

# 机器学习导论习题二

221300066, 季千<sup>①</sup>, qkjia@smail.nju.edu.cn

2024 年 4 月 19 日

## 作业注意事项

1. 作业所需的 LaTeX 及 Python 环境配置要求请参考: [Link];
2. 请在 LaTeX 模板中第一页填写个人的学号、姓名、邮箱;
3. 本次作业需提交的文件与对应的命名方式为:
  - (a) 作答后的 LaTeX 代码 — HW2.tex;
  - (b) 由 (a) 编译得到的 PDF 文件 — HW2.pdf;
  - (c) 编程题代码 — main.py;请将以上文件**打包为“学号\_姓名.zip”**(例如“221300001\_张三.zip”)后提交;
3. 若多次提交作业, 则在命名“.zip”文件时加上版本号, 例如“221300001\_张三\_v1.zip”(批改时以版本号最高, 提交时间最新的文件为准);
4. 本次作业提交截止时间为 **4 月 19 日 23:59:59**. 未按照要求提交作业, 提交作业格式不正确, **作业命名不规范**, 将会被扣除部分作业分数; 除特殊原因(如因病缓交, 需出示医院假条)逾期未交作业, 本次作业记 0 分; **如发现抄袭, 抄袭和被抄袭双方成绩全部取消**;
5. 学习过程中, 允许参考 ChatGPT 等生成式语言模型的生成结果, 但必须在可信的信息源处核实信息的真实性; **不允许直接使用模型的生成结果作为作业的回答内容**, 否则将视为作业非本人完成并取消成绩;
6. 证明题请给出关键证明步骤, 计算题请列出算式及中间结果, 否则不予计分; 撰写数学表达式请遵循符号约定.
7. 本次作业提交地址为 [Link], 请大家预留时间提前上交, 以防在临近截止日期时, 因网络等原因无法按时提交作业.

## 符号约定

[线性代数变量符号] 本次作业的部分题目涉及矩阵形式的数学推导, 请注意符号规范.

提示: 可以在导言区通过 `\def`, `\newcommand` 等命令自定义常用符号以简化编写, 例如 `\Real`, `\norm` 等.

- 矩阵请使用**粗体正体字母**, 例如  $\mathbf{A}, \mathbf{X}, \mathbf{O}$  (零矩阵);
- 向量请使用**粗体斜体字母**, 例如  $\mathbf{u}, \mathbf{v}, \mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{0}$  (零向量);
- 标量 (以及离散变量) 请使用**斜体小写字母**, 例如  $a, b, c, \lambda, \mu, \sigma$ ;
- 操作符或者函数名请使用**正体有衬线体字母**, 例如  $\min, \max, \text{Ent}, \text{MI}$ ;
- 转置矩阵 (Transposed Matrix) 以及埃尔米特矩阵 (Hermitian Matrix) 的**角标**请使用**正体无衬线体字母**, 例如  $\mathbf{A}^T, \mathbf{A}^H$ .

[奇异值分解与特征值分解] 由于  $\mathbb{R}^n \times \mathbb{R}^m$  上存在全序关系, 我们可以约定, 存在一种奇异值分解过程, 在此奇异值分解过程的意义下, 任何矩阵的奇异值分解是唯一的, 进而**任何 (半) 正定矩阵的特征值分解也是唯一的**. 如果没有符号冲突, 可以不加说明的使用以下记号:

- 矩阵  $\mathbf{A} \in \mathbb{R}^{n \times m}$  的**奇异值分解**是  $\mathbf{A} = \sum_{i=1}^{\min\{m,n\}} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ ,  $\sigma_i \geq \sigma_{i+1}$ ;
- **(半) 正定矩阵**  $\mathbf{A} \in \mathbb{R}^{n \times n}$  的**特征值分解**是  $\mathbf{A} = \sum_{i=1}^n \lambda_i \mathbf{v}_i \mathbf{v}_i^T$ ,  $\lambda_i \geq \lambda_{i+1}$ ;
- 另记  $\sigma_{\max}(\mathbf{A}) = \sigma_1(\mathbf{A})$ ,  $\lambda_{\max}(\mathbf{A}) = \lambda_1(\mathbf{A})$  是矩阵  $\mathbf{A}$  最大的奇异值.

[决策树] 为简化表述, 我们引入指示函数  $\mathbb{I}(e)$ : 若事件  $e$  发生, 则  $\mathbb{I}(e) = 1$ , 否则  $\mathbb{I}(e) = 0$ . 对于**表格数据**, 我们约定:

- 每个样例  $\mathbf{x}$  都具有  $d$  个属性  $\{1, \dots, d\}$ ;
- 第  $j$  个属性有  $n_j$  种可能的取值, 记作  $\mathbf{x}^j \in \{v_1^j, \dots, v_{n_j}^j\}$ ;
- 对于多分类问题, 样例的标签可以记作  $\mathbf{x}^\ell$ , 共有  $N$  种可能取值, 即  $\mathbf{x}^\ell \in \{1, \dots, N\}$ .

例如: 对于数据集  $D = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ , 事件  $\mathbf{x}^j = v_k^j$  的频率值可以记作  $\hat{p}(\mathbf{x}^j = v_k^j) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(\mathbf{x}_i^j = v_k^j)$ ; 再如, 对于“西瓜数据集 2.0” (详见教材表 4.1), 记号示范如下:

- $d = 6$ ;
- $\mathbf{x}^1 \in \{v_1^1 = \text{乌黑}, v_2^1 = \text{青绿}, v_3^1 = \text{浅白}\}$ , 即第 1 个属性色泽,  $n_1 = 3$ ;
- $\mathbf{x}^2 \in \{v_1^2 = \text{蜡缩}, v_2^2 = \text{稍蜷}, v_3^2 = \text{硬挺}\}$ , 即第 2 个属性根蒂,  $n_2 = 3$ ;
- $\mathbf{x}^3 \in \{v_1^3 = \text{清脆}, v_2^3 = \text{浊响}, v_3^3 = \text{沉闷}\}$ , 即第 3 个属性敲声,  $n_3 = 3$ ;
- $\mathbf{x}^4 \in \{v_1^4 = \text{清晰}, v_2^4 = \text{稍糊}, v_3^4 = \text{模糊}\}$ , 即第 4 个属性纹理,  $n_4 = 3$ ;
- $\mathbf{x}^5 \in \{v_1^5 = \text{凹陷}, v_2^5 = \text{稍凹}, v_3^5 = \text{平坦}\}$ , 即第 5 个属性脐部,  $n_5 = 3$ ;
- $\mathbf{x}^6 \in \{v_1^6 = \text{软粘}, v_2^6 = \text{硬滑}\}$ , 即第 6 个属性触感,  $n_6 = 2$ .

## 1 [20pts] 岭回归

在本题中, 我们假设所有凸优化问题均有解.

