机器学习导论
习题二参考答案

秦天

2021 年 4 月 19 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) 请在LaTeX模板中**第一页填写个人的姓名、学号、邮箱信息**；

(2) 本次作业需提交该pdf文件， pdf文件名格式为**学号_姓名.pdf**， 例如190000001_张三.pdf，**需通过教学立方提交**。

(3) 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**；

(4) 本次作业提交截止时间为**4月16日23:55:00。**

[1] 参考尹一通老师高级算法课程中对学术诚信的说明。

1 [40pts] Linear Discriminant Analysis

课本中介绍的 Fisher 判别分析 (Fisher Discriminant Analysis, FDA) 没有对样本分布进行假设. 当假设各类样本的协方差矩阵相同时, FDA 退化为线性判别分析 (Linear Discriminant Analysis, LDA). 考虑一般的K分类问题, $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}$为训练集, 其中, 第 k 类样本从正态分布 $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ 中独立同分布采样得到 ($k = 1, 2, \cdots, K$, 各类共享协方差矩阵), 记该类样本数量为 m_k, 类概率 $\Pr(y = k) = \pi_k$. 若 $X \in \mathbb{R}^d \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 则其概率密度函数为

$$p(x) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right). \tag{1}$$

请回答下列问题:

(1) **[6pts]** (贝叶斯最优分类器) 从贝叶斯决策论的角度出发, 对样本 $\boldsymbol{x}$ 做出的最优预测应为 $\arg\max_y \Pr(y \mid \boldsymbol{x})$. 因此, 只需考察 $\ln \Pr(y = k \mid \boldsymbol{x})$的大小, 即可得到贝叶斯最优分类器, 这也正是推导LDA的一种思路. 请证明: 在题给假设下, $\arg\max_y \Pr(y \mid \boldsymbol{x}) = \arg\max_k \delta_k(\boldsymbol{x})$, 其中 $\delta_k(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \ln \pi_k$ 为LDA在分类时的判别式.

(2) **[6pts]** 假设 $K = 2$, 记 $\hat{\pi}_k = \frac{m_k}{m}$, $\hat{\boldsymbol{\mu}}_k = \frac{1}{m_k} \sum_{y_i = k} \boldsymbol{x}_i$, $\hat{\boldsymbol{\Sigma}} = \frac{1}{m - K} \sum_{k=1}^{K} \sum_{y_i = k} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)^T$. LDA使用这些经验量替代真实参数, 计算判别式 $\delta_k(\boldsymbol{x})$ 并按照第(1)问中的准则做出预测. 请证明: 在 $\boldsymbol{x}^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) > \frac{1}{2} (\hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1)^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) - \ln(m_2/m_1)$ 时 LDA 将样本预测为第 2 类.

(3) **[16pts]** (线性回归) 考虑第(2)问中的二分类问题, 并将第 1 类样本的标记 y 设为 $-\frac{m}{m_1}$, 将第 2 类样本的标记 y 设为 $\frac{m}{m_2}$. 仿照线性回归, 得到下列优化问题:

$$\min_{\boldsymbol{\beta}, \beta_0} \quad \sum_{i=1}^{m} (y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta})^2. \tag{2}$$

请证明: 上述优化问题的最优解满足 $\boldsymbol{\beta}^* \propto \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$, 即通过线性回归解得的 $\boldsymbol{x}$ 系数与第(2)问中LDA的判别规则表达式中的 $\boldsymbol{x}$ 系数同向.

(4) **[6pts]** (对率回归) 通过课本的介绍可知对率回归假设对数几率为特征 $\boldsymbol{x}$ 的线性函数, 而由第(1)问可知, 在LDA 中, 对数几率 $\ln \frac{\Pr(y=k|\boldsymbol{x})}{\Pr(y=l|\boldsymbol{x})}$ 也可以写成 $\beta_0 + \boldsymbol{x}^T \boldsymbol{\beta}$ 的形式, 从这一角度来看, 这两种模型似乎是相同的? 哪种模型做出的假设更强? 请说明理由.

(5) **[6pts]** (二次判别分析) 假设各类样本仍服从正态分布, 但第 k 类样本从 $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ 中独立同分布采样得到, 即不假设各类的协方差矩阵相同. 请按照第(1)问中的思路, 给出分类应采用的判别式 $\delta_k(\boldsymbol{x})$, 使得 $\arg\max_y \Pr(y \mid \boldsymbol{x}) = \arg\max_k \delta_k(\boldsymbol{x})$. 此时判别式是一个关于$\boldsymbol{x}$的二次函数, 这一做法被称为二次判别分析 (Quadratic Discriminant Analysis, QDA).

