机器学习导论
习题二

191300020, 黄彦骁, AdrianHuang@smail.nju.edu.cn

2021 年 4 月 16 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 (**??**) 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 (**??**) 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 (**??**) 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 (**??**) 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 (**??**) 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)

$$\begin{aligned}
\arg\max_y \Pr(y \mid \boldsymbol{x}) &= \arg\max_y \ln \Pr(y = k \mid x) \\
&= \arg\max_k \ln \frac{P(X = x|Y = k) * P(Y = k)}{P(X = x)} \\
&= \arg\max_k \ln \frac{P_k(x) \cdot \pi_k}{P(X = x)} \\
&= \arg\max_k \ln P_k(x) \cdot \pi_k \\
&= \arg\max_k \ln \frac{1}{(2\pi)^{\frac{d}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} - \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) + \ln \pi_k \\
&= \arg\max_k \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \ln \pi_k \\
&= \arg\max_k \delta_k(\boldsymbol{x})
\end{aligned}$$

(2) 要说明 LDA 将样本预测为第 2 类，即只需说明 $\delta_2(\boldsymbol{x}) > \delta_1(\boldsymbol{x})$ 即可：

$$\begin{aligned}
\delta_2(\boldsymbol{x}) > \delta_1(\boldsymbol{x}) &= \boldsymbol{x}^\top \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_2 - \frac{1}{2} \hat{\boldsymbol{\mu}}_2^\top \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_2 + \ln \pi_2 - \boldsymbol{x}^\top \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_1 + \frac{1}{2} \hat{\boldsymbol{\mu}}_1^\top \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_1 - \ln \pi_1 \\
&= \boldsymbol{x}^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) > \frac{1}{2} (\hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1)^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) - \ln (m_2/m_1) \\
&> 0
\end{aligned}$$

故可以证得。

(3) 令 $f = \sum_{i=1}^m (y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta})^2$，对其求偏导得:

$$\frac{\partial f}{\partial \beta_0} = -2 \sum_{i=1}^m \left(y_i - \beta_0 - \boldsymbol{x}_i^\top \boldsymbol{\beta} \right)$$

$$\frac{\partial f}{\partial \boldsymbol{\beta}} = -2 \sum_{i=1}^m \left(y_i - \beta_0 - \boldsymbol{x}_i^\top \boldsymbol{\beta} \right)^\top \boldsymbol{x}_i^\top$$

为求其最优解，令其偏导等于零可得：

$$\beta_0 = \frac{1}{m} \sum_{i=1}^m (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2$$

由于 $\sum_{i=1}^m y_i = m_1 * (-\frac{m}{m_1}) + m_2 * \frac{m}{m_2} = 0$，所以有：

$$\beta_0 = -\frac{m_1 \hat{\boldsymbol{\mu}_1} + m_2 \hat{\boldsymbol{\mu}}_2}{m} \boldsymbol{\beta} = -\hat{\boldsymbol{\mu}} \boldsymbol{\beta}$$

代入原式得：

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

对 $\boldsymbol{\beta}$ 求导得：

$$\frac{\partial f}{\partial \boldsymbol{\beta}} = 2(m-2)\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - 2m(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$$

令其等于可以得到：

$$\boldsymbol{\beta} = \frac{m}{m-2}\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$$

所以有 $\beta^* \propto \hat{\Sigma}^{-1}(\hat{\mu}^1 - \hat{\mu}_2)$.

(4) 我认为是 LDA 模型更强，对数几率回归中是对 y 进行了类后验概率 $p(y=i|x)$ 的假设，而 LDA 中需要我们对理论参数进行估计，用样本参量代入理论模型来进行模拟，样本在估计时的参与率更高。

(5) 与第一题类似：

$$\begin{aligned} \arg\max_y \Pr(y \mid \boldsymbol{x}) &= \arg\max_y \ln \Pr(y = k \mid x) \\ &= \arg\max_k \ln \frac{P(X = x|Y = k) * P(Y = k)}{P(X = x)} \\ &= \arg\max_k \ln \frac{P_k(x) \cdot \pi_k}{P(X = x)} \\ &= \arg\max_k \ln P_k(x) \cdot \pi_k \\ &= \arg\max_k \ln \frac{1}{(2\pi)^{\frac{d}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} - \frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{\mu}_k\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}_k\right) + \ln \pi_k \\ &= \arg\max_k -\frac{1}{2}\ln(|\boldsymbol{\Sigma}|_2^2) + \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln \pi_k \end{aligned}$$

