	-	+	-
输出值	0.9	0.3	0.1	0.7	0.4

(1) **[10pts]** 请画出其对应的 ROC 图像，并计算对应的 AUC 和 ℓ_{rank} 的值（提示：可使用TikZ包作为 LaTeX 中的画图工具）；

(2) **[10pts]** 根据书上第 35 页中的公式 (2.20) 和公式 (2.21)，试证明

$$\mathrm{AUC} + \ell_{rank} = 1.$$

图 1: ROC-plot

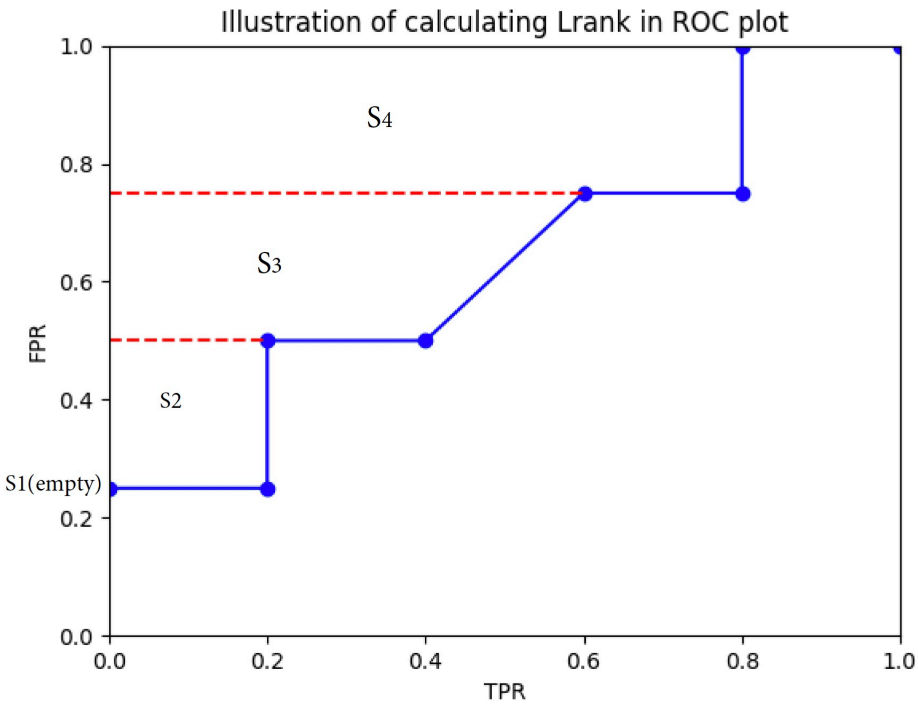


Solution. *(1) AUC* $= \frac{5}{6}$, $\ell_{rank} = \frac{1}{6}$. *The ROC-plot is Figure 1.*

(2) *Obviously, AUC = area of the part below the ROC-curve, so it suffices to prove that* $\ell_{rank} =$ *Area* S_{above} *where* $S_{above} \triangleq$ *the part above the ROC-curve.*

We will prove this by 3 steps. In addition, the Figure 2 could be used for illustration of the proof.

图 2: Illustration



Step 1. Decomposition of S_{above}.

Note that, when plotting AUC curve, for each cut-point-value we are currently considering and given the previous cut point's coordinate (x, y) *:*

If there is only one corresponding cut-point in the samples:

- *if the cut point is true positive, then the previous coordinate will "go up" by* $\frac{1}{m^+}$*, which results in a new point* $(x, y + \frac{1}{m^+})$*;*
- *if the cut point is false positive, then the precious coordinate will "go right" by* $\frac{1}{m^-}$*, which results in a new point* $x + \frac{1}{n^-}, y)$

Otherwise, there are multiple samples that share the same prediction value. In this case, the new point added will "go up" by $n_1 * \frac{1}{m^+}$*, and "go right" by* $n_2 * \frac{1}{m^-}$*, with corresponding coordinate* $(x + n_2 * \frac{1}{m^-}, y + n_1 * \frac{1}{m^+})$*, where* $n_1 (n_2)$ *is the number of positive (negative) samples have ranked all the same with the current cut-off-value.*

We could index the samples in $\mathcal{D}^+$ *(the set of positive samples) in decreasing order of prediction values as below shows:*

$$\mathcal{D}^+ = \{\mathbf{x}_k^+ \mid f(\mathbf{x}_k^+) \geq f(\mathbf{x}_{k+1}^+) \text{ for } k = 1, \cdots, m-1, \text{ and } \mathbf{x}_k^+ \text{is predicted positive}\},$$

and let

$$(x_k^+, y_k^+) = \text{ the coordinate of } \mathbf{x}_k^+,$$

$$S_k = \{(x,y) \mid y_{k-1}^+ < y \leq y_k^+, 0 < x < x_k^+\} \qquad (\textit{set } y_0^+ = 0\ .$$

Therefore, $S_{above} = \bigcup_{k=1}^{m^+} S_k$.

Step 2. Calculation of Area (S_k).

