机器学习导论习题课

第一次作业

题目1：基本概率和统计

随机变量X的概率分布如下：

$$f_X(x) = \begin{cases} \frac{1}{2} & 0 < x < 1; \\ \frac{1}{6} & 2 < x < 5; \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

（1）请给出X的累积分布函数$\mathrm{F_X}(x)$；

（2）定义随机变量Y满足$\mathrm{Y} = 1/X^2$，请给出Y的概率密度函数$f_Y(y)$；

（3）对于非负随机变量Z，请证明以下两个公式等价：

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

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

同时，请通过计算变量X和Y的期望来验证你的证明。

题目1：基本概率和统计

(1) 请给出X的累积分布函数$\mathbf{F_X}(x)$;

答：考概率分布的基本概念

$$F_X(x) = \begin{cases} 0 & x \leq 0; \\ \frac{1}{2}x & 0 < x \leq 1; \\ \frac{1}{2} & 1 < x \leq 2; \\ \frac{1}{6}x + \frac{1}{6} & 2 < x \leq 5; \\ 1 & x > 5; \end{cases}$$

这一问大部分同学都可以拿到分数。

题目1：基本概率和统计

(2) 请给出$\mathrm{Y} = 1/X^2$的概率密度函数$f_Y(y)$；

答：通过对x的取值分别讨论得到$F_Y(y)$，再通过对$F_Y(y)$求导得到$f_Y(y)$，需要注意$y = 0$点不可导，所以$f_Y(0)$不存在。

$$f_Y(y) = \begin{cases} \frac{1}{12}y^{-\frac{3}{2}} & \frac{1}{25} < y < \frac{1}{4}; \\ \frac{1}{4}y^{-\frac{3}{2}} & y > 1; \\ 0 & otherwise; \end{cases}$$

在这一问中凡是没有写推导过程直接写密度函数的答案都要扣分。

题目1：基本概率和统计

(3) 请证明以下两个公式等价；

答：从公式(1.3)出发

$$\begin{aligned}\mathbb{E}[Z] &= \int_{z=0}^{\infty} \Pr[Z \geq z] \mathrm{d}z \\ &= \int_{x=0}^{\infty} \Pr[Z \geq x] \mathrm{d}x \\ &= \int_{x=0}^{\infty} \int_{z=x}^{\infty} f(z) \mathrm{d}z \mathrm{d}x\end{aligned}$$

然后积分换限得，

$$\begin{aligned}&= \int_{z=0}^{\infty} \int_{x=0}^{z} f(z) \mathrm{d}x \mathrm{d}z \\ &= \int_{z=0}^{\infty} z f(z) \mathrm{d}z\end{aligned}$$

经计算可得$\mathrm{E}[X] = 2, \mathrm{E}[Y]$不存在，由此可知式(1.2)和(1.3)等价的一个条件就是期望必须存在。

$\int F_z(Z)\, dz = zF_z(Z)|$ 不成立，凡是使用该条规则证明的都会扣分

有不少同学由于忘记求解X和Y的期望而被扣分。

题目2：强凸

$D \in R^2$是一个有限集。定义函数$E: R^3 \to R$

$$E(a,b,c) = \sum_{x \in \mathcal{D}} (ax_1^2 + bx_1 + c - x_2)^2. \qquad (2.1)$$

(1) 函数E是凸的吗？

(2) 是否存在一个集合D使得函数E是强凸函数？证明或举出反例。

题目2：强凸

(1) 答：可以算的函数的Hessian矩阵为：

$$D^2E = 2\sum_{x\in D}\begin{pmatrix} x_1^4 & x_1^3 & x_1^2 \\ x_1^3 & x_1^2 & x_1 \\ x_1^2 & x_1 & 1 \end{pmatrix}$$

只要能证明Hessian矩阵半正定，就可以证明函数的凸性。

(2) 答：Hessian矩阵可以表示成以下形式

$$D^2E = 2\sum_{x\in D}\begin{pmatrix} x_1^2 \\ x_1 \\ 1 \end{pmatrix}\begin{pmatrix} x_1^2 & x_1 & 1 \end{pmatrix}$$

