机器学习导论
习题二

191250026, 丁云翔, 191250026@smail.nju.edu.cn

2021 年 4 月 15 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**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. 此处用于写解答(中英文均可)

(1)

由贝叶斯定理，

$$\begin{aligned}\Pr(y=k\mid\boldsymbol{x}) &= \frac{\Pr(\boldsymbol{x}\mid y=k)\cdot\Pr(y=k)}{\sum_{j=1}^{K}\Pr(\boldsymbol{x}\mid y=j)\cdot\Pr(y=j)}\\ &= \frac{\frac{1}{(2\pi)^{\frac{d}{2}}\det(\boldsymbol{\Sigma})^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)\cdot\pi_k}{\sum_{j=1}^{K}\frac{1}{(2\pi)^{\frac{d}{2}}\det(\boldsymbol{\Sigma})^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right)\cdot\pi_j}\\ &= \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)\cdot\pi_k}{\sum_{j=1}^{K}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right)\cdot\pi_j}\end{aligned}$$

由于分母对任意k均为常数，故在优化目标中可以省略。又取对数不影响优化函数单调性，所以对分子取对数。又取对数后化简得到的$-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$项为常数，故在优化目标中可以省略；另外有$\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k=\frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$(常数的转置等于其自身)，故优化目标最终化简如下：

$$\begin{aligned}\arg\max_y\Pr(y\mid\boldsymbol{x}) &= \arg\max_k\ln\Pr(y=k\mid\boldsymbol{x})\\ &= \arg\max_k\ln(\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)\cdot\pi_k)\\ &= \arg\max_k\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)+\ln\pi_k\right)\\ &= \arg\max_k\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}+\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k+\frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}+-\frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k+\ln\pi_k\right)\\ &= \arg\max_k\left(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k-\frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k+\ln\pi_k\right)\\ &= \arg\max_k\delta_k(\boldsymbol{x})\end{aligned}$$

故在题给假设下，$\arg\max_y\Pr(y\mid\boldsymbol{x})=\arg\max_k\delta_k(\boldsymbol{x})$，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)

将$K=2$和题给经验量代入判别式：

$$\begin{aligned}\delta_k(\boldsymbol{x}) &= \boldsymbol{x}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_k-\frac{1}{2}\hat{\boldsymbol{\mu}}_k^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_k+\ln\hat{\pi}_k\\ &= \boldsymbol{x}^T(\frac{1}{m-2}\sum_{k=1}^{2}\sum_{y_i=k}\left(\boldsymbol{x}_i-\frac{1}{m_k}\sum_{y_i=k}\boldsymbol{x}_i\right)\left(\boldsymbol{x}_i-\frac{1}{m_k}\sum_{y_i=k}\boldsymbol{x}_i\right)^T)^{-1}(\frac{1}{m_k}\sum_{y_i=k}\boldsymbol{x}_i)\\ &\quad-\frac{1}{2}(\frac{1}{m_k}\sum_{y_i=k}\boldsymbol{x}_i)^T(\frac{1}{m-2}\sum_{k=1}^{2}\sum_{y_i=k}\left(\boldsymbol{x}_i-\frac{1}{m_k}\sum_{y_i=k}\boldsymbol{x}_i\right)\left(\boldsymbol{x}_i-\frac{1}{m_k}\sum_{y_i=k}\boldsymbol{x}_i\right)^T)^{-1}(\frac{1}{m_k}\sum_{y_i=k}\boldsymbol{x}_i)\\ &\quad+\ln\frac{m_k}{m}\end{aligned}$$

故由第(1)问中的准则，预测样本$\boldsymbol{x}$的类别为$\arg\max_k\delta_k(\boldsymbol{x})$，即若$\delta_1(\boldsymbol{x})>\delta_2(\boldsymbol{x})$预测为第一类，若$\delta_2(\boldsymbol{x})>\delta_1(\boldsymbol{x})$预测为第二类，若$\delta_1(\boldsymbol{x})=\delta_2(\boldsymbol{x})$预测为第一类或第二类。