Solution. (1)

$$\begin{aligned}
\arg\max_y \Pr(y \mid \boldsymbol{x}) &= \arg\max_y \ln \Pr(y \mid \boldsymbol{x}) \\
&= \arg\max_y \ln \Pr(y) \Pr(\boldsymbol{x} \mid y) \\
&= \arg\max_y \ln \pi_y + \ln \Pr(\boldsymbol{x} \mid y) \\
&= \arg\max_y \ln \pi_y - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_y)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_y) \\
&= \arg\max_y \ln \pi_y + \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_y - \frac{1}{2}\boldsymbol{\mu}_y^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_y \\
&= \arg\max_k \delta_k(\boldsymbol{x}).
\end{aligned}$$

(2) 要预测为第 2 类, 只需 $\delta_2(\boldsymbol{x}) > \delta_1(\boldsymbol{x})$. 令$\hat{\delta}_2(\boldsymbol{x}) > \hat{\delta}_1(\boldsymbol{x})$, 化简整理可得

$$\boldsymbol{x}^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) > \frac{1}{2} (\hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1)^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) - \ln(m_2/m_1).$$

(3) 首先化简 $\hat{\boldsymbol{\Sigma}}$,

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}} &= \frac{1}{m-2} \sum_{k=1}^{2} \sum_{y_i=k} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)^T \\
&= \frac{1}{m-2} \left(\sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^T - m_1 \hat{\boldsymbol{\mu}}_1 \hat{\boldsymbol{\mu}}_1^T - m_2 \hat{\boldsymbol{\mu}}_2 \hat{\boldsymbol{\mu}}_2^T \right).
\end{aligned}$$

对目标函数关于 β_0 和 $\boldsymbol{\beta}$ 分别求导并令其等于0, 可得

$$\sum_{i=1}^{m} (y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta}) = 0, \tag{3}$$

$$\sum_{i=1}^{m} (y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta}) \boldsymbol{x_i} = \mathbf{0}. \tag{4}$$

将 $\boldsymbol{y}$ 的值代入(3), 得 $\beta_0 = -\frac{1}{m} \sum_{i=1}^{m} \boldsymbol{x}_i^T \boldsymbol{\beta}$. 将 β_0 和 $\boldsymbol{y}$ 代入(4), 注意到 $m = m_1 + m_2$, 得

$$\begin{aligned}
m(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) &= \sum_{i=1}^{m} y_i \boldsymbol{x}_i \\
&= \left(-\frac{1}{m} \sum_{i=1}^{m} \boldsymbol{x}_i \sum_{j=1}^{m} \boldsymbol{x}_j^T + \sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^T \right) \boldsymbol{\beta} \\
&= \left(-\frac{1}{m} (m_1 \hat{\boldsymbol{\mu}}_1 + m_2 \hat{\boldsymbol{\mu}}_2)(m_1 \hat{\boldsymbol{\mu}}_1 + m_2 \hat{\boldsymbol{\mu}}_2)^T + \sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^T \right) \boldsymbol{\beta} \\
&= \left((m-2)\hat{\boldsymbol{\Sigma}} + m_1 \hat{\boldsymbol{\mu}}_1 \hat{\boldsymbol{\mu}}_1^T + m_2 \hat{\boldsymbol{\mu}}_2 \hat{\boldsymbol{\mu}}_2^T - \frac{1}{m} (m_1 \hat{\boldsymbol{\mu}}_1 + m_2 \hat{\boldsymbol{\mu}}_2)(m_1 \hat{\boldsymbol{\mu}}_1 + m_2 \hat{\boldsymbol{\mu}}_2)^T \right) \boldsymbol{\beta} \\
&= \left((m-2)\hat{\boldsymbol{\Sigma}} + \frac{m_1 m_2}{m} (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)^T \right) \boldsymbol{\beta}
\end{aligned}$$

因为 $m(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$ 和 $\frac{m_1 m_2}{m} (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)^T \boldsymbol{\beta}$ 的方向均与 $\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1$ 的方向相同, 所以

$$(m-2)\hat{\boldsymbol{\Sigma}} \boldsymbol{\beta} \propto \hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1,$$

