机器学习导论
作业一

参考解答

2018 年 4 月 4 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) 随机变量X的累积分布函数$F_X(x)$为，

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

(2) 随机变量Y的概率密度函数$f_Y(y)$为，

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

(3) 证明：

$$\text{式}(1.3) = \mathbb{E}[Z] = x\Pr[Z \geq x]$$
$$= x\Pr[Z \geq x]|_{x=0}^{\infty} - \int_{x=0}^{\infty} x \,\mathrm{d}\Pr[Z \geq x]$$
$$= 0 + \int_{x=0}^{\infty} x f(x)\,\mathrm{d}x = \text{式}(1.2).$$

$\mathbb{E}[X] = \frac{25}{8}$,$\mathbb{E}[Y]$不存在.

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

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

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，

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

Proof.

首先我们证明的方向是式(2.2) $\to$式(2.1)，

假设可微函数f满足式(2.2),取可行域内任意两点$\mathbf{x}_1$和$\mathbf{x}_2$。令$\mathbf{y} = \alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2$,并取$\boldsymbol{\theta} \in \partial f(\mathbf{y})$，由式(2.2)可知，

$$f(\mathbf{x}_1) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^{\mathrm{T}}(\mathbf{x}_1-\mathbf{y}) + \frac{\lambda}{2}\|\mathbf{x}_1-\mathbf{y}\|^2,$$
$$f(\mathbf{x}_2) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^{\mathrm{T}}(\mathbf{x}_2-\mathbf{y}) + \frac{\lambda}{2}\|\mathbf{x}_2-\mathbf{y}\|^2.$$

将上述两式分别乘以α和$(1-\alpha)$相加，并将$\mathbf{y}$根据定义代入，可得到所需的结论。

下面，我们来证明另一方向，即式(2.1) $\to$式(2.2)，

假设可微函数f满足式(2.1), 经过简单变形，可以得到

$$\frac{f(\mathbf{x}+\alpha(\mathbf{y}-\mathbf{x})) - f(\mathbf{x})}{\alpha} \leq [f(\mathbf{y}) - f(\mathbf{x})] - \frac{\lambda}{2}(1-\alpha)\|\mathbf{x}-\mathbf{y}\|^2.$$

令$\alpha \to 0$, 我们可以得到，

$$f'(\mathbf{x};\mathbf{y}-\mathbf{x})\|\mathbf{y}-\mathbf{x}\| \leq f(\mathbf{y}) - f(\mathbf{x}) - \frac{\lambda}{2}\|\mathbf{x}-\mathbf{y}\|^2, \tag{2.3}$$

其中$f'(\mathbf{x};\mathbf{y}-\mathbf{x})$是函数$f$在$\mathbf{x}$点的关于$y-x$方向的方向导数。又因为，

$$f'(\mathbf{x};\mathbf{y}-\mathbf{x}) = \frac{\nabla f(\mathbf{x})^{\mathrm{T}}(\mathbf{y}-\mathbf{x})}{\|\mathbf{x}-\mathbf{y}\|}. \tag{2.4}$$

结合式(2.3)和式(2.4)可得式(2.1) $\to$式(2.2)。 □

备注. 本答案中式(2.1) →式(2.2)的推导我们参考了李奕萱同学（151160031）的解题方法，因为我们认为她的解题思路比原答案的解题思路更好。在此我们向李同学表示感谢。这里是李奕萱同学的参考文献链接*https://arxiv.org/pdf/1803.06573.pdf,*大家有兴趣可以阅读。

本题中式(2.2) →式(2.1)的证明*10*分，式(2.1) →式(2.2)的证明*10*分。

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*(double stochastic matrix)*。

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

(1) **[10pts]** 定义矩阵$\mathbf{X}$的统计量$H(\mathbf{X}) = -\sum_{i=1}^{d}\sum_{j=1}^{d} x_{ij}\log x_{ij}$. 试证明：$H(\mathbf{X}) \leq d\log d$.