由题给条件$\boldsymbol{x}^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_2-\hat{\boldsymbol{\mu}}_1)>\frac{1}{2}(\hat{\boldsymbol{\mu}}_2+\hat{\boldsymbol{\mu}}_1)^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_2-\hat{\boldsymbol{\mu}}_1)-\ln(m_2/m_1)$可得：

$$\begin{aligned}\boldsymbol{x}^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_2-\hat{\boldsymbol{\mu}}_1)&>\frac{1}{2}(\hat{\boldsymbol{\mu}}_2+\hat{\boldsymbol{\mu}}_1)^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_2-\hat{\boldsymbol{\mu}}_1)-\ln(m_2/m_1)\\ \boldsymbol{x}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_2-\frac{1}{2}\hat{\boldsymbol{\mu}}_2^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_2+\ln\frac{m_2}{m}&>\boldsymbol{x}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_1-\frac{1}{2}\hat{\boldsymbol{\mu}}_1^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_1+\ln\frac{m_1}{m}\\ \boldsymbol{x}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_2-\frac{1}{2}\hat{\boldsymbol{\mu}}_2^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_2+\ln\hat{\pi}_2&>\boldsymbol{x}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_1-\frac{1}{2}\hat{\boldsymbol{\mu}}_1^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_1+\ln\hat{\pi}_1\\ \delta_2(\boldsymbol{x})&>\delta_1(\boldsymbol{x})\end{aligned}$$

即有$\arg\max_k\delta_k(\boldsymbol{x})=2$，故此时LDA将样本预测为第二类。

所以在 $\boldsymbol{x}^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_2-\hat{\boldsymbol{\mu}}_1)>\frac{1}{2}(\hat{\boldsymbol{\mu}}_2+\hat{\boldsymbol{\mu}}_1)^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_2-\hat{\boldsymbol{\mu}}_1)-\ln(m_2/m_1)$ 时 LDA 将样本预测为第 2 类.

(3)

令$\mathbf{E}_{\boldsymbol{\beta},\beta_0}=\sum_{i=1}^m(y_i-\beta_0-\boldsymbol{x}_i^T\boldsymbol{\beta})^2$,并让$\mathbf{E}_{\boldsymbol{\beta},\beta_0}$对$\boldsymbol{\beta},\beta_0$分别求导：

$$\frac{\partial\mathbf{E}_{\boldsymbol{\beta},\beta_0}}{\partial\boldsymbol{\beta}}=\sum_{i=1}^m(\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\beta}+\boldsymbol{x}_i\beta_0-\boldsymbol{x}_iy_i)=0 \tag{3}$$

$$\frac{\partial\mathbf{E}_{\boldsymbol{\beta},\beta_0}}{\partial\boldsymbol{\beta}}=\sum_{i=1}^m(\boldsymbol{x}_i^T\boldsymbol{\beta}+\beta_0-y_i)=0 \tag{4}$$

又$\sum_{i=1}^m\boldsymbol{x}_iy_i=m_1\hat{\boldsymbol{\mu}}_1(-\frac{m}{m_1})+m_2\hat{\boldsymbol{\mu}}_2\frac{m}{m_2}=m(\hat{\boldsymbol{\mu}}_2-\hat{\boldsymbol{\mu}}_1)$，故由(3)式可得：

$$\sum_{i=1}^m(\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\beta}^*+\boldsymbol{x}_i\beta_0^*)=m(\hat{\boldsymbol{\mu}}_2-\hat{\boldsymbol{\mu}}_1) \tag{5}$$