回顾教材第三章第二节, 多元线性回归相当于求解如下的无约束优化问题 (1.1). 其中, 对于  $\mathbf{v} \in \mathbb{R}^d$ , 定义  $\|\mathbf{v}\|_2^2 = \mathbf{v}^\top \mathbf{v}$ .

$$\min_{\hat{\mathbf{w}}} \|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|_2^2 \quad (1.1)$$

在多元线性回归中增加约束项, 即成为岭回归 (Ridge Regression), 即求解有约束优化问题 (1.2), 其中  $\rho > 0$  是待确定的超参数.

$$\begin{aligned} \min_{\hat{\mathbf{w}}} \quad & \|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|_2^2 \\ \text{s.t.} \quad & \|\hat{\mathbf{w}}\|_2^2 \leq \rho^2 \end{aligned} \quad (1.2)$$

岭回归也可以写成无约束优化问题 (1.3). 其中,  $\lambda \geq 0$  是待确定的超参数.

$$\min_{\hat{\mathbf{w}}} \|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|_2^2 + \lambda \|\hat{\mathbf{w}}\|_2^2 \quad (1.3)$$

- (1) [5pts] 相比于多元线性回归, 岭回归引入了额外的归纳偏好 (Inductive Bias). 回顾教材第一章第四节, 请简要回答: 岭回归引入了怎样的归纳偏好? 这样的归纳偏好有怎样的作用?

提示: 回顾过拟合 (Overfitting)、“奥卡姆剃刀” (Occam’s Razor) 原则等知识点; 结合特殊情形回答, 例如矩阵  $\mathbf{X}$  不满秩、数据中存在异常值等.

- (2) [5pts] 请证明岭回归的两种形式 (1.2) 和 (1.3) 等价.

提示: 考虑 KKT 条件 (Karush-Kuhn-Tucker Conditions).

- (3) [5pts] 对于无约束优化形式 (1.3), 假设  $\lambda$  已经确定, 此时岭回归的解记作  $\mathbf{w}^*$ , 请推导出  $\mathbf{w}^*$  的表达式.

- (4) [5pts] 在 (3) 的基础上, 请推导出  $\lambda$  的下界 (关于  $\rho$  的函数), 并据此回答:  $\rho$  减小时, 若希望保持 (1.2) 和 (1.3) 的解一致, 需要怎样调整  $\lambda$ ?

提示: 你可能需要引入  $\sigma_{\max}(\mathbf{X})$ .

**Solution.** 此处用于写解答 (中英文均可)

(1) 归纳偏好: 岭回归希望让解向量  $\hat{\mathbf{w}}$  在满足能够最小化  $\|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|_2^2$  的同时, 更加偏好于那些在欧几里得空间中的模长尽量小, 不超过一个超参数  $\rho$  的解向量.

作用:

1. 首先最小化  $\|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|_2^2$  是一个无约束的优化问题, 那么可转化为求解能够使  $2\mathbf{X}^T(\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}) = \mathbf{0}$  的  $\hat{\mathbf{w}}$ . 那么如果  $\mathbf{X}^T\mathbf{X}$  满秩的话, 有唯一的闭式解; 如果不满秩的话, 那么满足这个式的  $\hat{\mathbf{w}}$  会构成一个解空间. 引入这个归纳偏好后, 岭回归算法会挑选 2 范数不超过超参数  $\rho$  的解 (根据奥卡姆剃刀原则, 某种程度上可以认为这种解的形式更加简单).

2. 原始的线性回归是通过最小化均方误差来得到解, 而均方误差对异常值很敏感, 容易导致出现对某一项指标的过分偏重. 而岭回归通过引入一个正则化项去惩罚过大的  $\hat{\mathbf{w}}$ , 从而减小了异常值对于最优解的影响, 以在存在较大噪声的前提下达到更好回归的效果.

(2) 证明:

(1.2) 式中的  $\|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|^2$  显然是关于  $\hat{\mathbf{w}}$  的凸函数，同理因为  $\lambda \geq 0$ ，(1.3) 式中的  $\|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|^2 + \lambda\|\hat{\mathbf{w}}\|^2$  也是关于  $\hat{\mathbf{w}}$  的一个凸函数。由式 (1.3) 是凸函数，可立即得到其最优解满足下面的式子：

$$\frac{\partial\|\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}\|^2 + \lambda\|\hat{\mathbf{w}}\|^2}{\partial\mathbf{w}} = 2\mathbf{X}^T(\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}) + 2\lambda\mathbf{w} = \mathbf{0}$$

通过整理可得最优解  $\hat{\mathbf{w}}$  满足下式：

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\hat{\mathbf{w}} = \mathbf{X}^T\mathbf{y}$$

由式 (1.2) 以及其约束条件可知，这是一个凸优化问题，所以满足 KKT 条件的点就是其最优解，其 KKT 条件如下：

$$\|\hat{\mathbf{w}}\|_2^2 - \rho^2 \leq 0$$

$$\mu \geq 0$$

$$\mu(\|\hat{\mathbf{w}}\|_2^2 - \rho^2) = 0$$

$$2\mathbf{X}^T(\mathbf{X}\hat{\mathbf{w}} - \mathbf{y}) + 2\mu\hat{\mathbf{w}} = \mathbf{0}$$

那么要证明这两个优化问题等价，即证明：

$$1. \forall \lambda \geq 0, \exists \rho > 0, s.t. \hat{\mathbf{w}}_1 = \hat{\mathbf{w}}_2$$

$$2. \forall \rho > 0, \exists \lambda \geq 0, s.t. \hat{\mathbf{w}}_2 = \hat{\mathbf{w}}_1$$

下面规定： $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2$  分别是 (1.2), (1.3) 取得最优解时  $\hat{\mathbf{w}}$  的值。

证明 1：

对  $\forall \lambda > 0$ ，因为有对于任意  $\hat{\mathbf{w}}$ ， $\hat{\mathbf{w}}^T \mathbf{X}^T \mathbf{X} \hat{\mathbf{w}} = (\hat{\mathbf{w}} \mathbf{X})^T (\hat{\mathbf{w}} \mathbf{X}) \geq 0$ ，则  $\mathbf{X}^T \mathbf{X}$  一定是半正定矩阵，所以  $\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}$  一定是正定矩阵，所以它可逆， $\hat{\mathbf{w}}_2$  有唯一解，所以有  $\hat{\mathbf{w}}_2 = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$ 。此时来观察 KKT 条件，只需令  $\mu = \lambda$ ， $\rho = \|(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}\|$ ，取  $\hat{\mathbf{w}}_1 = \hat{\mathbf{w}}_2$  即可满足 KKT 条件，所以有  $\hat{\mathbf{w}}_1 = \hat{\mathbf{w}}_2$  是  $\rho = \|(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}\|$  的条件下问题 (1.2) 的解。

当  $\lambda = 0$  时，令  $\mu = 0$ ，因为  $\mathbf{X}^T \mathbf{X}$  半正定，所以解出来的  $\hat{\mathbf{w}}_2$  的集合是一个子空间，对于该解空间中任意的  $\hat{\mathbf{w}}_2$ ，只需令  $\rho = \|\hat{\mathbf{w}}_2\|$  即可满足 KKT 条件，即可以得到  $\hat{\mathbf{w}}_1 = \hat{\mathbf{w}}_2$ 。

证明 2：

对  $\forall \rho > 0$ ，取 KKT 条件中  $\|\hat{\mathbf{w}}_2\| = \rho^2$ ，必须让  $\hat{\mathbf{w}}_1$  满足  $2\mathbf{X}^T(\mathbf{X}\hat{\mathbf{w}}_1 - \mathbf{y}) + 2\mu\hat{\mathbf{w}}_1 = \mathbf{0}$ ，且此时  $\mu$  依赖于  $\rho$  的值。根据之前的分析可以解得  $\hat{\mathbf{w}}_1 = (\mathbf{X}^T \mathbf{X} + \mu \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$ ，因为此时  $\|\hat{\mathbf{w}}_1\| = \rho^2$ ，取能够满足  $(\mathbf{X}^T \mathbf{X} + \mu \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}^2 = \rho^2$  的大于 0 的  $\mu$  即可，此时再令  $\lambda = \mu$ ，即可得到  $\hat{\mathbf{w}}_2 = \hat{\mathbf{w}}_1$ 。

