

机器学习导论

作业一

学号, 作者姓名, 邮箱

2018 年 3 月 20 日

学术诚信

本课程非常重视学术诚信规范，助教老师和助教同学将不予余力地维护作业中的学术诚信规范的建立。希望所有选课学生能够对此予以重视。¹

- (1) 允许同学之间的相互讨论，但是**署你名字的工作必须由你完成**，不允许直接照搬任何已有的材料，必须独立完成作业的书写过程；
- (2) 在完成作业过程中，对他人工作（出版物、互联网资料）中文本的直接照搬（包括原文的直接复制粘贴及语句的简单修改等）都将视为剽窃，剽窃者成绩将被取消。**对于完成作业中有关键作用的公开资料，应予以明显引用；**
- (3) 如果发现作业之间高度相似将被判定为互相抄袭行为，**抄袭和被抄袭双方的成绩都将被取消**。因此请主动防止自己的作业被他人抄袭。

作业提交注意事项

- (1) 请严格参照课程网站作业提交方法一节提交作业；
- (2) 未按照要求提交作业，或提交作业格式不正确，将会被扣除部分作业分数；
- (3) 除非有特殊情况（如因病缓交），否则截止时间后不接收作业，本次作业记零分。

¹参考尹一通老师高级算法课程中对学术诚信的说明。

1 [25pts] Basic Probability and Statistics

随机变量 X 的概率密度函数如下,

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

- (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。

$$\mathbb{E}[Z] = \int_{z=0}^{\infty} z f(z) dz. \quad (1.2)$$

$$\mathbb{E}[Z] = \int_{z=0}^{\infty} \Pr[Z \geq z] dz. \quad (1.3)$$

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. 此处用于写解答(中英文均可)

1.

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

2.

$$f_Y(y) = \begin{cases} \frac{3}{8y^2} & \frac{1}{5} < y < \frac{1}{3}; \\ \frac{1}{4y^2} & y > 1; \\ 0 & \text{otherwise.} \end{cases} \quad (1.5)$$

3. 证明:

$$\begin{aligned} \int_{z=0}^{\infty} \Pr[Z \geq z] dz &= \int_{z=0}^{\infty} (1 - F_Z(z)) dz = \int_{z=0}^{\infty} \int_{v=z}^{\infty} f_Z(v) dv dz \\ &= \iint_{0 \leq z \leq v < \infty} f_Z(v) dv dz = \int_{v=0}^{\infty} f_Z(v) \int_{z=0}^v dz dv \\ &= \int_{v=0}^{\infty} v f_Z(v) dv \end{aligned} \quad (1.6)$$

由此可得(1.2)与(1.3)等价。

验证:

$$\mathbb{E}[X] = \int_{x=0}^{\infty} x f(x) dx = \int_0^1 \frac{1}{4}x dx + \int_3^5 \frac{3}{8}x dx = \frac{25}{8} \quad (1.7)$$

$$\begin{aligned}\mathbb{E}[X] &= \int_{x=0}^{\infty} \Pr[X \geq x] \, dx = \int_0^1 \left(1 - \frac{1}{4}x\right) \, dx + \int_1^3 \frac{3}{4} \, dx + \int_3^5 \left(\frac{15}{8} - \frac{3}{8}x\right) \, dx \\ &= \frac{25}{8}\end{aligned}\tag{1.8}$$

$$\mathbb{E}[Y] = \int_{y=0}^{\infty} y f(y) \, dy = \int_{\frac{1}{5}}^{\frac{1}{3}} \frac{3}{8y} \, dy + \int_1^{\infty} \frac{1}{4y} \, dy = \infty\tag{1.9}$$

$$\mathbb{E}[Y] = \int_{y=0}^{\infty} \Pr[Y \geq y] \, dy = \infty\tag{1.10}$$

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 $f: \mathcal{K} \rightarrow \mathbb{R}$ ，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

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

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，

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

Proof. 此处用于写证明(中英文均可)