又$y_1=-\frac{m}{m_1}$，$y_2=\frac{m}{m_2}$，有$\sum_{i=1}^m y_i=m_1(-\frac{m}{m_1})+m_2\frac{m}{m_2}=0$，且$\sum_{i=1}^m\beta_0=m\beta_0$，故由(4)式可得：

$$\begin{aligned}\sum_{i=1}^m\boldsymbol{x}_i^T\boldsymbol{\beta}^*+m\beta_0^*&=0\\ \beta_0^*&=-\frac{1}{m}\sum_{i=1}^m\boldsymbol{x}_i^T\boldsymbol{\beta}^*\end{aligned} \tag{6}$$

将(6)式代入(5)式，得：

$$\begin{aligned}\sum_{i=1}^m(\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\beta}^*+\boldsymbol{x}_i(-\frac{1}{m}\sum_{i=1}^m\boldsymbol{x}_i^T\boldsymbol{\beta}^*))&=m(\hat{\boldsymbol{\mu}}_2-\hat{\boldsymbol{\mu}}_1)\\ \sum_{i=1}^m(\boldsymbol{x}_i\boldsymbol{x}_i^T-\frac{1}{m}\boldsymbol{x}_i\sum_{i=1}^m\boldsymbol{x}_i^T)\boldsymbol{\beta}^*&=m(\hat{\boldsymbol{\mu}}_2-\hat{\boldsymbol{\mu}}_1)\end{aligned} \tag{7}$$

同样定义“类内散度矩阵”$\boldsymbol{S}_w$和“类间散度矩阵”$\boldsymbol{S}_b$：

$$\begin{aligned}\boldsymbol{S}_w &= \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 \\ &= \sum_{y_i=1}(\boldsymbol{x}_i - \boldsymbol{\mu}_1)(\boldsymbol{x}_i - \boldsymbol{\mu}_1)^T + \sum_{y_i=2}(\boldsymbol{x}_i - \boldsymbol{\mu}_2)(\boldsymbol{x}_i - \boldsymbol{\mu}_2)^T \\ &= 2\hat{\boldsymbol{\Sigma}} \\ &= \sum_{y_i=1}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1)^T + \sum_{y_i=2}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_2)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_2)^T \\ \boldsymbol{S}_b &= (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \\ &= (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)^T\end{aligned}$$

将(7)式和上述定义式展开计算后发现有如下关系：

$$\sum_{i=1}^{m}(\boldsymbol{x}_i\boldsymbol{x}_i^T - \frac{1}{m}\boldsymbol{x}_i\sum_{i=1}^{m}\boldsymbol{x}_i^T) = \boldsymbol{S}_w + \frac{m_1 m_2}{m}\boldsymbol{S}_b \tag{8}$$

将(8)式代入(7)式：

$$(\boldsymbol{S}_w + \frac{m_1 m_2}{m}\boldsymbol{S}_b)\boldsymbol{\beta}^* = m(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$$

注意到$\frac{m_1 m_2}{m}\boldsymbol{S}_b\boldsymbol{\beta}^*$的方向恒为$\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1$，故$\boldsymbol{S}_w\boldsymbol{\beta}^*$的方向也应该为$\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1$，故有

$$\begin{aligned}\boldsymbol{S}_w\boldsymbol{\beta}^* &\propto (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) \\ 2\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta}^* &\propto (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) \\ \boldsymbol{\beta}^* &\propto \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)\end{aligned}$$

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)

两种模型并不相同。虽然二者形式一致，但模型参数的求解方法不同，并且LDA在求解参数时还考虑了每个类别的类内样本分布情况，对率回归模型则没有，所以可能会导致参数不相同，从而使模型不相同。而且模型自身做出的假设也并不相同，根据定义，LDA假设了各类样本的协方差矩阵相同且满秩，而对率回归则无需事先假设数据分布。故LDA做出的假设更强。

(5)