Note that $\sum_{\mathbf{x}^- \in \mathcal{D}^-} \mathbb{I}(f(\mathbf{x}_k^+) < f(\mathbf{x}_k^-)) + \frac{1}{2}\mathbb{I}(f(\mathbf{x}_k^+) = f(\mathbf{x}_k^-))$ *is the number of negative samples ranked before or the same as* $\mathbf{x}_k^+$, *therefore* $x_k^+ = \frac{1}{m^-}\sum_{\mathbf{x}^- \in \mathcal{D}^-} \mathbb{I}(f(\mathbf{x}_k^+) < f(\mathbf{x}_k^-)) + \mathbb{I}(f(\mathbf{x}_k^+) = f(\mathbf{x}_k^-))$.

If $\sum_{\mathbf{x}^- \in \mathcal{D}^-} \mathbb{I}(f(\mathbf{x}_k^+) = f(\mathbf{x}_k^-)) = 0$, *which suggests no negative samples are ranked the same as* $\mathbf{x}_k^+$, *then* S_k *is a trapezoid; otherwise, it is a rectangle.*

In either way,

$$\textit{Area } (S_k) = \frac{1}{m^+}\frac{1}{m^-}\sum_{\mathbf{x}^- \in \mathcal{D}^-} (\mathbb{I}(f(\mathbf{x}_k^+) < f(\mathbf{x}_k^-)) + \frac{1}{2}\mathbb{I}(f(\mathbf{x}_k^+) = f(\mathbf{x}_k^-)).$$

Step 3. Calculaiton of Area (S_{above}).

Since $\{S_k\}_{k=1}^{m^+}$ *are mutually disjoint,*

$$\begin{aligned}
\textit{Area } (S_{above}) &= \bigcup_{k=1}^{m} \textit{Area } (S_k) \\
&= \sum_{k=1}^{m^+} \frac{1}{m^+}\frac{1}{m^-}\sum_{\mathbf{x}^- \in \mathcal{D}^-} (\mathbb{I}(f(\mathbf{x}_k^+) < f(\mathbf{x}_k^-)) + \frac{1}{2}\mathbb{I}(f(\mathbf{x}_k^+) = f(\mathbf{x}_k^-)) \\
&= \frac{1}{m^+}\frac{1}{m^-}\sum_{\mathbf{x}^+ \in \mathcal{D}^+}\sum_{\mathbf{x}^- \in \mathcal{D}^-} (\mathbb{I}(f(\mathbf{x}_k^+) < f(\mathbf{x}_k^-)) + \frac{1}{2}\mathbb{I}(f(\mathbf{x}_k^+) = f(\mathbf{x}_k^-)) \\
&= \ell_{rank}.
\end{aligned}$$

□

6 [附加题 10pts] Expected Prediction Error

对于最小二乘线性回归问题，我们假设其线性模型为：

$$y = \mathbf{x}^T \boldsymbol{\beta} + \epsilon, \tag{6.1}$$

其中 ϵ 为噪声满足 $\epsilon \sim N(0, \sigma^2)$。我们记训练集 $\mathcal{D}$ 中的样本特征为 $\mathbf{X} \in \mathbb{R}^{p \times n}$，标记为 $\mathbf{Y} \in \mathbb{R}^n$，其中 n 为样本数，p 为特征维度。已知线性模型参数的估计为：

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{Y}. \tag{6.2}$$

对于给定的测试样本 $\mathbf{x}_0$，记 $\mathbf{EPE}(\mathbf{x}_0)$ 为其预测误差的期望 (Expected Predication Error)，试证明,

$$\mathbf{EPE}(\mathbf{x}_0) = \sigma^2 + \mathbb{E}_{\mathcal{D}}[\mathbf{x}_0^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{x}_0\sigma^2].$$

要求证明中给出详细的步骤与证明细节。(提示：$\mathbf{EPE}(\mathbf{x}_0) = \mathbb{E}_{y_0|\mathbf{x}_0}\mathbb{E}_{\mathcal{D}}[(y_0 - \hat{y}_0)^2]$，可以参考书中第 45 页关于方差-偏差分解的证明过程。)

Proof. *We will give a proof by three steps.*

Step 1. Preparation

First, we need to clarify some notations.

- $\hat{\boldsymbol{\beta}}$ *is the least-square-estimate coefficient, with*

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{Y} = (\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\beta} + \epsilon) = \boldsymbol{\beta} + (\boldsymbol{X}\boldsymbol{X}^T)^{-1}X\epsilon.$$

  $\hat{\boldsymbol{\beta}}$ *is an unbiased estimator of* $\boldsymbol{\beta}$ *since*

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbb{E}[\boldsymbol{\beta} + (\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\epsilon] = \boldsymbol{\beta}.$$

- y_0 *is the observation value, with*

$$y_0 = \boldsymbol{x}_0^T\beta + \epsilon. \tag{6.3}$$

- $\hat{y}_0$ *is the prediction value, with*

$$\hat{y}_0 = \boldsymbol{x}_0^T\hat{\beta} = \boldsymbol{x}_o^T\boldsymbol{\beta} + \boldsymbol{x}_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\epsilon. \tag{6.4}$$