强凸的条件是Hessian矩阵正定
严格凸的条件是半正定且至少一个特征值大于0
搞混强凸和严格凸的同学仍然会被扣分

题目2：强凸

下面证明强凸性

对于β为一个3维向量，

$$\beta^{\mathrm{T}} D^2 E \beta = 2 \sum_{x \in D} \left(\beta^{\mathrm{T}} \begin{pmatrix} x_1^2 \\ x_1 \\ 1 \end{pmatrix} \right)^2$$

对于任意数据集$D = \{(x_{11}, x_{21}), \dots (x_{1n}, x_{2n})\}$

如果$D^2 E$不强凸，那么存在非零β

$$\beta^{\mathrm{T}} D^2 E \beta = 2 \sum_{i=1}^{n} \left(\beta^{\mathrm{T}} \begin{pmatrix} x_{1i}^2 \\ x_{1i} \\ 1 \end{pmatrix} \right)^2 = 0$$

如果上式等于0，那么非零β对于所有i都有

$$\beta^{\mathrm{T}} \begin{pmatrix} x_{1i}^2 \\ x_{1i} \\ 1 \end{pmatrix} = 0$$

题目2：强凸

即

$$\begin{pmatrix} x_{11}^2 & x_{11} & 1 \\ \dots & \dots & \dots \\ x_{1n}^2 & x_{1n} & 1 \end{pmatrix} \beta = \mathrm{A}\beta = 0$$

根据齐次方程组特性，如果矩阵A满秩则方程组只有0解，这个时候就不存在非零β使得$\beta^{\mathrm{T}} D^2 E \beta = 0$。

不妨设$n = 3$,如果此时矩阵的行列式>0,则秩为3

$\begin{vmatrix} x_{11}^2 & x_{11} & 1 \\ \dots & \dots & \dots \\ x_{13}^2 & x_{13} & 1 \end{vmatrix}$是一个范德蒙行列式，它的解为

$$(x_{13} - x_{12})\ (x_{12} - x_{12})\ (x_{13} - x_{11})$$

基于此我们可以很容易找到一组x_{11}，x_{12}，x_{13}使得A满秩，此时齐次方程组只有0解，Hessian矩阵强凸。

> n阶范德蒙矩阵的行列式可以表示为：
>
> $$\det(V) = \prod_{1 \le i < j \le n} (\alpha_j - \alpha_i)$$
>
> 当α_i各不相同时，$\det(V)$不为零。

题目3：转移概率矩阵

假设第k年选择课程A的学生比例是 x_k，余下 $y_k = 1 - x_k$ 的学生选择课程B。

第k+1年，之前1/5选择课程A的学生改选了课程B，同时1/10选择课程B的学生改选了课程A。

不同的初始概率：

例如：$\lim_{k\to\infty} P^k \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1/3 \\ 2/3 \end{bmatrix}$

从而马尔可夫链模型的状态转移矩阵收敛到的稳定概率分布与初始状态概率分布无关。任意的概率分布样本 ⟶ 对应稳定概率分布的样本

注意：

1）非周期的：状态转化没有循环的，如果循环则不会收敛；

2）任何两个状态连通：任意一个状态可以通过有限步到达其它任意状态，不会出现条件概率一直为0导致不可达的情况

3）状态数既可以有限也可以无限，所以可以用于连续和离散概率分布

题目3：转移概率矩阵（解法2）

所以，P符合马尔可夫链收敛性质：

1. We calculate the eigenvalues of P.
Solve $|P - \lambda I| = 0$, we have the two eigenvalues are $\lambda_1 = 1, \lambda_2 = 7/10$. Solve $Px = \lambda_i x$, we have the corresponding eigenvectors are $x_1 = (1,2)^\top, x_2 = (1,-1)^\top$. Then we have

$$\lim_{k\to\infty} P^k \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix} = \lim_{k\to\infty} \begin{pmatrix} 1 & (\frac{7}{10})^k \cdot 1 \\ 2 & (\frac{7}{10})^k \cdot (-1) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix}$$

Hence

$$\lim_{k\to\infty} P^k [1 \quad 0]^\top = \begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix}^{-1} [1 \quad 0]^\top = (1/3, 2/3)^\top$$