同样由贝叶斯定理，

$$\begin{aligned}\Pr(y=k \mid \boldsymbol{x}) &= \frac{\Pr(\boldsymbol{x} \mid y=k)\cdot\Pr(y=k)}{\sum_{j=1}^{K}\Pr(\boldsymbol{x} \mid y=j)\cdot\Pr(y=j)} \\ &= \frac{\frac{1}{(2\pi)^{\frac{d}{2}}\det(\boldsymbol{\Sigma}_k)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)\cdot\pi_k}{\sum_{j=1}^{K}\frac{1}{(2\pi)^{\frac{d}{2}}\det(\boldsymbol{\Sigma}_j)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right)\cdot\pi_j} \\ &= \frac{\frac{1}{\det(\boldsymbol{\Sigma}_k)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)\cdot\pi_k}{\sum_{j=1}^{K}\frac{1}{\det(\boldsymbol{\Sigma}_j)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right)\cdot\pi_j}\end{aligned}$$

由于分母对任意k均为常数，故在优化目标中可以省略。又取对数不影响优化函数单调性，所以对分子取对数。另外有$\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k = \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{x}$(常数的转置等于其自身)，故优化目标最终

化简如下：

$$\begin{aligned}
\arg\max_y \Pr(y \mid \boldsymbol{x}) &= \arg\max_k \ln \Pr(y=k \mid \boldsymbol{x}) \\
&= \arg\max_k \ln \left(\frac{1}{\det(\boldsymbol{\Sigma}_k)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) \right) \cdot \pi_k \right) \\
&= \arg\max_k \left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) + \ln \pi_k - \ln\left(\det(\boldsymbol{\Sigma}_k)^{\frac{1}{2}} \right) \right) \\
&= \arg\max_k \left(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{\Sigma}_k^{-1}\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k + \frac{1}{2}\boldsymbol{\mu}_k^T \boldsymbol{\Sigma}_k^{-1}\boldsymbol{x} + -\frac{1}{2}\boldsymbol{\mu}_k^T \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k \right. \\
&\quad \left. + \ln \pi_k - \ln\left(\det(\boldsymbol{\Sigma}_k)^{\frac{1}{2}} \right) \right) \\
&= \arg\max_k \left(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{\Sigma}_k^{-1}\boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k \right. \\
&\quad \left. + \ln \pi_k - \ln\left(\det(\boldsymbol{\Sigma}_k)^{\frac{1}{2}} \right) \right) \\
&= \arg\max_k \delta_k(\boldsymbol{x})
\end{aligned}$$

故此时$\delta_k(\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{x}^T \boldsymbol{\Sigma}_k^{-1}\boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k + \ln \pi_k - \ln\left(\det(\boldsymbol{\Sigma}_k)^{\frac{1}{2}} \right)$ 为QDA在分类时的判别式。

2 [30pts] Generalized Rayleigh Quotient

在面对多类样本时, FDA 需要求解广义瑞利商:

$$\max_{\boldsymbol{w}} \quad \frac{\boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}}. \tag{9}$$

(1) **[15pts]** 请证明: 瑞利商满足

$$\lambda_{\min}(\boldsymbol{A}) \le \frac{\boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{w}} \le \lambda_{\max}(\boldsymbol{A}), \tag{10}$$

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

$$\lambda_{\min}(\boldsymbol{B}^{-1}\boldsymbol{A}) \le \frac{\boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{B} \boldsymbol{w}} \le \lambda_{\max}(\boldsymbol{B}^{-1}\boldsymbol{A}). \tag{11}$$

Solution. 此处用于写解答(中英文均可)

(1)