下面讨论一种特殊情况，即找不到  $\mu > 0$  满足  $\|(\mathbf{X}^T \mathbf{X} + \mu \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}\| = \rho^2$ ，因为变大  $\mu$  实际上是在缩小  $\mathbf{X}^T \mathbf{y}$  的模长，所以出现这种情况说明  $\rho$  的值很大，那么就取不到  $\|\hat{\mathbf{w}}_1\|_2^2 = \rho^2$ ，因而此时有  $\|\hat{\mathbf{w}}_1\|_2^2 < \rho^2$ ，令  $\mu = 0, \lambda = 0$ ，就可以找到  $\hat{\mathbf{w}}_2 = \hat{\mathbf{w}}_1$ 。

综上，可以证明这两个条件是等价的。

(3) 通过 (2) 中的推导，易得：

当  $\lambda > 0$  时：

$$\mathbf{w}^* = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

当  $\lambda = 0$  时:

$$\mathbf{w}^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

(4)  $\lambda$  应该调大。

从 (2) 的分析中可以得到, 当这两个问题等价时, 有:

$$\rho \geq \|(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}\|_2$$

因为  $\mathbf{X}^T \mathbf{y}$  得到的是一个向量, 又因为  $\mathbf{X}^T \mathbf{X}$  是半正定矩阵, 选取其最大特征值  $\sigma_{\max}(\mathbf{X}^T \mathbf{X})$ , 根据高等代数的相关结论, 有  $(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1}$  的最小特征值为  $\frac{1}{\lambda + \sigma_{\max}(\mathbf{X}^T \mathbf{X})}$ , 那么显然有:

$$\rho \geq \|(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}\|_2 \geq \frac{\|\mathbf{X}^T \mathbf{y}\|_2}{\lambda + \sigma_{\max}(\mathbf{X}^T \mathbf{X})}$$

在这里第二个不等号能够成立是因为  $(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1}$  相当于对向量做了正交变换后伸缩其基的长度, 那么当该向量与被缩短最明显的向量平行时, 其模长最小. 不难得到  $\lambda$  的下界为:

$$\lambda \geq \frac{\|\mathbf{X}^T \mathbf{y}\|_2}{\rho} - \sigma_{\max}(\mathbf{X}^T \mathbf{X})$$

显然  $\lambda$  的下界是关于  $\rho$  的减函数, 所以当  $\rho$  减小时,  $\lambda$  的下界会变大, 此时应该将  $\lambda$  相应地调大。

## 2 [20pts] 决策树的构建流程

**[注意事项]** 本题可使用 PowerPoint®, Visio® 等软件绘制决策树, 导出为图片或 PDF 插入到作业文档中; 亦允许手绘决策树, 但是请确保文字与线条的工整. 如果插入照片, 请确保照片内容清晰可读, 否则将扣除部分分数.

考虑如下的表格数据集: 在诊断疾病 RD 时, 通常采用 UW-OCT, OCT-PD, US 三种检测手段, 其中 1 代表阳性, 0 代表阴性. 假设总共收集了 16 位病人的检测结果, 其中 8 条用于训练, 如表格 1 所示, 8 条用于验证, 如表格 2 所示.

表 1: 训练集  $D_{\text{train}}$

编号	UW-OCT	OCT-PD	US	RD
1	1	1	0	1
2	1	1	1	1
3	0	1	1	1
4	0	1	1	1
5	0	0	1	0
6	1	0	0	0
7	1	0	1	0
8	0	1	0	0

表 2: 验证集  $D_{\text{valid}}$

编号	UW-OCT	OCT-PD	US	RD
9	1	1	1	1
10	0	1	1	1
11	0	1	0	0
12	1	0	1	0
13	0	1	1	1
14	1	1	0	0
15	1	0	0	0
16	0	0	0	0

- (1) [10pts] 回顾教材第四章第一, 二节, 请使用基尼指数作为划分准则, 通过训练集中的数据训练决策树. 在 HW2.pdf 中展示**最终得到的决策树**, 并给出**每个节点划分时的计算和比较过程**.
- (2) [5pts] 回顾教材第四章第三节, 在 (1) 的基础上, 请判断每个节点是否需要预剪枝. 在 HW2.pdf 中展示**预剪枝后的决策树**, 并给出**每个节点预剪枝时的计算和比较过程**.
- (3) [5pts] 对一个学习任务来说, 给定属性集, 其中有些属性可能很关键, 很有用, 称为“相关特征”, 另一些属性则可能用处不大, 称为“无关特征”. 请简要回答如下问题:

- (a) 比较 (1,2) 的结果, 指出当前训练集和验证集划分下的无关特征, 并说明理由.
- (b) 如果不给出数据集, 只给出决策树和剪枝后的决策树, 应该怎样挑选无关特征?
- (c) 如果不给出数据集, 也不给出剪枝后的决策树, 只给出未剪枝的决策树, 还能挑选无关特征吗? 请简要给出理由.

**Solution.** 此处用于写解答 (中英文均可)

(1) 在还是根结点时, 当前属性集合为  $\{UW - OCT, OCT - PD, US\}$ , 对每个属性, 都有  $\{0, 1\}$  两个取值, 所以对于所有属性的基尼值:

$$Gini(D, UW - OCT = 1) = 1 - (1/2)^2 - (1/2)^2 = 1/2$$

$$Gini(D, UW - OCT = 0) = 1 - (1/2)^2 - (1/2)^2 = 1/2$$

$$Gini(D, OCT - PD = 1) = 1 - (4/5)^2 - (1/5)^2 = 8/25$$

$$Gini(D, OCT - PD = 0) = 1 - (0 + 1) = 0$$

$$Gini(D, US = 1) = 1 - (3/5)^2 - (2/5)^2 = 12/25$$

$$Gini(D, US = 0) = 1 - (1/3)^2 - (2/3)^2 = 4/9$$

可以求得各属性的基尼系数为:

$$Gini\_index(D, UW - OCT) = 4/8 \times 1/2 + 4/8 \times 1/2 = 1/2$$

$$Gini\_index(D, OCT - PD) = 5/8 \times 8/25 + 3/8 \times 0 = 1/5$$

$$Gini\_index(D, US) = 5/8 \times 12/25 + 3/8 \times 4/9 = 7/15$$

因为 OCT-PD 的基尼系数最小, 可以设置为划分属性, 得到以下决策树:

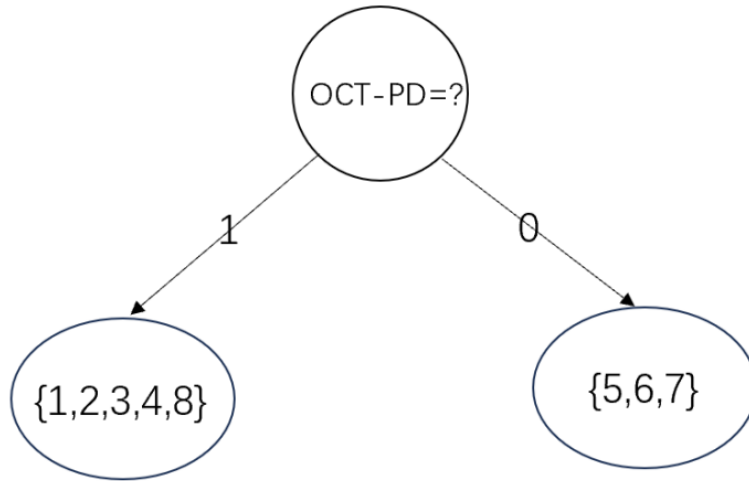


图 1: 决策树构建流程 1

可以发现右子树中样例都是阴性, 可直接设为阴性。接下来, 我们计算左边集合各个属性的基尼值:

$$Gini(D1, UW - OCT = 1) = 1 - (1 + 0) = 0$$

$$Gini(D1, UW - OCT = 0) = 1 - \frac{2}{3} \times \frac{2}{3} = \frac{4}{9}$$

$$Gini(D1, US = 0) = 1 - (1 + 0) = 0$$

$$Gini(D1, US = 1) = 1 - \frac{1}{2} \times \frac{1}{2} = \frac{1}{2}$$

可以计算得到各属性的基尼系数为：

$$Gini\_index(D, UW - UCT) = \frac{2}{5} \times 0 + \frac{3}{5} \times \frac{4}{9} = \frac{4}{15}$$

$$Gini\_index(D, US) = \frac{3}{5} \times 0 + \frac{2}{5} \times \frac{1}{2} = \frac{1}{5}$$

可以发现左结点中 US 的基尼系数最小，以它为最优划分属性，可以得到以下决策树，又发现左边结点都为阳性，所以有：

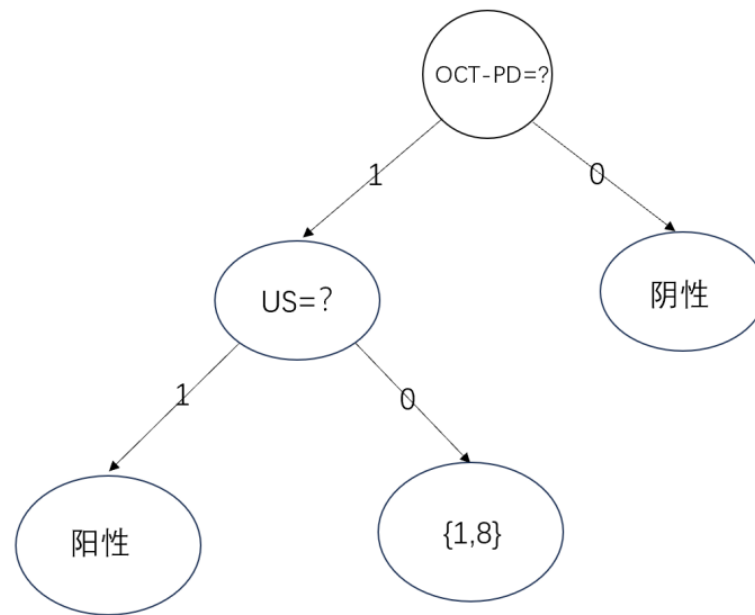


图 2: 决策树构建流程 2

最后只剩下 UW-UCT 一个属性，而且恰好一个阳性一个阴性，所以可以得到最终决策树如下：



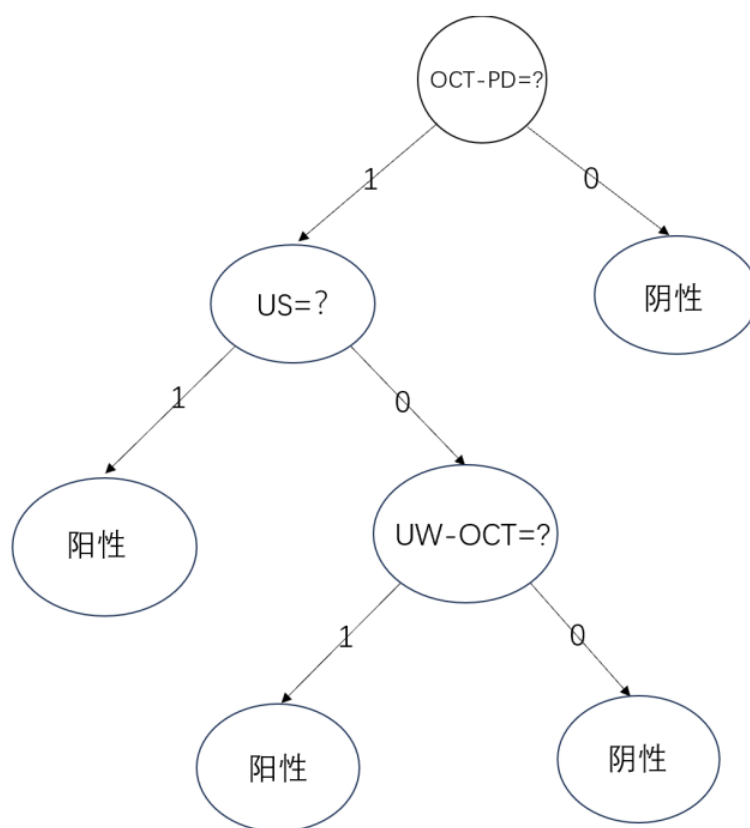


图 3: 决策树

(2) 考虑根节点，如果不划分，因为训练集中阳性阴性个数相同，假设被标记为阳性，则划分前验证集精度为  $acc = 3/8$  (若标记成阴性则精度为  $acc = 5/8$ )，划分后，右边被标记成阴性，左边被标记成阳性，那么有  $acc = 3 + 3/8 = 3/4 \geq 5/8$ ，所以该结点应该被划分。接下来是结点  $US = ?$ ，划分前，有  $\{9, 10, 11, 13, 14\}$  在这个结点处，被划分成阳性，精度为  $acc = 3/5$ ，划分后，分类全部正确，精度为  $acc = 1$ ，所以应该划分。最后考虑  $UW-OCT = ?$  这个结点，不划分的话，精度为 1(由上可知)，而划分后有一个被错误归类，精度为  $1/2$ ，所以这个结点不该被划分。最终得到预剪枝决策树如下：

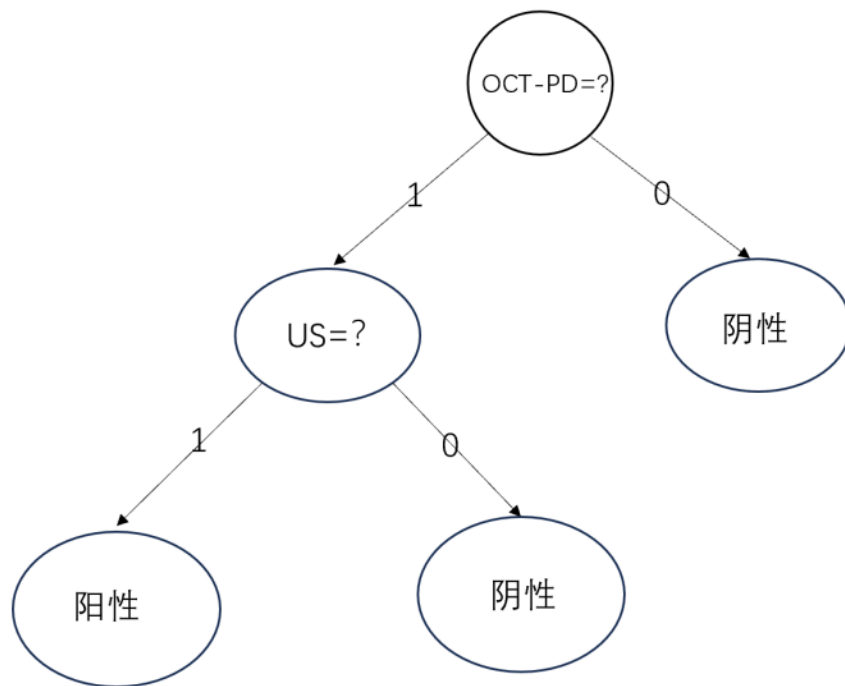


图 4: 预剪枝决策树

(3)

(a) 因为剪枝之后，UW-OCT 特征消失了，所以当前训练集和验证集划分下的无关特征是 UW-OCT。

(b) 可以比较剪枝前后的决策树结构，剪枝后消失的特征可以视为无关或冗余的特征。

(c) 可以尝试进行挑选。如果决策树深度越大，包含样本数更少的特征就很可能是无关特征，但这也只是概率上的可能，并不能完全确定该特征就是无关特征。

### 3 [15pts] 决策树的划分准则

- (1) [5pts] 回顾教材第四章第一节, 请结合决策树基本算法中递归返回的条件, 简要回答: 如果要求“只要训练集不含冲突数据 (即特征向量完全相同但标记不同), 就必须获得与训练集一致 (即训练误差为 0) 的决策树”, 那么纯度函数需要满足怎样的要求?
- (2) [5pts] 回顾教材第四章第二节, 信息增益可以重写为互信息 (Mutual Information)

$$MI(E, F) = \sum_{e \in E} \sum_{f \in F} p(e, f) \log \frac{p(e, f)}{p(e)p(f)},$$

其中  $E, F$  都是事件的集合. 请给出此时  $E, F$  的定义, 列出必要的推导步骤, 并使用  $MI(E, F)$ ,  $Ent(E)$ ,  $Ent(F)$  等符号表示增益率.

- (3) [5pts] 考虑归一化互信息 (Normalized Mutual Information) 的一种定义

$$NMI(E, F) = \frac{MI(E, F)}{(MI(E, E) + MI(F, F)) / 2} = \frac{2 \cdot MI(E, F)}{Ent(E) + Ent(F)}.$$

在 (2) 的基础上, 如果使用归一化互信息作为划分准则, 与使用增益率作为划分准则产生的决策树相同吗? 请给出证明或举出反例.

提示: 已知数学性质  $0 \leq MI(E, F) \leq \min\{Ent(E), Ent(F)\}$ .

**Solution.** 此处用于写解答 (中英文均可)

(1) 当训练轮数充足时, 纯度函数只需要满足能够使所有属性都可以在某一轮被选为最优属性即可, 而且纯度函数不能有随机性, 对同分布的属性和类别信息能够输出一致的结果; 因为该决策树是基于训练集的, 所以不可能出现落入  $D_v$  为空的情况。所以只需考虑  $A$  为空或者  $D$  在  $A$  上取值相同的情况, 那么可以看出生成叶节点只会在样本属于同一类或者所有特征值都用完的时候, 又因为同一节点中的样本, 其路径上的特征值都是完全相同的, 而又因为训练集上没有冲突数据, 则每个节点上训练误差都为 0。

(2) 定义  $E$  为样本集中所有标签的取值情况, 每个元素  $e$  表示样本标签为  $e$  这个事件, 定义  $F$  为属性集合  $a$  的所有取值情况, 每个元素  $f$  表示样本在属性  $a$  中取值为属性  $f$  的事件。(此处的属性  $a$  针对增益率公式中的属性  $a$ )

下面定义:  $Ent(E|F) = \sum_{f \in F} p(f) Ent(E|F = f)$

那么对于信息增益公式, 不难发现有:

$$Gain(D, a) = Ent(D) - \sum_{v=1}^V \frac{|D^v|}{|D|} Ent(D^v) = Ent(E) - Ent(E|F)$$

互信息公式的推导过程如下:

$$\begin{aligned} MI(E, F) &= \sum_{e \in E} \sum_{f \in F} p(e, f) \log \frac{p(e, f)}{p(e)p(f)} \\ &= \sum_{e \in E} \sum_{f \in F} p(e, f) \log \frac{p(e, f)}{p(f)} - \sum_{e \in E} \sum_{f \in F} p(e, f) \log p(e) \\ &= \sum_{e \in E} \sum_{f \in F} p(e | f) \log p(e | f) - \sum_{e \in E} \sum_{f \in F} p(e, f) \log p(e) \end{aligned}$$

$$\begin{aligned}
&= \sum_{f \in F} p(f) \sum_{e \in E} p(e | f) \log p(e | f) - \sum_{e \in E} \log p(e) \sum_{f \in F} p(e, f) \\
&= - \sum_{f \in F} p(f) Ent(E | F = f) - \sum_{e \in E} \log p(e) p(e) \\
&= Ent(E) - Ent(E | F) = Gain(E, F)
\end{aligned}$$

增益率表示如下：

用  $E_f$  表示样本在属性  $a$  上取值为  $f$  的事件令  $IV(F) = - \sum_{f \in F} \frac{|E_f|}{|E|} \log \frac{|E_f|}{|E|} = Ent(F)$ ，则有：

$$Gain\_ratio(E, F) = \frac{Gain(E, F)}{IV(F)} = \frac{MI(E, F)}{IV(F)} = \frac{MI(E, F)}{Ent(F)}$$

(3) 可能会不一致，假设有一个数据集，其中有两个特征  $F1$  和  $F2$ ，以及一个目标变量  $T$ 。 $F1$  的取值有很多，而  $F2$  的取值只有两个，同时假设  $F1$  和  $T$  之间的互信息很高，而  $F2$  和  $T$  之间的互信息较低。在这种情况下，如果使用  $NMI$  作为划分准则，可能会优先选择  $F1$  作为划分特征；而如果使用增益率作为划分准则，可能会优先选择  $F2$  作为划分特征，因为增益率会对取值较多的特征进行惩罚。

具体反例构造如下：

假设有两个标签  $0, 1$ ，和两个属性  $b_0, b_1$ ，且它们都有  $0, 1$  两个取值，下面让： $p(l = 0, b_0 = 0) = 0.05$ ， $p(l = 0, b_0 = 1) = 0.15$ ， $p(l = 1, b_0 = 0) = 0.70$ ， $p(l = 1, b_0 = 1) = 0.10$  令  $F^{b_0}$  表示属性集合  $b_0$  的所有取值情况。所以有：

$$Ent(E) \approx 0.722, Ent(F^{b_0}) \approx 0.810, MI(E, F^{b_0}) \approx 0.200$$

再让  $p(l = 0, b_1 = 0) = 0.03$ ， $p(l = 0, b_1 = 1) = 0.17$ ， $p(l = 1, b_1 = 0) = 0.63$ ， $p(l = 1, b_1 = 1) = 0.17$  那么有：

$$Ent(E) \approx 0.722, Ent(F^{b_1}) \approx 0.915, MI(E, F^{b_1}) \approx 0.199$$

所以此时：

$$\frac{MI(E, F^{b_0})}{Ent(F^{b_0})} \approx 0.247 < 0.266 \approx \frac{MI(E, F^{b_1})}{Ent(F^{b_1})}$$

而： $NMI(E, F^{b_0}) > NMI(E, F^{b_1})$

因此这两种划分将产生不同的决策树，反例构造成功。

## 4 [20(+5)pts] 线性判别分析

回顾教材第三章第四节, LDA (Linear Discriminant Analysis) 有两个优化目标: 最小化类内散度  $w^T S_w w$  与最大化类间散度  $w^T S_b w$ , 目的是使得同类样例的投影点尽可能接近, 异类样例的投影点尽可能远离. 在 LDA 之外, 课堂上还介绍了 PCA (Principal Components Analysis, 主成分分析). 事实上, PCA 可以写成类似 LDA 的形式, 但 PCA 只有一个目标, 即最大化全局散度:  $\max_w w^T S_t w$ .

- (1) [5pts] 教材图 3.3 中, “+”, “-” 代表数据点, 任务需要把二维数据降维到一维, 直线  $y = w^T x$  代表 LDA 的投影方向. 请画出图 5 数据集上 PCA 的大致投影方向 (可以使用蓝色直线), 并在 HW2.pdf 中展示.

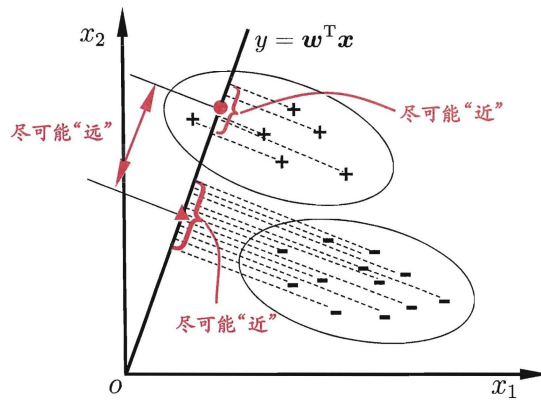


图 5: 教材图 3.3

- (2) [5pts] 请参考题干中的介绍与 (1) 中的现象, 简要回答:
- (a) 对照题干中 LDA 的优化目的, PCA 的优化目的是什么?
  - (b) PCA 相较于 LDA 有什么显著的不同点?
- (3) [5pts] 下面, 我们先回顾教材第三章第四节中多分类 LDA 优化问题的矩阵形式. 考虑总类内散度是各个类别散度之和, 其矩阵形式为:  $S_w = \sum_{i=1}^N S_{w_i}$ . 对于第  $i$  个类别的类内散度矩阵定义如下:  $S_{w_i} = \sum_{x \in X_i} (x - \mu_i)(x - \mu_i)^T$ . 类似的, 总类间散度是各个类别中心相对于全局中心的散度之和, 其矩阵形式为:  $S_b = \sum_{i=1}^N S_{b_i}$ . 对于第  $i$  个类别的中心相对于全局中心的散度矩阵定义如下:  $S_{b_i} = m_i(\mu_i - \mu)(\mu_i - \mu)^T$ . LDA 事实上是在最小化平均类内散度和最大化平均类间散度, 其矩阵形式如 (4.1) 所示. 其中,  $d'$  是降维后的维度, 严格小于数据维度  $d$ .

$$\max_w J(W) = \frac{\text{tr}(W^T S_b W)}{\text{tr}(W^T S_w W)} = \frac{\text{tr}\left(W^T \left(\sum_{i=1}^N S_{b_i}\right) W\right)}{\text{tr}\left(W^T \left(\sum_{i=1}^N S_{w_i}\right) W\right)} = \frac{\frac{1}{N} \sum_{i=1}^N \text{tr}(W^T S_{b_i} W)}{\frac{1}{N} \sum_{i=1}^N \text{tr}(W^T S_{w_i} W)}$$

$$\text{s.t. } W^T W = I_{d'} \quad (4.1)$$

根据教材中的介绍, (4.1) 可通过广义特征值分解进行求解. 然而, 在某些现实场景下, 我们应用 LDA 的目的是提高分类准确率, 那么通常进一步希望每个类别散度尽可能

小, 每个类别中心相对于全局中心的散度尽可能大, 而非平均散度. 因此, 考虑 LDA 的一种拓展形式:

$$\begin{aligned} \max_{\mathbf{W}} \quad & \left( \min_{i,j} J_{i,j}(\mathbf{W}) \right) = \frac{\min_j \{ \text{tr}(\mathbf{W}^T \mathbf{S}_{b_j} \mathbf{W}) \}}{\max_i \{ \text{tr}(\mathbf{W}^T \mathbf{S}_{w_i} \mathbf{W}) \}} \\ \text{s.t.} \quad & \mathbf{W}^T \mathbf{W} = \mathbf{I}_{d'} \end{aligned} \quad (4.2)$$

请指出拓展形式 (4.2) 无法直接沿用原有形式 (4.1) 的广义特征值求解算法的原因.

提示: 指出求解时存在变量间的循环依赖关系.

- (4) [5pts] 在线性代数中, 对于 (半) 正定矩阵  $\mathbf{A}, \mathbf{B}$ , 若  $(\mathbf{A} - \mathbf{B})$  是正定矩阵, 则通常记作  $\mathbf{A} \succ \mathbf{B}$  或  $\mathbf{B} \prec \mathbf{A}$ ; 若  $(\mathbf{A} - \mathbf{B})$  是半正定矩阵, 则通常记作  $\mathbf{A} \succeq \mathbf{B}$  或  $\mathbf{B} \preceq \mathbf{A}$ . 在优化问题中, 凸优化问题有多项式时间复杂度的理论保证, 能够高效求解. 凸优化问题的定义是: (i) 最小化的目标函数是凸函数, 或者最大化的目标函数是凹函数, 而且 (ii) 可行域是凸集合. 可行域是所有满足约束条件的控制变量取值 (又称可行解) 构成的集合. 拓展形式 (4.2) 不能沿用原有形式 (4.1) 的求解算法, 也不是凸优化问题. 为了高效求解, 需要探索一种将其转化成凸优化问题的方法. 已知原有形式 (4.1) 可以松弛成如下凸优化问题:

$$\begin{aligned} \max_{\mathbf{W}, r} \quad & r \\ \text{s.t.} \quad & r \cdot \text{tr}(\mathbf{S}_w \mathbf{M}) - \text{tr}(\mathbf{S}_b \mathbf{M}) \leq 0 \\ & -r \leq 0 \\ & \mathbf{0} \preceq \mathbf{M} \preceq \mathbf{I}_d \\ & \text{tr}(\mathbf{M}) = d' \end{aligned} \quad (4.3)$$

请仿照原有形式 (4.1) 的松弛形式 (4.3), 给出拓展形式 (4.2) 的松弛形式, 并证明拓展形式的松弛形式是凸优化问题, 即同时满足条件 (i) 和条件 (ii).

- (5) [5pts] (本题为附加题, 得分计入卷面分数, 但本次作业总得分不超过 100 分) 请证明:

- (a) 松弛形式 (4.1) 和原有形式 (4.3) 的约束条件不等价;
- (b) 当  $r \cdot \text{tr}(\mathbf{S}_w \mathbf{M}) - \text{tr}(\mathbf{S}_b \mathbf{M}) = 0$  时, (4.3) 的可行域是 (4.1) 可行域的凸包 (Convex Hull). 即: (4.3) 的可行解可以表示成 (4.1) 的可行解的线性组合.

进而, (4.3) 的可行域是包含 (4.1) 的可行域的最小凸集合, 即 (4.3) 对 (4.1) 的放松程度是最小的, 因而能够使得凸问题 (4.3) 的解尽可能的接近原问题 (4.1) 的解.

**Solution.** 此处用于写解答 (中英文均可)

(1)

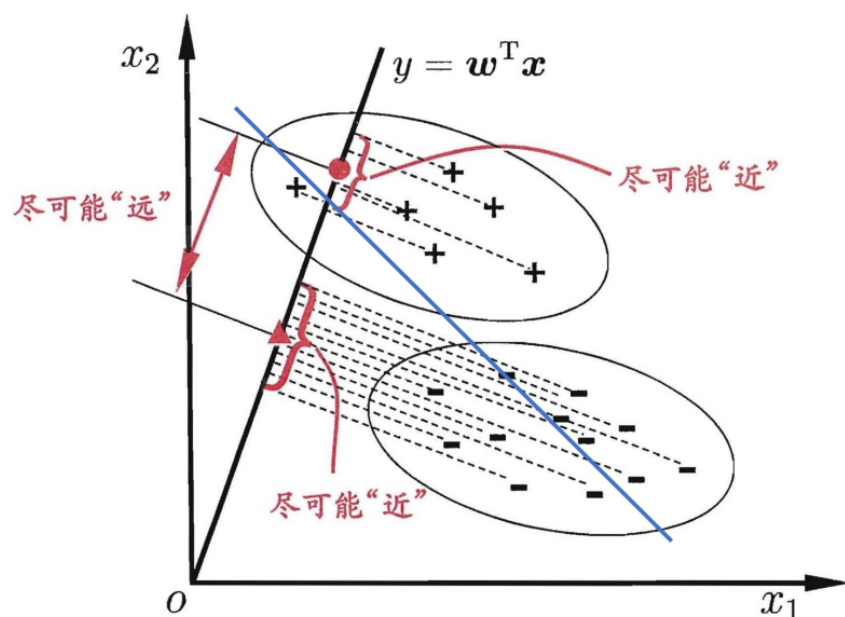


图 6: PCA 的大致投影方向

(2)

(a) PCA 的优化目的是：找到一个超平面，使得样本点在投影到这个超平面上后它们之间的投影能尽可能分开，即最大化投影后样本点的方差。

(b) 显著不同点：

1. LDA 需要考虑类间的距离，也就是说 LDA 用到了样本点的标签，是一种监督学习的方法。但 PCA 并不需要用到样本点是属于哪一类的标签，是一种无监督学习的方法。

2. LDA 会选择分类性能最好的投影方向，而 PCA 会选择样本点投影具有最大方差的方向。

(3) 按照原来 (4.1) 的形式，我们将其转化成下面的优化形式：

$$\min_{\mathbf{W}} -\text{tr}(\mathbf{W}^T \mathbf{S}_b \mathbf{W})$$

$$\text{s.t.} \quad -\text{tr}(\mathbf{W}^T \mathbf{S}_w \mathbf{W}) = 1$$

写出其拉格朗日函数然后对  $\mathbf{W}$  求导后可以通过广义特征值来求解，其形式如下：

$$\mathbf{S}_b \mathbf{W} = \lambda \mathbf{S}_w \mathbf{W}$$

注意到这里的形式相较于课本里的形式还多了个  $\mathbf{W}^T \mathbf{W} = \mathbf{I}_{d'}$  的约束，但实际上这个约束可以说是不起作用的，只起到了标准化模长的作用，因为  $(\mathbf{S}_w)^{-1} \mathbf{S}_b$  是一个实对称矩阵，那么其特征向量都是正交的，所以求得的  $\mathbf{W}$  必然满足  $\mathbf{W}^T \mathbf{W} = \mathbf{I}_{d'}$ ，所以说这个约束实际上不起作用。

(4) 问题 (4.2) 的松弛形式如下：

$$\begin{aligned} \max_{\mathbf{W}, r} \quad & r \\ \text{s.t.} \quad & r \cdot w - b \leq 0 \end{aligned}$$

$$w \leq \text{tr}(\mathbf{S}_{wi}\mathbf{M})$$

$$-b \leq \text{tr}(\mathbf{S}_{bj}\mathbf{M})$$

$$-r \leq 0$$

$$\mathbf{O} \preceq \mathbf{M} \preceq \mathbf{I}_d$$

$$\text{tr}(\mathbf{M}) = d'$$

证明：由于优化目标和优化函数都是仿射的，所以这个问题显然是凸优化问题。

(5)

(a) 要证明松弛形式 (4.1) 和原有形式 (4.3) 的约束条件不等价，需要比较两个问题中的约束。

在 (4.1) 中，约束是  $\mathbf{W}^\top \mathbf{W} = \mathbf{I}_{d'}$ ，这是一个关于  $\mathbf{W}$  的正交约束，即要求  $\mathbf{W}$  是一个正交矩阵。

而在 (4.3) 中，约束是  $r \cdot \text{tr}(\mathbf{S}_w \mathbf{M}) - \text{tr}(\mathbf{S}_b \mathbf{M}) \leq 0$ ， $-r \leq 0$ ， $\mathbf{O} \preceq \mathbf{M} \preceq \mathbf{I}_{d'}$ ，以及  $\text{tr}(\mathbf{M}) = d'$ 。这里引入了额外的变量  $r$  和  $\mathbf{M}$ ，并且  $\mathbf{M}$  被要求是一个介于零矩阵  $\mathbf{O}$  和单位矩阵  $\mathbf{I}_{d'}$  之间的半正定矩阵，且迹为  $d'$ 。

显然，(4.1) 中的正交约束不能保证 (4.3) 中关于  $r$  和  $\mathbf{M}$  的所有约束都得到满足，反之亦然。因此，这两个问题的约束条件是不等价的。

(b) 当  $r \cdot \text{tr}(\mathbf{S}_w \mathbf{M}) - \text{tr}(\mathbf{S}_b \mathbf{M}) = 0$  时，我们来考虑 (4.3) 的可行域与 (4.1) 可行域的关系。由于  $r \leq 0$ ，当  $r \cdot \text{tr}(\mathbf{S}_w \mathbf{M}) - \text{tr}(\mathbf{S}_b \mathbf{M}) = 0$  时，实际上有  $r = 0$ 。在这种情况下，(4.3) 中的约束简化为  $\mathbf{O} \preceq \mathbf{M} \preceq \mathbf{I}_{d'}$  和  $\text{tr}(\mathbf{M}) = d'$ 。

此时，任何满足  $\mathbf{W}^\top \mathbf{W} = \mathbf{I}_{d'}$  的  $\mathbf{W}$  都可以扩展为一个满足  $\mathbf{O} \preceq \mathbf{M} \preceq \mathbf{I}_{d'}$  和  $\text{tr}(\mathbf{M}) = d'$  的  $\mathbf{M}$ ，通过令  $\mathbf{M} = \mathbf{W}\mathbf{W}^\top$ 。这样的  $\mathbf{M}$  会自动满足  $\text{tr}(\mathbf{M}) = d'$  因为  $\text{tr}(\mathbf{W}\mathbf{W}^\top) = \text{tr}(\mathbf{W}^\top \mathbf{W}) = d'$ 。因此，(4.3) 的可行解可以表示成 (4.1) 的可行解的线性组合，即 (4.3) 的可行域是 (4.1) 可行域的凸包。

综上所述，(4.3) 的可行域是包含 (4.1) 的可行域的最小凸集合，这意味着 (4.3) 对 (4.1) 的放松程度是最小的，从而使得凸问题 (4.3) 的解尽可能接近原问题 (4.1) 的解。



## 5 [25pts] 编程实验: LDA 与多分类

**[注意事项]** 请不要修改或提交 `utils.py`; 在合理范围内, 运行时间和错误率不作为评分依据. 实现过程中只允许使用 NumPy 和 SciPy 提供的矩阵运算接口, 否则相应题目不计入分数.

此外, 对于题 (3,4,5), 如果调用 Sci-Kit Learn 实现二分类模型, 基于二分类模型实现多分类模型, 并且画出相应图像, 可计入题 (4,5) 得分, 但不计入题 (3) 得分.

**[符号约定]**  $\mathbf{x}$  是形状为  $(m, d)$  的矩阵, 其元素为 32 位浮点数; 在题 (1,3,4,5) 中,  $\mathbf{y}$  是形状为  $(m,)$  的向量, 其元素为 32 位整数; 在题 (2) 中,  $\mathbf{y}$  是形状为  $(m, 2)$  的向量, 其元素为 32 位浮点数. 其中:  $m$  为样例个数,  $d$  为样例维度; 标签从 0 开始, 例如共 20 类时, 标签依次是  $\{0, \dots, 19\}$ .

- (1) [5pts] 根据 `main.py` 中的框架代码, 实现 LDA 降维, 通过 `lda_sanity_check` 测试, 并在 `HW2.pdf` 中展示运行后的终端输出截图.
- (2) [5pts] 基于 (1) 分别把训练数据和测试数据降至两维, 并绘制在同一张散点图上, 在 `HW2.pdf` 中展示. 注意: 同类别的点应当使用同一颜色, 不同类别的数据点应当使用不同颜色.
- (3) [5pts] 分类任务可以被归结为一种特殊的回归任务, 可参考 `sklearn` 中的内容: [Link]. 对于二分类任务, 我们任选一类作为正类, 另一类成为负类. 对于正类样本  $\mathbf{x}_+$ , 约定  $y_+ = 1$ , 对于负类样本  $\mathbf{x}_-$ , 约定  $y_- = -1$ , 对于训练得到的分类器  $f$  和测试样例  $\mathbf{x}$ , 如果  $f(\mathbf{x}) \geq 0$  预测为正类, 否则预测为负类.  
根据框架代码, 按照上述约定实现基于岭回归的二分类模型, 通过 `classifier_2_sanity_check` 测试, 并在 `HW2.pdf` 中展示运行后的终端输出截图.
- (4) [5pts] 基于 (3) 中的二分类模型, 通过 OvR 策略将其拓展为多分类模型, 通过 `classifier_n_sanity_check` 测试, 最后在 `HW2.pdf` 中展示运行后的终端输出截图.  
提示: 判断测试样例的预测结果时, 可以依照教材实现, 即若有多个分类器预测为正类或者没有分类器预测为正类, 则考虑各分类器的预测置信度 ( $f(\mathbf{x})$  之值), 选择置信度最大的类别标记作为分类结果.
- (5) [5pts] 基于 (4) 绘制并在 `HW2.pdf` 中展示训练错误率和测试错误率随  $\lambda$  变化的折线图. 注意: 图像横轴为  $\lambda$ ; 训练错误率和测试错误率应当使用不同颜色的曲线.

**Solution.** 此处用于写解答 (中英文均可)

(1)

```

main x
C:\Users\86182\anaconda3\python.exe C:\Users\86182\Desktop\221300066季千焜ml\code\main.py
[2024-04-18 22:50:50.763307] 校验数据集开始...
[2024-04-18 22:50:50.776586] 校验数据集成功! (耗时0.0133秒钟)
[2024-04-18 22:50:50.776586] 加载数据集开始...
[2024-04-18 22:50:51.167698] 加载数据集成功! (耗时0.3911秒钟)
[2024-04-18 22:50:51.167698] 检查LDA的输入输出格式开始...
[2024-04-18 22:50:51.778356] 检查LDA的输入输出格式成功! (耗时0.6107秒钟)
[2024-04-18 22:50:51.778356] 检查Ridge2的输入输出格式开始...

```

图 7: LDACheck

(2)

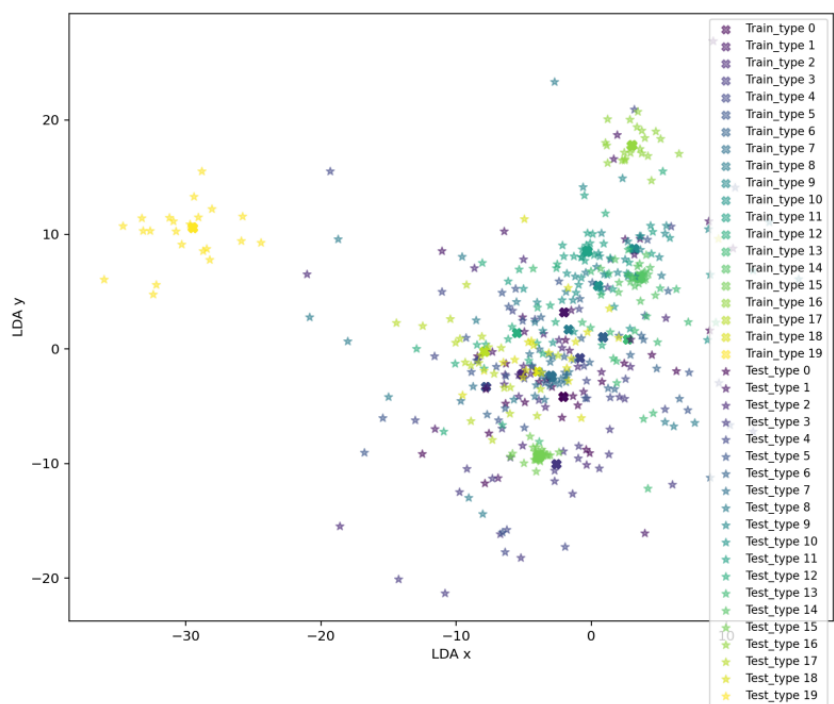


图 8: LDA

(3)(4)

```

[2024-04-18 22:53:11.647695] 检查Ridge2的输入输出格式开始...
[2024-04-18 22:53:11.667820] 检查Ridge2的输入输出格式成功! (耗时0.0201秒钟)
[2024-04-18 22:53:11.667820] 检查RidgeN的输入输出格式开始...
[2024-04-18 22:53:11.969104] 检查RidgeN的输入输出格式成功! (耗时0.3013秒钟)

```

图 9: classifierCheck

(5)

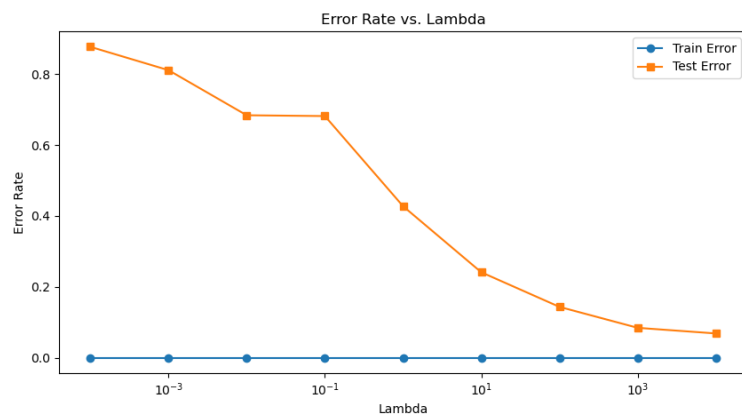


图 10: 训练错误率和测试错误率随  $\lambda$  变化的折线图

## Acknowledgments

允许与其他同样未完成作业的同学讨论作业的内容, 但需在此注明并加以致谢; 如在作业过程中, 参考了互联网上的资料或大语言模型的生成结果, 且对完成作业有帮助的, 亦需注明并致谢.