即可得到。

2 [30pts] Generalized Rayleigh Quotient

在面对多类样本时, FDA 需要求解广义瑞利商:

$$\max_{\boldsymbol{w}} \quad \frac{\boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}}. \tag{3}$$

(1) **[15pts]** 请证明: 瑞利商满足

$$\lambda_{\min}(\boldsymbol{A}) \leq \frac{\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{w}} \leq \lambda_{\max}(\boldsymbol{A}), \tag{4}$$

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

$$\lambda_{\min}(\boldsymbol{B}^{-1}\boldsymbol{A}) \leq \frac{\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{B}\boldsymbol{w}} \leq \lambda_{\max}(\boldsymbol{B}^{-1}\boldsymbol{A}). \tag{5}$$

Solution. (1) 由于 A 为实对称矩阵，不妨设其维度为 n，设其特征值为 $\lambda_1, \lambda_2 \ldots \lambda_n$ 同时有：

$$\lambda_{min} \leqslant \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n \leqslant \lambda_{max}$$

所以存在正定矩阵使得 $\boldsymbol{U}$ 和对角矩阵 $\boldsymbol{M}$ 使得 $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^T$,其中 $M = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$, 将上式代入瑞利商得：

$$\frac{\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{w}} = \frac{(\boldsymbol{U}^T\boldsymbol{w})^T\boldsymbol{M}(\boldsymbol{U}^T\boldsymbol{w})}{\boldsymbol{w}^T\boldsymbol{w}}$$

设 $\boldsymbol{p} = \boldsymbol{U}^T\boldsymbol{w}$，同时 p_i 为 $\boldsymbol{p}$ 的第 i 个元素，那么上式即可写为：

$$\frac{\boldsymbol{p}^T\boldsymbol{M}\boldsymbol{p}}{\boldsymbol{w}^T\boldsymbol{w}} = \frac{\sum_{i=1}^n \lambda_i |p_i|^2}{\sum_{i=1}^n |w_i|^2}$$

由特征值的大小关系有：

$$\frac{\lambda_1 \sum_{i=1}^n |p_i|^2}{\sum_{i=1}^n |w_i|^2} \leqslant \frac{\sum_{i=1}^n \lambda_i |p_i|^2}{\sum_{i=1}^n |w_i|^2} \leqslant \frac{\lambda_n \sum_{i=1}^n |p_i|^2}{\sum_{i=1}^n |w_i|^2}$$

设 $\boldsymbol{U}$ 的第 i 行第 j 列的元素为 u_{ij}, 那么:

$$p_i = \sum_{j=1}^n u_{ij}x_j, p_i^T = \sum_{j=1}^n x_j u_{ij}$$

$$|p_i|^2 = p_i^T p_i = \sum_{j=1}^n \sum_{k=1}^n x_j u_{ij} u_{ki} x_k$$

于是有：

$$\sum_{i=1}^n |p_i|^2 = \sum_{j=1}^n \sum_{k=1}^n \left(\sum_{i=1}^n u_{ki}u_{ij} \right) x_j x_k$$

由 $\boldsymbol{U}$ 为正定矩阵有 $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$，那么上式即可等价为：

$$I_{jk} = \sum_{i=1}^n u_{ji}u_{ik}$$

当 $j = k$ 时，$I_{jk} = 1$，否则 $I_{jk} = 0$，所以得到：

$$\sum_{i=1}^n |p_i|^2 = \sum_{i=1}^n |x_i|^2$$

将上述代入原式得：

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

(2) 我们不妨令 $\boldsymbol{w}=\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{w}_1$, 那么便有：

$$\frac{\boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{B}\boldsymbol{w}}=\frac{\boldsymbol{w}^T\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{w}}{\boldsymbol{w}_1^T\boldsymbol{w}_1}$$

所以有：

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

而由于 $(\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}})^T=(\boldsymbol{B}^{-\frac{1}{2}})^T(\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}})^T=\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}=\boldsymbol{B}^{-1}\boldsymbol{A}$ 所以 $\boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{-\frac{1}{2}}$ 与 $\boldsymbol{B}^{-1}\boldsymbol{A}$ 特征值相等，所以有：

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

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**??**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:

$$\mathrm{Gain}(D,a)=\mathrm{Ent}(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\mathrm{Ent}(D^v)\geq 0, \tag{6}$$

即信息增益非负.