由$\boldsymbol{A}$为实对称矩阵，可知存在矩阵$\boldsymbol{P}$满足$\boldsymbol{P}^T = \boldsymbol{P}^{-1}$，使$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{C}\boldsymbol{P}^T$，其中$\boldsymbol{C} = diag(\lambda_1, \lambda_2, ..., \lambda_n)$，$\lambda_1, \lambda_2, ..., \lambda_n$为$\boldsymbol{A}$的特征值。故(10)式可改写为：

$$\begin{aligned}
\frac{\boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{w}} &= \frac{\boldsymbol{w}^T \boldsymbol{P}\boldsymbol{C}\boldsymbol{P}^T \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{w}} \\
&= \frac{(\boldsymbol{P}^T \boldsymbol{w})^T \boldsymbol{C} (\boldsymbol{P}^T \boldsymbol{w})}{\boldsymbol{w}^T \boldsymbol{w}}
\end{aligned} \tag{12}$$

由于有$\boldsymbol{P}^T = \boldsymbol{P}^{-1}$，所以有$\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{E}$，记$\boldsymbol{P}$的第$i$列为$\boldsymbol{P}_i$，则有$\sum_{i=1}^n \boldsymbol{P}_i\boldsymbol{P}_i^T = \boldsymbol{E}$。又$\boldsymbol{C}$为对角阵，故：

$$\begin{aligned}(\boldsymbol{P}^T\boldsymbol{w})^T\boldsymbol{C}(\boldsymbol{P}^T\boldsymbol{w}) &= \sum_{i=1}^n \lambda_i(\boldsymbol{P}_i^T\boldsymbol{w})^T(\boldsymbol{P}_i^T\boldsymbol{w}) \\ &= \sum_{i=1}^n \lambda_i \boldsymbol{w}^T\boldsymbol{P}_i\boldsymbol{P}_i^T\boldsymbol{w} \end{aligned} \tag{13}$$

因为$\boldsymbol{w}^T\boldsymbol{P}_i\boldsymbol{P}_i^T\boldsymbol{w} = (\boldsymbol{P}_i^T\boldsymbol{w})^2 > 0$，所以有以下不等式成立：

$$\sum_{i=1}^n \lambda_{\min}(\boldsymbol{A})\boldsymbol{w}^T\boldsymbol{P}_i\boldsymbol{P}_i^T\boldsymbol{w} \le \sum_{i=1}^n \lambda_i \boldsymbol{w}^T\boldsymbol{P}_i\boldsymbol{P}_i^T\boldsymbol{w} \le \sum_{i=1}^n \lambda_{\max}(\boldsymbol{A})\boldsymbol{w}^T\boldsymbol{P}_i\boldsymbol{P}_i^T\boldsymbol{w} \tag{14}$$

又

$$\sum_{i=1}^n \lambda_{\min}(\boldsymbol{A})\boldsymbol{w}^T\boldsymbol{P}_i\boldsymbol{P}_i^T\boldsymbol{w} = \lambda_{\min}(\boldsymbol{A})\boldsymbol{w}^T\sum_{i=1}^n \boldsymbol{P}_i\boldsymbol{P}_i^T\boldsymbol{w} = \lambda_{\min}(\boldsymbol{A})\boldsymbol{w}^T\boldsymbol{w} \tag{15}$$

$$\sum_{i=1}^n \lambda_{\max}(\boldsymbol{A})\boldsymbol{w}^T\boldsymbol{P}_i\boldsymbol{P}_i^T\boldsymbol{w} = \lambda_{\max}(\boldsymbol{A})\boldsymbol{w}^T\sum_{i=1}^n \boldsymbol{P}_i\boldsymbol{P}_i^T\boldsymbol{w} = \lambda_{\max}(\boldsymbol{A})\boldsymbol{w}^T\boldsymbol{w} \tag{16}$$

将(15)式和(16)式代入(14)式得：

$$\begin{aligned}\lambda_{\min}(\boldsymbol{A})\boldsymbol{w}^T\boldsymbol{w} \le \sum_{i=1}^n \lambda_i \boldsymbol{w}^T\boldsymbol{P}_i\boldsymbol{P}_i^T\boldsymbol{w} &\le \lambda_{\max}(\boldsymbol{A})\boldsymbol{w}^T\boldsymbol{w} \\ \lambda_{\min}(\boldsymbol{A}) \le \frac{\sum_{i=1}^n \lambda_i \boldsymbol{w}^T\boldsymbol{P}_i\boldsymbol{P}_i^T\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{w}} &\le \lambda_{\max}(\boldsymbol{A}) \end{aligned} \tag{17}$$

