机器学习导论
作业一

2018 年 6 月 30 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遗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，

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

(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. *此处用于写解答(中英文均可)*

(1) 由$F_X(x) = \int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x$得:

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

(2) 随机变量Y的累计分布函数$F_Y(y) = \Pr[Y \leq y]$

$$\begin{aligned} &= \Pr[1/X \leq y] \\ &= \begin{cases} 0 & y \leq \frac{1}{5}; \\ \Pr[X \geq 1/y] & \textit{otherwise.} \end{cases} \\ &= \begin{cases} 0 & y \leq \frac{1}{5}; \\ 1 - F_X(1/y) & \textit{otherwise.} \end{cases} \\ &= \begin{cases} 0 & y \leq \frac{1}{5}; \\ -\frac{3}{8y} + \frac{15}{8} & \frac{1}{5} \leq y < \frac{1}{3}; \\ \frac{3}{4} & \frac{1}{3} \leq y \leq 1; \\ 1 - \frac{1}{4y} & \textit{otherwise.} \end{cases} \end{aligned}$$

由$f_Y(y) = \dfrac{dF_Y(y)}{dy}$得：

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

(3) 随机变量Z的概率密度函数为$f(z)$,展开式 (1.3)得到二重积分：

$$\mathbb{E}[Z] = \int_0^\infty \int_{z'}^\infty f_Z(z)\,\mathrm{d}z\,\mathrm{d}z'.$$

交换上式对于z和z'的积分次序得：

$$\begin{aligned} \mathbb{E}[Z] &= \int_0^\infty \int_0^z f_Z(z)\,\mathrm{d}z'\,\mathrm{d}z \\ &= \int_0^\infty z f(z)\,\mathrm{d}z. \end{aligned}$$

利用两种期望公式计算得到如下一致的结果：

$$\begin{aligned} \mathbb{E}[X] &= \frac{25}{8} \\ \mathbb{E}[Y] &= \infty. \end{aligned}$$

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. 此处用于写证明(中英文均可)

把式 (1.2)左边泰勒展开得：

$$\begin{aligned}&f((1-\alpha)\mathbf{x}+\alpha\mathbf{y})\\=&f(\mathbf{x}+\alpha(\mathbf{y}-\mathbf{x}))\\=&f(\mathbf{x})+\nabla f(\mathbf{x})^{\mathrm{T}}(\alpha(\mathbf{y}-\mathbf{x}))+o(\alpha(\mathbf{y}-\mathbf{x}))\end{aligned} \tag{2.3}$$

式 (2.3)与式 (2.1)联立得：

$$\begin{aligned}&f(\mathbf{x})+\nabla f(\mathbf{x})^{\mathrm{T}}(\alpha(\mathbf{y}-\mathbf{x}))+o(\alpha(\mathbf{y}-\mathbf{x}))\\\leq&(1-\alpha)f(\mathbf{x})+\alpha f(\mathbf{y})-\frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{x}-\mathbf{y}\|^2.\end{aligned} \tag{2.4}$$

考虑$\alpha\to0$时，式 (2.4)移项，左右两端同时除以α即可得到：

$$f(\mathbf{y})\geq f(\mathbf{x})+\nabla f(\mathbf{x})^{\mathrm{T}}(\mathbf{y}-\mathbf{x})+\frac{\lambda}{2}\|\mathbf{y}-\mathbf{x}\|^2. \tag{2.5}$$

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]** 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*)*

证明简洁起见，不妨设题目中的*log*的底数为自然常数e，即$log\mathbf{x} = ln\mathbf{x}$

$$H(\mathbf{X}) = -\sum_{i=1}^{d}\sum_{j=1}^{d} x_{ij} log(x_{ij}) \tag{3.3}$$

下面利用拉格朗日乘子法求解式 (**??**)在题目中式 (3.1)和式 (3.2)的约束下的极（大）值。记拉格朗日函数为:

$$\begin{aligned} L(\mathbf{X}, \lambda, \mu) = &-\sum_{i=1}^{d}\sum_{j=1}^{d} x_{ij} log(x_{ij}) \\ &+ \sum_{i=1}^{d} \lambda_i (\sum_{j=1}^{d} x_{ij} - 1) \\ &+ \sum_{j=1}^{d} \mu_i (\sum_{i=1}^{d} x_{ij} - 1) \end{aligned} \tag{3.4}$$

其中，$\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_d)$,$\mu = (\mu_1, \mu_2, \cdots, \mu_d)$.

拉格朗日函数L对各变量求导并置*0*得:

$$\frac{dL}{dx_{ij}} = -1 - lnx_{ij} + \lambda_i + \mu_j = 0 \tag{3.5a}$$

$$\frac{dL}{d\lambda_i} = \sum_{j=1}^{d} x_{ij} - 1 = 0 \tag{3.5b}$$

$$\frac{dL}{d\mu_j} = \sum_{i=1}^{d} x_{ij} - 1 = 0 \tag{3.5c}$$

for $i, j = 1, 2, \cdots, d$.

联立以上$d^2 + 2d$个方程，求解得到导数为0时变量的取值：

$$x_{ij} = e^{\lambda_i + \mu_j - 1} = \frac{1}{d} \tag{3.6}$$

for $i, j = 1, 2, \cdots, d$.