- *Define the expected prediction value* $\bar{y}_0$ *with*

$$\bar{y}_0 \triangleq \mathbb{E}_{\mathcal{D}}[\hat{y}_0]. \tag{6.5}$$

  Then substituting (6.4) into (6.5), we obtain

$$\bar{y}_0 = \mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_o^T\boldsymbol{\beta} + \boldsymbol{x}_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\epsilon] = \boldsymbol{x}_o^T\boldsymbol{\beta} + \boldsymbol{x}_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\epsilon. \tag{6.6}$$

Step 2. Decomposition of $\mathbf{EPE}(\boldsymbol{x}_0)$

$$
\begin{aligned}
\mathbf{EPE}(\boldsymbol{x}_0) &= \mathbb{E}_{y_0|\boldsymbol{x}_0}\mathbb{E}_{\mathcal{D}}[(y_0-\hat{y}_0)^2 \\
&= \mathbb{E}_{y_0|\boldsymbol{x}_0}\mathbb{E}_{\mathcal{D}}[(y_0-\bar{y}_0+\bar{y}_0-\hat{y}_0)^2 \\
&= \mathbb{E}_{y_0|\boldsymbol{x}_0}\mathbb{E}_{\mathcal{D}}[(y_0-\bar{y}_0)^2+\mathbb{E}_{y_0|\boldsymbol{x}_0}\mathbb{E}_{\mathcal{D}}[(\bar{y}_0-\hat{y}_0)^2+2\mathbb{E}_{y_0|\boldsymbol{x}_0}\mathbb{E}_{\mathcal{D}}[(y_0-\bar{y}_0)(\bar{y}_0-\hat{y}_0)
\end{aligned}
\tag{6.7}
$$

- *The first item*

  Using (6.3 and (6.6), we have

$$
\begin{aligned}
\mathbb{E}_{y_0|\boldsymbol{x}_0}\mathbb{E}_{\mathcal{D}}[(y_0-\bar{y}_0)^2 &= \mathbb{E}_{y_0|\boldsymbol{x}_0}\mathbb{E}_{\mathcal{D}}[(\boldsymbol{x}_0^T\boldsymbol{\beta}+\epsilon-\boldsymbol{x}_0^T\boldsymbol{\beta})^2] \\
&= \mathbb{E}_{y_0|\boldsymbol{x}_0}\mathbb{E}_{\mathcal{D}}[\epsilon^2] \\
&= \sigma^2
\end{aligned}
\tag{6.8}
$$

- *The second item*

  Using (6.6) and (6.4), we have

$$
\begin{aligned}
\mathbb{E}_{y_0|\boldsymbol{x}_0}\mathbb{E}_{\mathcal{D}}[(\bar{y}_0-\hat{y}_0)^2] &= \mathbb{E}_{y_0|\boldsymbol{x}_0}\mathbb{E}_{\mathcal{D}}[(\boldsymbol{x}_0^T\boldsymbol{\beta}-\boldsymbol{x}_0^T\boldsymbol{\beta}+\boldsymbol{x}_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\epsilon)^2] \\
&= \mathbb{E}_{y_0|\boldsymbol{x}_0}\mathbb{E}_{\mathcal{D}}[(\boldsymbol{x}_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\epsilon)^2] \\
&= \mathbb{E}_{y_0|\boldsymbol{x}_0}\mathbb{E}_{\mathcal{D}}[(\boldsymbol{x}_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\epsilon)(\boldsymbol{x}_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\epsilon)] \\
&= \mathbb{E}_{y_0|\boldsymbol{x}_0}\mathbb{E}_{\mathcal{D}}[(\boldsymbol{x}_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X})(\boldsymbol{x}_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X})\epsilon^2]
\end{aligned}
$$

  Since $\boldsymbol{X}\boldsymbol{X}^T$ *is symmetric, which suggests* $\boldsymbol{x}_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}$ *is a quadratic form, we have*

$$
\boldsymbol{x}_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}=\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{x}_0
$$

  . Therefore,

$$
\begin{aligned}
\mathbb{E}_{y_0|\boldsymbol{x}_0}\mathbb{E}_{\mathcal{D}}[(\bar{y}_0-\hat{y}_0)^2] &= \mathbb{E}_{y_0|\boldsymbol{x}_0}\mathbb{E}_{\mathcal{D}}[(\boldsymbol{x}_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X})(\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{x}_0)\epsilon^2] \\
&= \mathbb{E}_{y_0|\boldsymbol{x}_0}\mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}(\boldsymbol{X}\boldsymbol{X}^T)(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{x}_0\epsilon^2] \\
&= \mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{x}_0\sigma^2]
\end{aligned}
\tag{6.9}
$$

- *The third item Because of the definition of* $\hat{y}_0$ *(6.5),*

$$
E_{\mathcal{D}}[(y_0-\bar{y}_0)(\bar{y}_0-\hat{y}_0)=0. \tag{6.10}
$$

Step 3. Putting together

Substituting (6.8), 6.9) and (6.10) into (6.7), we obtain the desired simplified decomposition of $\mathbf{EPE}(\boldsymbol{x}_0)$*.*

□