将(12)式和(13)式代入(17)式得：

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

证明完毕。

(2)

由于$\boldsymbol{B}$为正定矩阵，故$\boldsymbol{B}$为正定的实对称矩阵，$\boldsymbol{B}^{-\frac{1}{2}}$存在且$(\boldsymbol{B}^{-\frac{1}{2}})^2 = \boldsymbol{B}^{-1}$。令$\boldsymbol{w} = \boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{x}$，证明(11)式即证明：

$$\begin{aligned}\lambda_{\min}(\boldsymbol{B}^{-1}\boldsymbol{A}) \le \frac{(\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{x})^T\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{x}}{(\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{x})^T\boldsymbol{B}\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{x}} &\le \lambda_{\max}(\boldsymbol{B}^{-1}\boldsymbol{A}) \\ \lambda_{\min}(\boldsymbol{B}^{-1}\boldsymbol{A}) \le \frac{\boldsymbol{x}^T\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{x}} &\le \lambda_{\max}(\boldsymbol{B}^{-1}\boldsymbol{A}) \end{aligned} \tag{18}$$

下面分析$\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{B}^{-\frac{1}{2}}$，由于$\boldsymbol{B}^{-\frac{1}{2}}$也为实对称矩阵，故存在可逆矩阵$\boldsymbol{Q}$，使得$\boldsymbol{B}^{-\frac{1}{2}} = \boldsymbol{Q}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{Q}^{-1}$，其中$\boldsymbol{D} = diag(\lambda_1, \lambda_2, ..., \lambda_n)$，$\lambda_1, \lambda_2, ..., \lambda_n$为$\boldsymbol{B}$的特征值。

故$\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{B}^{-\frac{1}{2}} = \boldsymbol{Q}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{Q}^{-1}\left(\left(\boldsymbol{Q}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{Q}^{-1}\right)^{-1}\left(\boldsymbol{Q}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{Q}^{-1}\right)^{-1}\right)\boldsymbol{Q}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{Q}^{-1} = \boldsymbol{E}$，故(18)式等价于：

$$\lambda_{\min}(\boldsymbol{B}^{-1}\boldsymbol{A}) \le \frac{\boldsymbol{x}^T\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} \le \lambda_{\max}(\boldsymbol{B}^{-1}\boldsymbol{A}) \tag{19}$$

同第(1)问的条件，$\boldsymbol{A}$为实对称矩阵，可知存在矩阵$\boldsymbol{P}$满足$\boldsymbol{P}^T = \boldsymbol{P}^{-1}$，使$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{C}\boldsymbol{P}^T$，其中$\boldsymbol{C} = diag(\lambda_1, \lambda_2, ..., \lambda_n)$，$\lambda_1, \lambda_2, ..., \lambda_n$为$\boldsymbol{A}$的特征值。所以$\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}} = \boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{P}\boldsymbol{C}\boldsymbol{P}^T\boldsymbol{B}^{-\frac{1}{2}}$，又$\boldsymbol{P}^T\boldsymbol{B}^{-\frac{1}{2}} = (\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{P})^T$，故$\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}$也为实对称矩阵。因此，应用第(1)问中的结论，有如下结果成立：

$$\lambda_{\min}(\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}) \leq \frac{\boldsymbol{x}^T\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} \leq \lambda_{\max}(\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}) \tag{20}$$

设λ为矩阵$\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}$的特征值，$\xi$为矩阵$\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}$的特征向量，则有：

$$\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}\xi = \lambda\xi \tag{21}$$