此时，$H(\mathbf{X})$取得极值$d \log d$.由表达式易知，该极值是$H(\mathbf{X})$的极大值。因此：

$$H(\mathbf{X}) \leq d \log d \tag{3.7}$$

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*)*

算法A到E的平均序值为:

算法	A	B	C	D	E
平均序值	3.8	4	3.2	1.2	2.8

表 2: 算法平均序值表

根据题目$k = 5, N = 5$,计算得统计变量:

$$\tau_{\chi^2} = \frac{N}{k(k+1)}\left(\sum_{i=1}^{k} r_i^2 - \frac{k(k+1)^2}{4}\right) = 9.92 \tag{4.1}$$

$$\tau_F = \frac{(N-1)\chi^2}{N(k-1)-\chi^2} = 3.937 > 3.007 \tag{4.2}$$

其中，3.007为$\alpha = 0.05$的F检验临界值，因此这些算法性能显著不同,应当将进行$Nemenyi$后续检验。

当$k = 5$时, $q_\alpha = 2.728$,计算得到对应平均序值差别的临界值域$CD = 2.728$只有算法B与最好算法D的差距$(4 - 1.2 = 2.8 > 2.728)$超过了临界值,因此最好的算法D只与算法B有显著差别。

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3所示：

表 3: 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

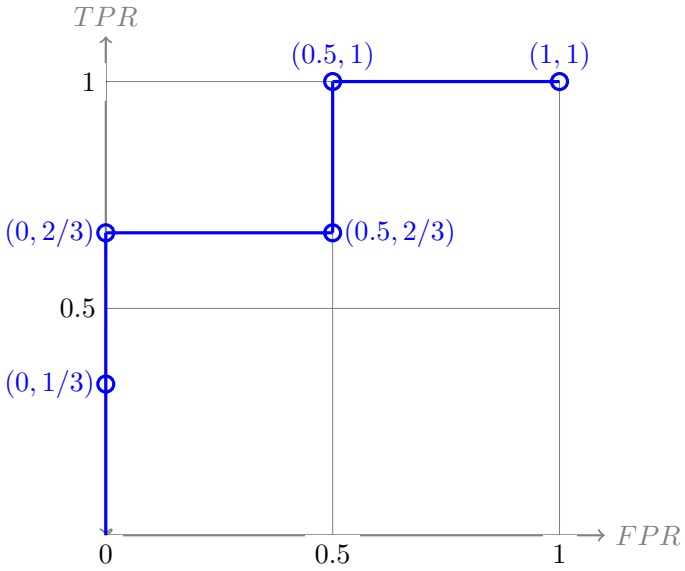
(1) **[10pts]** 请画出其对应的ROC图像，并计算对应的AUC和ℓ_{rank}的值（提示：可使用TikZ包作为LaTeX中的画图工具）；

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

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

Solution. 此处用于写解答(中英文均可)

*(1) ROC*图像：



容易计算ROC曲线下面积$AUC = 2/3 \times 1/2 + 1 \times 1/2 = 5/6$.

由于只有一种正例的的预测值小于反例的情况(x_2, x_5),因此损失$\ell_{rank} = 1/6$.

(2) 设ROC曲线上正例的对应的标记点依次为$((x_1, y_1), (x_2, y_2), \cdots, (x_{m^+}, y_{m^+}))$, 则$x_i$ $(i = 1, 2, \cdots, m^+)$恰是排序在其之前的反例所占的比例，即预测值大于该正例的反例的数量为$m^- \times x_i$.该正例的对应的惩罚为 $\frac{1}{m^+ m^-} \times m^- \times x_i = \frac{x_i}{m^+}$,因此所有的惩罚和为：

$$\sum_{i=1}^{m^+} \frac{x_i}{m^+} = \sum_{i=1}^{m^+} x_i dy \tag{5.1}$$

式(5.1)表明ℓ_{rank}可以看做ROC曲线对$y(TPR)$轴的积分，而AUC表示ROC曲线对$x(FPR)$轴的积分，所以：

$$AUC + \ell_{rank} = 1. \tag{5.2}$$

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. 此处用于写证明*(中英文均可)*

根据式 (6.1)，有：

$$\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{\beta} + \varepsilon \tag{6.3}$$

其中，$\varepsilon = [\epsilon_1, \epsilon_2, \ldots, \epsilon_n]^T$,$\epsilon_i \sim N(0, \sigma^2)(i = 1, 2, \ldots, n)$.

模型的期望预测为：

$$\hat{y}_0 = \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} \tag{6.4}$$

联立式(6.3)(6.4)得：

$$\hat{y}_0 = \boldsymbol{x}_0^T\boldsymbol{\beta} + \boldsymbol{x}_0^T(\boldsymbol{X}^T)^{-1}\varepsilon \tag{6.5}$$

由机器学习参考书中第45页关于方差-偏差分解的证明过程可知：

$$\mathbf{EPE}(\boldsymbol{x}_0) = \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}\hat{y}_0 - \hat{y}_0)^2] + \mathbb{E}_{y_0|\boldsymbol{x}_0}[(\mathbb{E}_{\mathcal{D}}\hat{y}_0 - y_0)^2] \tag{6.6}$$

又由式(6.5)可知，$\mathbb{E}_{\mathcal{D}}\hat{y}_0 = \mathbb{E}_{y_0|\boldsymbol{x}_0}y_0 = \boldsymbol{x}_0^T\boldsymbol{\beta}$, 所以(6.6)可以改写为：

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

由y和$\hat{y}$的表达式(6.1)(6.5)易得：

$$var_{\mathcal{D}}(\hat{y}_0) = \mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{x}_0\sigma^2] \tag{6.8a}$$

$$var_{y_0|\boldsymbol{x}_0}(y_0) = \sigma^2 \tag{6.8b}$$

联立(6.7)(6.8a)(6.8b)得：

$$\mathbf{EPE}(\boldsymbol{x}_0) = \sigma^2 + \mathbb{E}_{\mathcal{D}}[\boldsymbol{x}_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{x}_0\sigma^2] \tag{6.9}$$