2. Use the convergence theorem of Markov chains:
If (1)the transition matrix P is irreducible, (2) every state in this Markov chain has period 1, and (3) this Markov chain has a stationary distribution π, then

$$\lim_{k\to\infty} P^k(i,j) = \pi(i)$$

Since every entry of P is positive, P is irreducible and aperiodic. Solve

$$\begin{cases} \pi = P\pi \\ \pi(1) + \pi(2) = 1 \end{cases} \tag{3.1}$$

, we have $\pi = (1/3, 2/3)^\top$. Hence

$$\lim_{k\to\infty} P^k [1 \quad 0]^\top = (1/3, 2/3)^\top$$

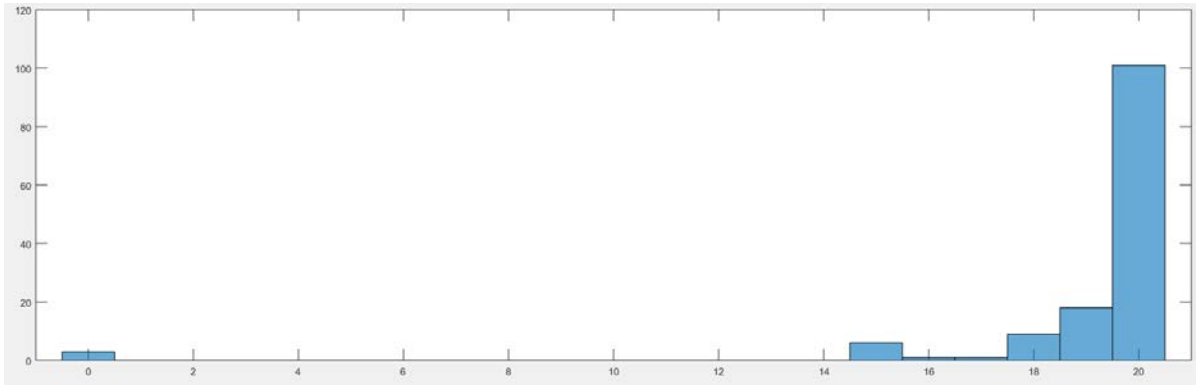
收敛性质：
1) 可能的状态数是有限的
2) 转移概率固定不变
3) 从任意状态能够转变到任意其他状态
4) 不是简单循环，例如不能全是A到B再从B到A

题目3：转移概率矩阵

应用：

- 语音识别：
  声学模型：HMM，语言模型：N-Gram
- PageRank
- 股指建模、时间序列分析
- ……

成绩直方图：



题目 4：假设检验

小明同学共抛硬币 50 次，其中有 35 次正面向上.
问该硬币是否出现正面向上的概率更大?

(1) 设 $\alpha = 0.05$ ，使用 z-test 方法给出必要的共识和计算过程;

(2) 计算 p-value 并解释其含义.

题目4：假设检验

(1)设显著性水平 $\alpha = 0.05$ ，使用 z-test 方法给出必要的共识和计算过程；

答：首先写出零假设(null hypothesis)和备择假设(alternative hypothesis)，如下：

$$H_0 : p \leq 0.5$$
$$H_1 : p > 0.5$$

其中 p 为正面向上的概率. 下面我们将在上述假设的基础上，进行单边 z-test。

通过查表可知，$\alpha = 0.05$ 对应的 临界值 (critical value) 为 $z_c = 1.645$. 并且，我们有 $\hat{p} = \frac{35}{50} = 0.7$ ，因此有 $z = \frac{\hat{p} - p}{\sqrt{p\frac{1-p}{n}}} = 2.828 > 1.645$.

所以，我们拒绝假设 H_0 。

题目 4：假设检验

(2) 计算 p-value 并解释其含义.

答：我们可以计算出 $p = \Pr(z > 2.828) = 0.0023$，由于 $p < \alpha = 0.05$，零假设 H_0 被拒绝.

因此，当显著性水平 $\alpha = 0.05$ 时，由观测到的证据，我们可以认为 该硬币出现正面向上的概率更大.