对(21)式等号两边同时左乘$\boldsymbol{B}^{-\frac{1}{2}}$：

$$\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}\xi = \boldsymbol{B}^{-\frac{1}{2}}\lambda\xi$$
$$\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}\xi = \lambda\boldsymbol{B}^{-\frac{1}{2}}\xi$$

即λ也为$\boldsymbol{B}^{-1}\boldsymbol{A}$的特征值。又$\boldsymbol{B}^{-1}\boldsymbol{A}$与$\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}$维度相同，故二者的特征值相同。所以有：

$$\lambda_{\min}(\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}) = \lambda_{\min}(\boldsymbol{B}^{-1}\boldsymbol{A}) \tag{22}$$
$$\lambda_{\max}(\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}) = \lambda_{\max}(\boldsymbol{B}^{-1}\boldsymbol{A}) \tag{23}$$

将(22)式和(23)式代入(20)式，即证明了(19)式：

$$\lambda_{\min}(\boldsymbol{B}^{-1}\boldsymbol{A}) \leq \frac{\boldsymbol{x}^T\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} \leq \lambda_{\max}(\boldsymbol{B}^{-1}\boldsymbol{A})$$

又(19)式等价于(18)式等价于(11)式，故：

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

证明完毕。

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)

(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:

$$\mathrm{Gain}(D, a) = \mathrm{Ent}(D) - \sum_{v=1}^{V}\frac{|D^v|}{|D|}\mathrm{Ent}(D^v) \geq 0, \tag{24}$$

即信息增益非负.

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

Solution. 此处用于写解答(中英文均可)

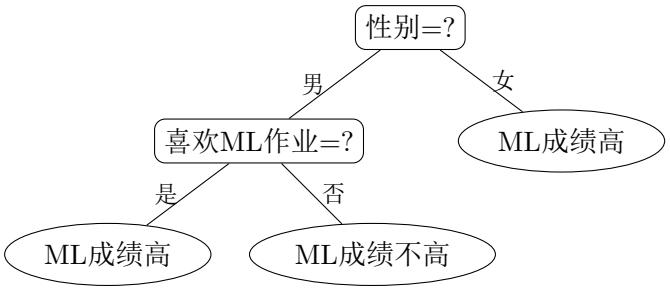
(1)

一定存在。假设训练集中不含冲突样本，且每一次属性划分都将该属性的属性值不同的样本完全分开，虽然由于样本数目可以是无穷大，每次划分可能会产生无穷多个分支，但能够满足将该属性的属性值不同的样本完全分开，则到最大深度后每一个叶结点对应唯一一个样本或所有属性和标记都相同的多个样本，叶结点的标记则与样本自身的标记一致，因此训练误差为0，决策树与训练集一致。又样本特征的数量有限，若每一次属性划分都将该属性的属性值不同的样本完全分开，则决策树深度的最大值为属性的数量，故决策树深度有限。所以存在与训练集一致的深度有限的决策树。

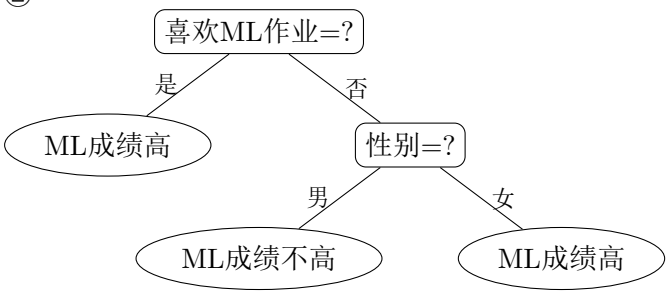
(2)

所有可能的使用信息增益为划分准则产生的决策树共两种，如下图：

①



②



(3)