(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) 利用信息熵函数是凹函数，由*Jensen*不等式，

$$\begin{aligned} H(\mathbf{X}) &= -\sum_{i=1}^{d}\sum_{j=1}^{d} x_{ij}\log x_{ij} \\ &\leq -d^2 \cdot \frac{s}{d^2}\log\frac{s}{d^2} = d\log d \end{aligned} \tag{3.3}$$

其中最后一个等式成立是因为$\mathbf{X}$是双随机矩阵，因此有$s = \sum_{i=1}^{d}\sum_{j=1}^{d} x_{ij} = d$. □

(2) 由关于非负矩阵的广义*Perron–Frobenius*定理可知，对于非负矩阵$\mathbf{X}$有

$$\min_{1\leq i\leq d}\sum_{j=1}^{d} x_{ij} \leq \rho(\mathbf{X}) \leq \max_{1\leq i\leq d}\sum_{j=1}^{d} x_{ij}. \tag{3.4}$$

因为$\mathbf{X}$是双随机矩阵，因此有$\rho(\mathbf{X}) = 1$。同时，注意到$\mathbf{X}\cdot\mathbf{1} = 1\cdot\mathbf{1}$，因此*1*是$\mathbf{X}$的特征值。 □

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:

Table 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. *检验结果为: Friedman Test 统计量$\tau_F = 3.9365$, p-value=0.0207. Nemenyi Test 的临界值域为: 2.7278.*
所以说这五种算法的性能显著不同. 通过Nemenyi test, 发现$BD = 2.728, 1.2 + 2.728 = 3.928 < 4$,所以算法$B$ 和算法D 在统计意义上显著不同.

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：

Table 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

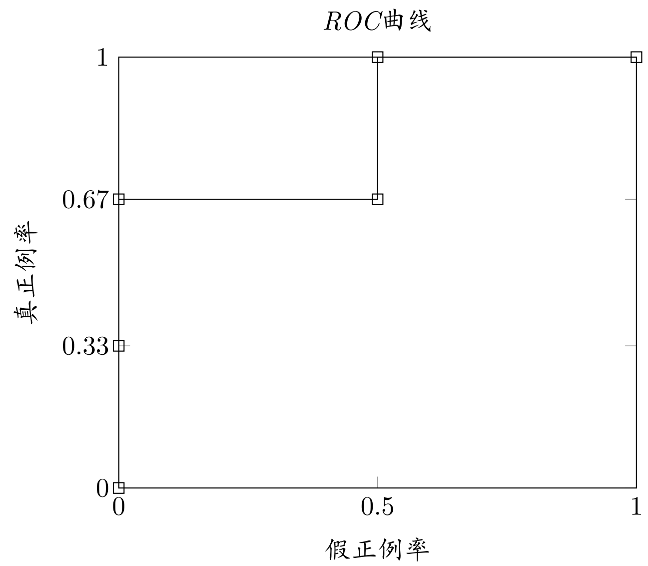
(1) **[10pts]** 请画出其对应的ROC图像，并计算对应的AUC和ℓ_{rank}的值（提示：可使用TikZ包作为LaTeX中的画图工具）；

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

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

Solution.

(1) 画出的ROC曲线如下图所示:



对ROC曲线下的区域求面积得到AUC大小为：

$$
\begin{aligned}
AUC &= \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1}x_i)\dot{(}y_i + y_{i+1}) \\
&= 0.5 \times \frac{2}{3} + 0.5 \times 1 \\
&= \frac{5}{6}
\end{aligned}
$$

根据ℓ_{rank}的定义可以求得：

$$
\begin{aligned}
\ell_{rank} &= \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}(f(x^+) < f(x^-)) + \frac{1}{2}\mathbb{I}(f(x^+) = f(x^-))\right) \\
&= \frac{1}{2\times 3} \times 1 \\
&= \frac{1}{6}
\end{aligned}
$$

(2)

由AUC公式可知，AUC值对应ROC曲线下面积.
下面分析ROC曲线下方的面积.

1. 对每单位纵向线上方格子：面积$S_1 = 0$;
2. 对每单位横向线上方格子：令此时增加的FP为y_p，横向线上方格子数即为预测值小于y_p的正例数，面积$S_2 = \sum_{i=1}^{m} II(f(x_i) < f(y_p))$，$x_i$为第$i$个正例；
3. 对每单位斜向线上方格子：令此时增加的FP为y_q，斜向线上方格子数即为预测值小于y_q的正例数和预测值等于y_q的正例所占格子的一半，面积$S_3 = \sum_{i=1}^{m} II(f(x_i) < f(y_q)) + \frac{1}{2}II(f(x_i) = f(y_q))$，$x_i$为第$i$个正例；

所以说$(m^+m^-)AUC = S_1+S_2+S_3 = \sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}(f(x^+) > f(x^-)) + \frac{1}{2}\mathbb{I}(f(x^+) = f(x^-))\right)$.
其中x_i为第i个真正例，y_j为第j个假正例.
而ℓ_{rank}的公式为：

$$\ell_{rank} = \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}(f(x^+) < f(x^-)) + \frac{1}{2}\mathbb{I}(f(x^+) = f(x^-))\right)$$

正因为此，$\ell_{rank} + AUC = 1$.