题目5:评价指标

给定8个样本和两分类器的预测结果如下:

y	1	0	1	1	1	0	0	0
y_{C_1}	0.5	0.3	0.6	0.22	0.4	0.51	0.2	0.33
y_{C_2}	0.04	0.1	0.68	0.22	0.4	0.11	0.8	0.53

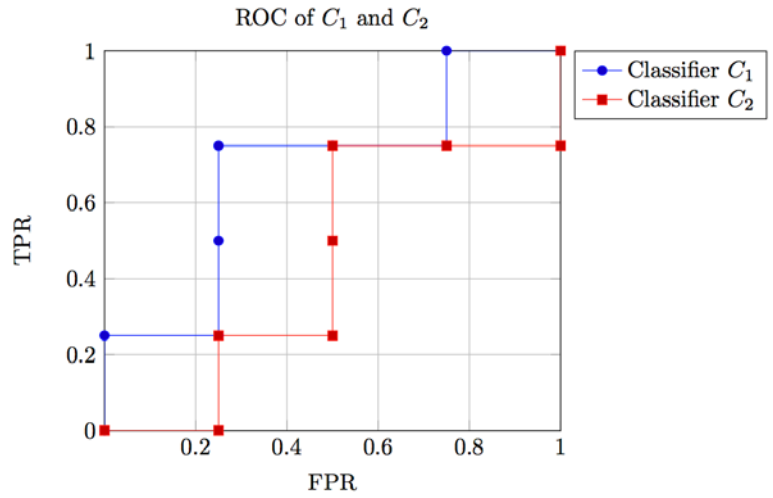
(1)画出分类器$\mathbf{C_1}$和$\mathbf{C_2}$的ROC曲线，并计算相应的AUC。

(2)对分类器$\mathbf{C_1}$，将分类阈值设为0.33，对分类器$\mathbf{C_2}$将分类阈值设为0.1，给出对应的混淆矩阵和F1 Score。

(3)证明课本第35页Eq.(2.22)：AUC = $\mathbf{1-\ell_{rank}}$。

ROC曲线和AUC

(1)答：根据分类器的预测结果对样本进行排序，
依次将每个样本划分为正例，当前若为真正例，则对应的坐标为(x,y+0.25)，若为假正例，则坐标为(x+0.25,y)。



C_1的AUC： $\frac{11}{16}$

C_2的AUC： $\frac{7}{16}$

混淆矩阵和F1 Score

(2)答：将C_1的阈值设为0.33，C_2的阈值设为0.1，根据预测值和真实值容易得到混淆矩阵如下：

C_1的混淆矩阵

	预测结果	
真实情况	正例	反例
正例	3	1
反例	1	3

C_2的混淆矩阵

	预测结果	
真实情况	正例	反例
正例	3	1
反例	3	1

根据混淆矩阵可以算出对应的P和R，然后带入公式即可得到F1分别为：0.75和0.6。

证明: $AUC = 1 - \ell_{rank}$

(3)答：分析 ROC 曲线上方的面积：

1) 对每单位纵向线上方的格子：面积为0

2) 对每单位横向线上方的格子：对应样本为x^-,上方单元格数为 $\sum_{x^+ \in D^+} I(f(x^+) < f(x^-))$。

3) 对每条斜线段上方的格子：上方单元格数为$\sum_{x^+ \in D^+} I(f(x^+) < f(x^-)) + \frac{1}{2} I(f(x^+ = f(x^-))$

AUC 对应的方格数为总数减去曲线上方的单元格数:

$$m^+ m^- - \sum_{x^+ \in D^+} \sum_{x^- \in D^-} I\big(f(x^+) < f(x^-)\big) + \frac{1}{2} I(f\big(x^+ = f(x^-)\big)$$

进行归一化可得$AUC = 1 - \ell_{rank}$，其中$\ell_{rank} = \frac{1}{m^+ m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} I\big(f(x^+) <$

题目6：预测误差的期望

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。)

题目6：预测误差的期望

Proof.

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

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

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

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

综合问题汇总

(1)很多同学都没有将姓名包含在自己的文件或者文件名中，有部分同学学号写错又找不到名字，只能记为0分。

(2)大部分手写同学的板书质量都非常糟糕，看不清的地方都直接扣分，所以请大家不要偷懒。

(3)请同学们注意推导的步骤一定要具体，凡是推导完全省略或是难以理解的回答都会被扣分。

课程QQ群: 891152744