令 $g(\mathbf{y}) = f(\mathbf{y}) - \frac{\lambda}{2}\|\mathbf{y}\|^2$ 。要证明(2.1)和(2.2)等价，先证明(2.2)和

$$g(\mathbf{y}) \text{ is convex} \quad (2.3)$$

等价。

因为

$$\begin{aligned} & g(\mathbf{y}) \text{ is convex} \\ \Leftrightarrow & g(\mathbf{y}) \geq g(\mathbf{x}) + \nabla g(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \\ \Leftrightarrow & f(\mathbf{y}) - \frac{\lambda}{2}\|\mathbf{y}\|^2 \geq f(\mathbf{x}) - \frac{\lambda}{2}\|\mathbf{x}\|^2 + (\nabla f(\mathbf{x})^T - \lambda\mathbf{x}^T)(\mathbf{y} - \mathbf{x}) \\ \Leftrightarrow & f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{\lambda}{2}\|\mathbf{y}\|^2 - \frac{\lambda}{2}\|\mathbf{x}\|^2 - \lambda\mathbf{x}^T(\mathbf{y} - \mathbf{x}) \\ \Leftrightarrow & f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{\lambda}{2}\|\mathbf{y} - \mathbf{x}\|^2 \end{aligned} \quad (2.4)$$

可得(2.2)与(2.3)等价，下面证明(2.1)与(2.3)等价

$$\begin{aligned} & g(\mathbf{y}) \text{ is convex} \\ \Leftrightarrow & g((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1-\alpha)g(\mathbf{x}) + \alpha g(\mathbf{y}) \\ \Leftrightarrow & f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) - \frac{\lambda}{2}\|(1-\alpha)\mathbf{x} + \alpha\mathbf{y}\|^2 \leq (1-\alpha)(f(\mathbf{x}) - \frac{\lambda}{2}\|\mathbf{x}\|^2) + \alpha(f(\mathbf{y}) - \frac{\lambda}{2}\|\mathbf{y}\|^2) \\ \Leftrightarrow & f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}) - \frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{x} - \mathbf{y}\|^2 \end{aligned} \quad (2.5)$$

得证，在函数 f 可微的情况下，式 (2.1) 与式 (2.2) 等价。 \square

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

$$\sum_{j=1}^d x_{ij} = 1, \quad i = 1, 2, \dots, d. \quad (3.1)$$

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

$$\sum_{i=1}^d x_{ij} = 1, \quad j = 1, 2, \dots, d. \quad (3.2)$$

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] 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. 此处用于写证明 (中英文均可)

1. 因为 $H(\mathbf{X}) = \sum_{i=1}^d \sum_{j=1}^d x_{ij} \log \frac{1}{x_{ij}}$, 所以要证明 $\sum_{i=1}^d \sum_{j=1}^d x_{ij} \log \frac{1}{x_{ij}} \leq d \log d$. 证明如下:

首先证明 $\sum_{j=1}^d x_{ij} \log \frac{1}{x_{ij}} \leq \log d$, 观察到 $f(x) = x \log \frac{1}{x}$ is concave function, 所以有 $\sum_{j=1}^d f(x_{ij}) \leq d f(\frac{\sum_{j=1}^d x_{ij}}{d})$, 又因为 (3.1), 所以进一步有 $\sum_{j=1}^d f(x_{ij}) \leq d \frac{1}{d} \log \frac{d}{1} = \log d$, 因此可得 $H(\mathbf{X}) = \sum_{i=1}^d \sum_{j=1}^d x_{ij} \log \frac{1}{x_{ij}} \leq d \log d$.

2. 首先易得 1 是 \mathbf{X} 的特征值之一, 对应的特征向量 \mathbf{v} 是全一向量: 因为 $\mathbf{X}\mathbf{v} = 1\mathbf{v}$.

根据 *Inequalities for Perron–Frobenius eigenvalue*: 对任意矩阵 \mathbf{A} 的特征值 λ , 一定有 $|\lambda| \leq \max_i \sum_j |\mathbf{A}_{ij}|$. 又因为 \mathbf{X} 是双随机矩阵所以对于其任意特征值 λ 都有 $|\lambda| \leq 1$, 又因为 1 是 \mathbf{X} 的特征值之一, 所以可得 $\rho(\mathbf{X}) = \max \{|\lambda_1|, \dots, |\lambda_d|\} = 1$.