Solution. (1) 不一定，考虑数据集 $\left\{(\frac{1}{i},(-1)^i)\right\}_{i=1}^{\infty}$。若决策树为有限深度，那么其只能进行有限次属性判断，区间只能被划分为有限个区域，但该数据集将区间划分为无穷多个区域。

(2) 结果见图 1

(3) 我们不妨设 $P(X)$ 为第 X 个标记样本占总体的比例，$P(K)$ 为第 K 个可能取值样本占总体的比例，$P(X,K)$ 为第 K 个可能取值样本中标记为 X 的样本占总体的比例有 $P(X|K)=\frac{P(X,K)}{P(K)},P(X)=\sum_K P(X,K),P(K)=\sum_X P(X,K)$。

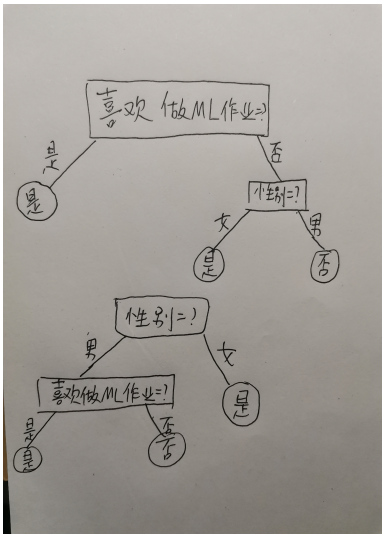

图 1: 决策树结果

证明.

$$
\begin{aligned}
Gain(D,a) &= \sum_X -P(X)\log_2 P(X) - \sum_K P(K)\sum_X \left(-P(X\mid K)\log_2 P(X\mid K)\right)\\
-Gaim(D,a) &= \sum_X P(X)\log_2 P(X) - \sum_K P(K)\sum_X \left(P(X\mid K)\log_2 P(X\mid K)\right)\\
&= \sum_X P(X)\log_2 P(X) - \sum_K P(K)\sum_X \left(P(X\mid K)\log_2 P(X\mid K)\right)\\
&= \sum_X\sum_K P(X,K)\log_2 P(X) - \sum_K\sum_X \left(P(K)P(X\mid K)\log_2 P(X\mid K)\right)\\
&= \sum_X\sum_K P(X,K)\log_2 P(X) - \sum_K\sum_X \left(P(X,K)\log_2 P(X\mid K)\right)\\
&= \sum_X\sum_K P(X,K)\left(\log_2 P(X) - \log_2 P(X\mid K)\right)\\
&= \sum_X\sum_K P(X,K)\left(\log_2\left(\frac{P(X)}{P(X\mid K)}\right)\right)\\
&= \sum_X\sum_K P(X\mid K)P(K)\left(\log_2\left(\frac{P(X)}{P(X\mid K)}\right)\right)\\
&= \sum_K P(K)\sum_X P(X\mid K)\left(\log_2\left(\frac{P(X)}{P(X\mid K)}\right)\right)\\
&\le \sum_K P(K)\left(\log_2\left(\sum_X \frac{P(X\mid K)P(X)}{P(X\mid K)}\right)\right)\ (From\ Jensen\ Inequality)\\
&\le \log_2\left(\sum_K\sum_X \frac{P(K)P(X\mid K)P(X)}{P(X\mid K)}\right)\ (From\ Jensen\ Inequality)\\
&\le \log_2\left(\sum_K\sum_X P(K)P(X)\right)\\
&\le \log_2\left(\sum_K P(K)\sum_X P(X)\right)\\
&\le \log_2\left(\sum_K P(K)\right)\\
&\le \log_2(1)\\
&\le 0
\end{aligned}
$$

□

所以有 $Gaim(D,a) \geqslant 0$.