从而

$$\boldsymbol{\beta}^* \propto \hat{\boldsymbol{\Sigma}}^{-1}\left(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1\right).$$

点评: 这一题的化简略有些复杂, 只有少数同学完整正确地给出证明, 只要完成了求导步骤, 就可以拿到 4 到 6 分. 很多同学没有对 $\hat{\boldsymbol{\Sigma}}$ 进行化简, 直接给出一个包含 $\hat{\boldsymbol{\Sigma}}$ 与 $\boldsymbol{\beta}$ 的等式 (并且多数是错误的), 这种情况即使等式正确, 也扣除了部分过程分. 还有一些同学直接更改了 $\hat{\boldsymbol{\Sigma}}$ 的定义, 这种情况一般会扣除 6 到 10 分. 另外, 绝大多数同学的符号使用极不规范, 常见情况包括: 省略估计量顶部的 hat 符号、不使用数学字体对标量和向量进行区分、将矩阵逆写为分数形式等, 请多加注意.

(4) LDA 的假设更强. LDA 假设类条件概率服从正态分布, 并且各类共享协方差矩阵, 从而使对数几率为线性函数, 而对率回归仅假设对数几率为线性函数, 所以 LDA 的假设更强. (当样本分布为正态分布时, 通过极大似然估计得到的协方差矩阵即为样本协方差矩阵, 所以两种方法实际上均可看作通过极大似然估计进行求解, 只不过对率回归需要进行迭代优化. 但是, 因为 LDA 需要计算样本均值与协方差, 所以其对离群点更为敏感, 对率回归更为鲁棒.)

点评: 这一题只要求回答 LDA 的假设更强 (3分) 以及原因是额外假设了高斯分布和共享的协方差矩阵 (3分). 有些同学认为 LDA 额外假设了样本是由 i.i.d. 采样得到的, 从而 LDA 假设更强, 而实际上 i.i.d. 是大多数机器学习算法对数据的共同要求, 否则测试集与训练集的分布不同, 无法对泛化性能给出理论保证.

(5)

$$\begin{aligned}
\arg\max_y \ \Pr\left(y \mid \boldsymbol{x}\right) &= \arg\max_y \ \ln \Pr\left(y \mid \boldsymbol{x}\right) \\
&= \arg\max_y \ \ln \Pr(y) \Pr(\boldsymbol{x} \mid y) \\
&= \arg\max_y \ \ln \pi_y + \ln \Pr(\boldsymbol{x} \mid y) \\
&= \arg\max_y \ \ln \pi_y - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_y)^T \boldsymbol{\Sigma}_y^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_y) - \frac{1}{2}\ln\det\left(\boldsymbol{\Sigma}_y\right).
\end{aligned}$$

所以 $\delta_k(\boldsymbol{x}) = \ln \pi_k - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) - \frac{1}{2}\ln\det\left(\boldsymbol{\Sigma}_k\right)$.

点评: 这一题过于简单, 所以只看最后给出的表达式 (6分). 常见错误包括: 漏掉行列式的符号, 正负号错误, 漏掉关于 $\boldsymbol{\Sigma}_k^{-1}$ 的项. 还有部分同学不约而同地将行列式写成了行列式的平方, **请相关同学引以为戒**.

2 [30pts] Generalized Rayleigh Quotient

在面对多类样本时, FDA 需要求解广义瑞利商:

$$\max_{\boldsymbol{w}} \quad \frac{\boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}}. \tag{5}$$

(1) **[15pts]** 请证明: 瑞利商满足

$$\lambda_{\min}(\boldsymbol{A}) \leq \frac{\boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{w}} \leq \lambda_{\max}(\boldsymbol{A}), \tag{6}$$

其中 $\boldsymbol{A}$ 为实对称矩阵, $\lambda(\boldsymbol{A})$ 为 $\boldsymbol{A}$ 的特征值.