□

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. 此处用于写解答(中英文均可)

*Friedman*检验：

由题目得 $k = 5$, $N = 5$ ，根据简单计算有 $r_1 = 3.8, r_2 = 4, r_3 = 3.2, r_4 = 1.2, r_5 = 2.8$ ，根据书上(2.34)公式可得 $\tau_{\chi^2} = 9.9200$ ，再根据书上(2.35)公式可得 $\tau_F = 3.9365$ ，查书上表2.6可得 $3.9365 > 3.007$ ，所以拒绝“所有算法性能相同”这个假设。

*Nemenyi*检验：

查表2.7可得当 $k = 5, \alpha = 0.05$ 时 $q_\alpha = 2.728$ ，根据书上(2.36)可得 $CD = 2.728$ 。观察题目中表1可得 $r_2 - r_4 = 2.8 > 2.728$ 所以认为算法D与算法B有显著区别，其他算法之间无显著区别。

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

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

Solution. 此处用于写解答(中英文均可)

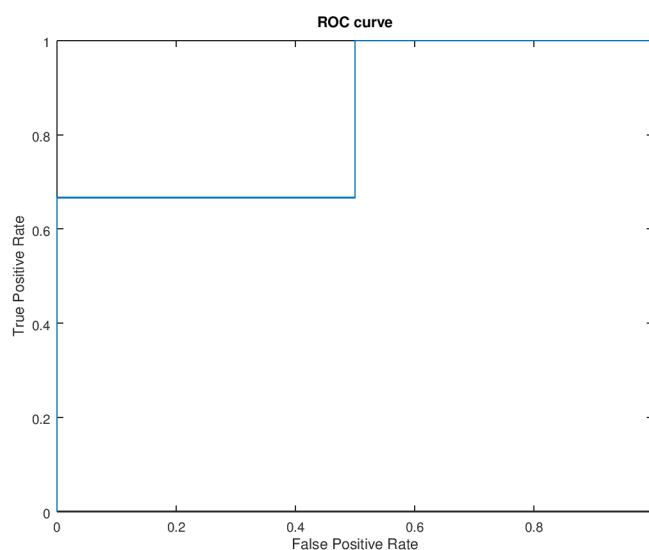


图 1: 根据表2中数据绘制的roc曲线

1.

$$\begin{aligned} \ell_{rank} &= \frac{1}{m^+m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \left(\mathbb{I}(f(x^+) < f(x^-)) + \frac{1}{2} \mathbb{I}(f(x^+) = f(x^-)) \right) \\ &= \frac{1}{6} = 0.1667 \end{aligned} \quad (5.1)$$

$$AUC = 1 - \ell_{rank} = 0.8333 \quad (5.2)$$

2. 要证明 $AUC + \ell_{rank} = 1$, 即证明:

$$AUC = \frac{1}{m^+m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \left(\mathbb{I}(f(x^+) > f(x^-)) + \frac{1}{2} \mathbb{I}(f(x^+) = f(x^-)) \right) \quad (5.3)$$

根据 AUC 的定义可得 AUC 表示的是 ROC 曲线下方围成的面积, 那么先定义一个 $unit$ 为大小为 $\frac{1}{m^+} \frac{1}{m^-}$ 的矩形块面积, 在已知上一个样例的坐标为 (x_0, y_0) 时, 如果当前样例是正样例, 那么当前新坐标就是 $(x_0, y_0 + \frac{1}{m^+})$, 并不增加面积。如果当前样例是负样例, 那么当前新坐标就是 $(x_0 + \frac{1}{m^-}, y_0)$, 会增加 ROC 之下的面积, 增加的面积大小为 k 个 $unit$ 的大小, k 是输出值大于当前负样例的正样例的个数。

由此可得, 对于任意两个一正一负的样例 $s_1 \in D^+$, $s_2 \in D^-$, 有以下三种情况:

- $f(s_1) > f(s_2)$, 那么 s_1 和 s_2 会为 AUC 贡献一个 $unit$ 的面积, 因为 s_1 贡献了 $\frac{1}{m^+}$ 的 y 轴长度, s_2 贡献了 $\frac{1}{m^-}$ 的 x 轴长度。
- $f(s_1) = f(s_2)$, 贡献一个三角形的面积, 大小为半个 $unit$ 大小。
- $f(s_1) < f(s_2)$, 不贡献面积。

综上所述, 可得:

$$AUC = \frac{1}{m^+m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \left(\mathbb{I}(f(x^+) > f(x^-)) + \frac{1}{2} \mathbb{I}(f(x^+) = f(x^-)) \right) \quad (5.4)$$

所以

$$AUC + \ell_{rank} = \frac{1}{m^+m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} 1 = 1 \quad (5.5)$$

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：

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

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：

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

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 Prediction 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. 此处用于写证明(中英文均可)

□