设K为标记的类别数，将$\mathrm{Ent}(D)$的定义$\mathrm{Ent}(D) = -\sum_{k=1}^{K} \boldsymbol{p}_k \log_2 \boldsymbol{p}_k$代入(24)得：

$$\mathrm{Gain}(D,a) = (-\sum_{k=1}^{K} \boldsymbol{p}_k \log_2 \boldsymbol{p}_k) - \sum_{v=1}^{V} \frac{|D^v|}{|D|}(-\sum_{k=1}^{K} \boldsymbol{p}_{vk} \log_2 \boldsymbol{p}_{vk}) \tag{25}$$

由于K和V均为常数，故求和次序可交换，(25)式可改写为：

$$\begin{aligned}\mathrm{Gain}(D,a) &= (-\sum_{k=1}^{K}\boldsymbol{p}_k\log_2\boldsymbol{p}_k)+\sum_{k=1}^{K}\sum_{v=1}^{V}\frac{|D^v|}{|D|}\boldsymbol{p}_{vk}\log_2\boldsymbol{p}_{vk}\\ &= \sum_{k=1}^{K}(\sum_{v=1}^{V}\frac{|D^v|}{|D|}\boldsymbol{p}_{vk}\log_2\boldsymbol{p}_{vk}-\boldsymbol{p}_k\log_2\boldsymbol{p}_k)\end{aligned} \tag{26}$$

又由相关参数的定义，我们有$\sum_{k=1}^{K}\boldsymbol{p}_k=1$，$\sum_{v=1}^{V}\frac{|D^v|}{|D|}=1$，$\sum_{k=1}^{K}\sum_{v=1}^{V}\frac{|D^v|}{|D|}\boldsymbol{p}_{vk}=1$，且$\boldsymbol{p}_k=\sum_{v=1}^{V}\frac{|D^v|}{|D|}\boldsymbol{p}_{vk}$。设$\mathrm{F}(\boldsymbol{p})=\boldsymbol{p}\log_2\boldsymbol{p}$，因为$0\le\boldsymbol{p}\le1$，结合$\mathrm{F}(\boldsymbol{p})$的图像，可知$\mathrm{F}(\boldsymbol{p})$在$[0,1]$为凸函数。由琴生不等式，得：

$$\begin{aligned}\mathrm{F}(\boldsymbol{p}_k)=\mathrm{F}(\sum_{v=1}^{V}\frac{|D^v|}{|D|}\boldsymbol{p}_{vk})&\le\sum_{v=1}^{V}\frac{|D^v|}{|D|}\mathrm{F}(\boldsymbol{p}_{vk})\\ \boldsymbol{p}_k\log_2\boldsymbol{p}_k=\sum_{v=1}^{V}\frac{|D^v|}{|D|}\boldsymbol{p}_{vk}\log_2(\sum_{v=1}^{V}\frac{|D^v|}{|D|}\boldsymbol{p}_{vk})&\le\sum_{v=1}^{V}\frac{|D^v|}{|D|}\boldsymbol{p}_{vk}\log_2\boldsymbol{p}_{vk}\end{aligned} \tag{27}$$

由(27)式可得：

$$\sum_{v=1}^{V}\frac{|D^v|}{|D|}\boldsymbol{p}_{vk}\log_2\boldsymbol{p}_{vk}-\boldsymbol{p}_k\log_2\boldsymbol{p}_k\ge0 \tag{28}$$

将(28)式代入(26)式得：

$$\mathrm{Gain}(D,a)=\sum_{k=1}^{K}(\sum_{v=1}^{V}\frac{|D^v|}{|D|}\boldsymbol{p}_{vk}\log_2\boldsymbol{p}_{vk}-\boldsymbol{p}_k\log_2\boldsymbol{p}_k)\ge0 \tag{29}$$

即信息增益非负，证明完毕。

4 参考文献

(1) 《机器学习》周志华著

(2) 《模式识别》吴建鑫著

(3) 《高等代数》北京大学数学系前代数小组编

(4) 《概率论与数理统计》盛骤等编

(5) 《Convex Optimization》by Stephen Boyd and Lieven Vandenberghe