(2) **[15pts]** 请证明: 如果 $\boldsymbol{A}$ 为实对称矩阵, $\boldsymbol{B}$ 为正定矩阵, 那么广义瑞利商满足

$$\lambda_{\min}(\boldsymbol{B}^{-1}\boldsymbol{A}) \leq \frac{\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{B}\boldsymbol{w}} \leq \lambda_{\max}(\boldsymbol{B}^{-1}\boldsymbol{A}). \tag{7}$$

Solution. (1) 设 $\boldsymbol{A}\boldsymbol{\xi}_i = \lambda_i\boldsymbol{\xi}_i$, $i = 1, 2, \cdots, n$, $\{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_n\}$ 构成一组单位正交基, 从而存在一组 $\{\alpha_i\}_{i=1}^n$ 使得

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i\boldsymbol{\xi}_i,$$

代入瑞利商的定义, 可得

$$\frac{\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{w}} = \frac{\left(\sum_i \alpha_i\boldsymbol{\xi}_i^T\right)\boldsymbol{A}\left(\sum_i \alpha_i\boldsymbol{\xi}_i\right)}{\left(\sum_i \alpha_i\boldsymbol{\xi}_i^T\right)\left(\sum_i \alpha_i\boldsymbol{\xi}_i\right)} = \frac{\sum_i \alpha_i^2\lambda_i}{\sum_i \alpha_i^2}.$$

可见瑞利商是 $\boldsymbol{A}$ 的特征值的加权平均, 所以

$$\lambda_{\min}(\boldsymbol{A}) \leq \frac{\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{w}} \leq \lambda_{\max}(\boldsymbol{A}).$$

(2) 对 $\boldsymbol{B}$ 进行正交对角化, 可得 $\boldsymbol{B} = \boldsymbol{P}^T\boldsymbol{\Lambda}\boldsymbol{P}$, 其中 $\boldsymbol{P}$ 为特征向量构成的正交矩阵, $\boldsymbol{\Lambda}$ 为对角线元素为相应特征值 (正数) 的对角矩阵. 于是,

$$\frac{\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{B}\boldsymbol{w}} = \frac{\tilde{\boldsymbol{w}}^T\boldsymbol{C}\tilde{\boldsymbol{w}}}{\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{w}}},$$

其中 $\tilde{\boldsymbol{w}} = \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{w}$, $\boldsymbol{C} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}$. 由前一问知

$$\lambda_{\min}(\boldsymbol{C}) \leq \frac{\tilde{\boldsymbol{w}}^T\boldsymbol{C}\tilde{\boldsymbol{w}}}{\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{w}}} \leq \lambda_{\max}(\boldsymbol{C}).$$

下面证明若 λ 是 $\boldsymbol{C}$ 的特征值, 那么其也是 $\boldsymbol{B}^{-1}\boldsymbol{A}$ 的特征值. 考察特征多项式, 可得

$$\begin{aligned}
\det\left(\lambda\boldsymbol{I} - \boldsymbol{C}\right) &= \det\left(\lambda\boldsymbol{I} - \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}\right) \\
&= \det\left(\left(\lambda\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{P} - \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{P}\boldsymbol{A}\right)\boldsymbol{P}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}\right) \\
&= \det\left(\boldsymbol{P}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}\right)\det\left(\lambda\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{P} - \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{P}\boldsymbol{A}\right) \\
&= \det\left(\lambda\boldsymbol{I} - \boldsymbol{P}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{P}\boldsymbol{A}\right) \\
&= \det\left(\lambda\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A}\right),
\end{aligned}$$

因此 $\boldsymbol{C}$ 与 $\boldsymbol{B}^{-1}\boldsymbol{A}$ 有相同的特征值, 所以

$$\lambda_{\min}(\boldsymbol{B^{-1}A}) \leq \frac{\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{B}\boldsymbol{w}} \leq \lambda_{\max}(\boldsymbol{B^{-1}A}).$$

点评: (除参考答案外, 也可以使用拉格朗日乘子法进行证明.) 第二问可以利用第一问的结论, 但需要证明 (或说明) $\boldsymbol{C}$ 与 $\boldsymbol{B}^{-1}\boldsymbol{A}$ 有相同的特征值 (基于特征多项式、矩阵相似等), 没有此步骤的扣 5 分. 另外, 一些同学想当然地使用关于矩阵运算的错误命题或照搬网上的错误资料, 写出如 $\boldsymbol{B}^{-1}\boldsymbol{A} = \boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}$ 的错误等式, 这种情况视前面的证明过程扣 5 到 10 分. 提醒各位同学: 网络资料良莠不齐, 参考需谨慎!