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：$\mathbf{EPE}(\mathbf{x}_0) = \mathbb{E}_{y_0|\mathbf{x}_0}\mathbb{E}_{\mathcal{D}}[(y_0-\hat{y}_0)^2]$，其中$y_0$为测试样本未知的真实标记，即$y_0 = \mathbf{x}_0^T\boldsymbol{\beta}+\epsilon$,而$\hat{y}_0$则是线性模型对于$y_0$的估计,$\mathbb{E}_{y_0|\mathbf{x}_0}$是$y_0$在$\mathbf{x}_0$给定时的条件期望。可以参考书中第45页关于方差-偏差分解的证明过程。)

Proof.

记$\boldsymbol{\mathcal{E}}$为训练集中所有样本的噪声形成的向量，即$\mathcal{E} = [\epsilon_1,\dots,\epsilon_n]^{\mathrm{T}}$，则有$\mathbf{Y} = \mathbf{X}^T\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}$.

$$\begin{aligned}\hat{y}_0 &= \mathbf{x}_0^T\hat{\boldsymbol{\beta}}\\ &= \mathbf{x}_0^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}(\mathbf{X}^T\boldsymbol{\beta}+\boldsymbol{\mathcal{E}})\\ &= \mathbf{x}_0^T\boldsymbol{\beta} + \mathbf{x}_0^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\boldsymbol{\mathcal{E}}\end{aligned} \tag{6.3}$$

$$\begin{aligned}Var_{\mathcal{D}}(\hat{y}_0) &= \mathbb{E}_{\mathcal{D}}[(\hat{y}_0 - \mathbb{E}_{\mathcal{D}}\hat{y}_0)^2]\\ &= \mathbb{E}_{\mathcal{D}}[\mathbf{x}_0^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}^T\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{x}_0]\\ &= \mathbb{E}_{\mathcal{D}}[\mathbf{x}_0^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{I}_p\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{x}_0\sigma^2]\\ &= \mathbb{E}_{\mathcal{D}}[\mathbf{x}_0^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{x}_0\sigma^2]\end{aligned} \tag{6.4}$$

$$\begin{aligned}
\mathbf{EPE}(\mathbf{x}_0) &= \mathbb{E}_{y_0|\mathbf{x}_0}\mathbb{E}_{\mathcal{D}}[(y_0-\hat{y}_0)^2] \\
&= \mathbb{E}_{y_0|\mathbf{x}_0}(y_0^2 - 2y_0\mathbb{E}_{\mathcal{D}}(\hat{y}_0) + \mathbb{E}_{\mathcal{D}}(\hat{y}_0^2)) \\
&= \mathbb{E}_{y_0|\mathbf{x}_0}\{\mathbb{E}_{\mathcal{D}}[(\hat{y}_0 - \mathbb{E}_{\mathcal{D}}(\hat{y}_0))^2] + (\mathbb{E}_{\mathcal{D}}(\hat{y}_0) - \mathbf{x}_0^T\boldsymbol{\beta})^2 + (\mathbf{x}_0^T\boldsymbol{\beta} - y_0)^2\} \\
&= \mathbb{E}_{\mathcal{D}}[(\hat{y}_0 - \mathbb{E}_{\mathcal{D}}(\hat{y}_0))^2] + (\mathbb{E}_{\mathcal{D}}(\hat{y}_0) - \mathbf{x}_0^T\boldsymbol{\beta})^2 + Var(y_0|\mathbf{x}_0) \\
&= \ Var_{\mathcal{D}}(\hat{y}_0) + Bia^2(\hat{y}_0) + Var(y_0|\mathbf{x}_0) \\
&= \mathbb{E}_{\mathcal{D}}[\mathbf{x}_0^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{x}_0\sigma^2] + 0 + \sigma^2
\end{aligned}$$

$\square$