3 [30+10*pts] Decision Tree

(1) **[15pts]** 对于不含冲突样本 (即特征相同但标记不同) 的训练集, 必存在与训练集一致 (即训练误差为 0) 的决策树. 如果训练集可以包含无穷多个样本, 是否一定存在与训练集一致的深度有限的决策树? 证明你的结论. (仅考虑单个划分准则仅包含一次属性判断的决策树)

(2) **[15pts]** 考虑如表1所示的人造数据, 其中“性别”、“喜欢ML作业”是特征, “ML成绩高”是标记. 请画出**所有**可能的使用信息增益为划分准则产生的决策树. (不需要写出计算过程)

表 1: 人造训练集

编号	性别	喜欢ML作业	ML成绩高
1	男	是	是
2	女	是	是
3	男	否	否
4	男	否	否
5	女	否	是

(3) **[10*pts]** 在决策树的生成过程中, 需要计算信息增益以生成新的结点. 设 a 为有V个可能取值$\{a^1, a^2, \cdots, a^V\}$的离散属性, 请证明:

$$\mathrm{Gain}(D, a) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v) \geq 0, \tag{8}$$

即信息增益非负.

Solution. (1) 不一定. 考虑数据集$\{(1/i, (-1)^i)\}_{i=1}^{\infty}$. 有限深度的决策树有有限次属性判断，只能将区间划分为有限个区域, 而该数据集将区间划分为无穷多个区域.

点评: 这一题的证明只要给出一个例子即可. 常见问题有: 证明过程写了很多, 却找不到要证明的结论是什么; 仍然考虑有冲突样本 (题目假设无冲突); 认为样本特征数可以为无穷 (这样的话连一个样本都无法用有限长度描述, 何谈训练模型?); 用词过于口语化, 术语概念不清, 使用混乱. 结论正确的最少可得 7 分, 结论错误的最多可得 5 分.

(2) 结果如图1所示.

(3) 定义随机变量 Y 为样本标记, 其分布即为落到相应决策树结点上的样本的标记的经验分布. 定义随机变量 A 为属性 a 的取值, 其分布即为落到相应决策树结点上的样本的属性 a 的经

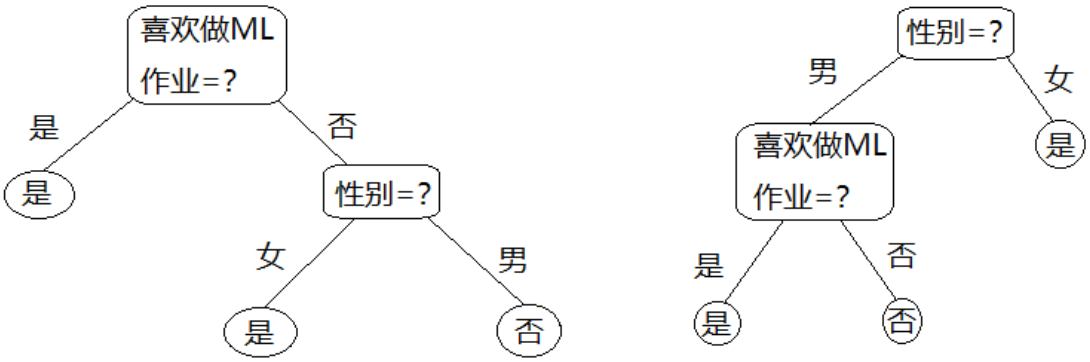

图 1: 决策树

验分布. 用 $H(\cdot)$ 表示信息熵, $I(\cdot;\cdot)$ 表示互信息, 有

$$\begin{aligned}\mathrm{Gain}(D,a) &= \mathrm{Ent}(D) - \sum_{v=1}^{V}\frac{|D^v|}{|D|}\mathrm{Ent}(D^v)\\ &= H(Y) - \sum_{v=1}^{V}\Pr(A=a^v)H(Y|A=a^v)\\ &= H(Y) - H(Y|A)\\ &= I(Y;A)\\ &\geq 0.\end{aligned}$$

点评: 这一题可直接利用信息论的相关结论, 也可定义相关符号化简后使用 Jensen 不等式证明. 有些同学照搬网络资料, 直接将该题等价于证明 KL 散度的非负性, 却不给出信息增益到 KL 散度的转化步骤 (或在不理解的情况下给出错误的关系, 正确关系应为 $\mathrm{Gain} = \mathrm{KL}(p(a,y)\|p(a)p(y))$ 而非 $\mathrm{KL}(p(a)\|p(y))$), 这种情况认为后续证明无效, 0 分处理.

此次作业成绩分布如下: