机器学习导论习题二

参考答案

2024 年 5 月 7 日

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 $\boldsymbol{v} \in \mathbb{R}^d$, 定义 $\|\boldsymbol{v}\|_2^2 = \boldsymbol{v}^\top \boldsymbol{v}$.

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

在多元线性回归中增加约束项, 即成为**岭回归 (Rigde 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.

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

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

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

(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 λ 已经确定, 此时岭回归的解记作 $\boldsymbol{w}^\star$, 请推导出 $\boldsymbol{w}^\star$ 的表达式.

(4) **[5pts] 在 (3) 的基础上**, 请推导出 λ 的下界 (关于 ρ 的函数), 并据此回答: ρ 减小时, 若希望保持 (1.2) 和 (1.3) 的解一致, 需要怎样调整 λ?
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: $\hat{\boldsymbol{w}}$ 接近于零向量. 归纳偏好作用: 能够**缓解**过拟合.
至少从一个角度, 解释为什么符合 "奥卡姆剃刀" 原则, 有正则化为什么比无正则化更 "简单".

- $\hat{\boldsymbol{w}}$ 接近于零向量, 相当于缩小了假设空间.
- $\hat{\boldsymbol{w}}$ 对训练数据更不敏感, **缓解**异常值带来的偏差, 相当于缩小了假设空间.
- 矩阵 $\mathbf{X}$ 不满秩时仍可以求逆, 无需引入 Moore-Penrose Pseudo Inverse.

(2) 列出 KKT 条件即得.

(3) 列出 Fermat's Optimality Condition ($\mathbf{0} \in \partial J(\hat{\boldsymbol{w}})$) 即得, $\boldsymbol{w}^\star = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\boldsymbol{y}$.

(4) $\rho \to 0, \lambda \to \infty$. 下界:

$$\lambda \geqslant \frac{\left\|\mathbf{X}^\top\boldsymbol{y}\right\|}{\|\boldsymbol{w}^\star\|} - \sigma^2_{\max}(\mathbf{X}) \geqslant \frac{\left\|\mathbf{X}^\top\boldsymbol{y}\right\|}{\rho} - \sigma^2_{\max}(\mathbf{X})$$

证明要点:

$$\left(\sigma^2_{\max}(\mathbf{X}) + \lambda\right)\|\boldsymbol{w}^\star\| \geqslant \left\|\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)\boldsymbol{w}^\star\right\| = \left\|\mathbf{X}^\top\boldsymbol{y}\right\|$$

上述证明基于赋范向量空间的基本性质 (三角不等式), 也可在 $\mathbb{R}^d$ 上基于特征值分解等初等方法证明.

2 [20pts] 决策树的构建流程

> **[注意事项]** 本题可使用 PowerPoint®, Visio® 等软件软件绘制决策树, 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

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

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) `OCT-PD=?`

$$\text{Gini\_index}(D_{\text{train}}, 1) = \frac{4}{8}\left(1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2\right)+\frac{4}{8}\left(1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2\right)$$

$$\text{Gini\_index}(D_{\text{train}}, 2) = \frac{5}{8}\left(1-\left(\frac{4}{5}\right)^2-\left(\frac{1}{5}\right)^2\right)+\frac{3}{8}\left(1-\left(\frac{3}{3}\right)^2-\left(\frac{0}{3}\right)^2\right)(\star)$$

$$\text{Gini\_index}(D_{\text{train}}, 3) = \frac{5}{8}\left(1-\left(\frac{3}{5}\right)^2-\left(\frac{2}{5}\right)^2\right)+\frac{3}{8}\left(1-\left(\frac{1}{3}\right)^2-\left(\frac{2}{3}\right)^2\right)$$

`OCT-PD=0`, 无需划分, 当前结点包含的样本 RD 诊断阴性.

`OCT-PD=1; US=?`

$\mathrm{Gini\_index}(D_{\text{train}}, 1) = \frac{2}{5}\left(1-\left(\frac{1}{2}\right)^2-\left(\frac{0}{2}\right)^2\right)+\frac{3}{5}\left(1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2\right)$

$\mathrm{Gini\_index}(D_{\text{train}}, 3) = \frac{3}{5}\left(1-\left(\frac{3}{3}\right)^2-\left(\frac{0}{3}\right)^2\right)+\frac{2}{5}\left(1-\left(\frac{1}{2}\right)^2-\left(\frac{0}{2}\right)^2\right)(\star)$

`OCT-PD=1; US=1`, 无需划分, 当前结点包含的样本 RD 诊断阳性.

`OCT-PD=1; US=0; UW-OCT=?` 只剩一个属性, 无需选择属性.

`OCT-PD=1; US=0; UW-OCT=1`, 无需划分, 当前结点包含的样本 RD 诊断阳性.

`OCT-PD=1; US=0; UW-OCT=0`, 无需划分, 当前结点包含的样本 RD 诊断阴性.

```
               OCT-PD=?
                   /   \
         OCT-PD=1 /     \ OCT-PD=0
                 /       \
              US=?      RD=0
              /   \
       US=1  /     \ US=0
            /       \
        RD=1  UW-OCT=0
                  /   \
       UW-OCT=1  /     \ UW-OCT=0
                /       \
            RD=1       RD=0
```

(2) `OCT-PD=?` 划分前 62.5%, 划分后 75%, 决策是划分.

`OCT-PD=1; US=?` 划分前 75%, 划分后 100%, 决策是划分.

`OCT-PD=1; US=0; UW-OCT=?` 划分前 100%, 划分后 87.5%, 决策是禁止划分.

```
               OCT-PD=?
                   /   \
         OCT-PD=1 /     \ OCT-PD=0
                 /       \
              US=?      RD=0
              /   \
       US=1  /     \ US=0
            /       \
        RD=1       RD=0
```

(3) UW-OCT 是无关特征; 剪枝抛弃的特征是无关特征; 深度越深越可能是无关特征.

可以认为数据集刻画了分布 $p(\boldsymbol{x}, y)$, 而模型刻画了分布 $p(y \mid \boldsymbol{x})$.

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, 信息增益可以重写成互信息 (Mutual Information)

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

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 $\mathrm{MI}(E, F)$, $\mathrm{Ent}(E)$, $\mathrm{Ent}(F)$ 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

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

在 (3) 的基础上, 如果使用归一化互信息作为划分准测, 与使用增益率作为划分准测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 $0 \leqslant \mathrm{MI}(E, F) \leqslant \min\{\mathrm{Ent}(E), \mathrm{Ent}(F)\}$.

Solution. 此处用于写解答 (中英文均可)

(1) 如果当前节点包含的样本不全属于同一类别, 那么划分后的纯度函数值**严格大于**划分前的纯度函数值.

(2) $E = \{\boldsymbol{x}^j = v_k^j \mid k \in [n_j]\}$, $F = \{\boldsymbol{x}^\ell = k \mid k \in [N]\}$; $\mathrm{Gain\_ratio} = \dfrac{\mathrm{MI}(E, F)}{\mathrm{Ent}(E)}$. 需要分别证明, $\mathrm{Gain} = \mathrm{Ent}(F) - \mathrm{Ent}(F \mid E) = \mathrm{MI}(E, F)$, $\mathrm{IV} = \mathrm{Ent}(E)$, 从而 $\mathrm{Gain\_ratio} = \dfrac{\mathrm{Gain}}{\mathrm{IV}}$.

(3) 未必相同.

得 5pts 的作答:

- 基于公式分析 NMI 和 Gain_ratio, 例如做差或者做商.
- 给出反例的数据集.

得 4pts 的作答:

- 分析过程正确, 误用 "糖水不等式", 得出了错误的结论.
- 给出反例, 但是没有给出相应的数据集. 原因: 五个量只有三个自由度, 没有说明取值的存在性.
  例如, 如下反例只能得 4pts,
  $\mathrm{MI}(E_1, F) = 1$, $\mathrm{MI}(E_2, F) = 49$, $\mathrm{Ent}(F) = 10$, $\mathrm{Ent}(E_1) = 2$, $\mathrm{Ent}(E_2) = 100$,
  $\dfrac{\mathrm{MI}(E_1, F)}{\mathrm{Ent}(E_1)} > \dfrac{\mathrm{MI}(E_2, F)}{\mathrm{Ent}(E_2)}, \dfrac{2 \cdot \mathrm{MI}(E_1, F)}{\mathrm{Ent}(E_1) + \mathrm{Ent}(F)} < \dfrac{2 \cdot \mathrm{MI}(E_2, F)}{\mathrm{Ent}(E_2) + \mathrm{Ent}(F)}$

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 $\boldsymbol{w}^\mathsf{T}\mathbf{S}_\mathrm{w}\boldsymbol{w}$ 与最大化类间散度 $\boldsymbol{w}^\mathsf{T}\mathbf{S}_\mathrm{b}\boldsymbol{w}$, **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: $\max_{\boldsymbol{w}} \boldsymbol{w}^\mathsf{T}\mathbf{S}_\mathrm{t}\boldsymbol{w}$.

(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 $y = \boldsymbol{w}^\mathsf{T}\boldsymbol{x}$ 代表 LDA 的投影方向. 请画出图 1 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.

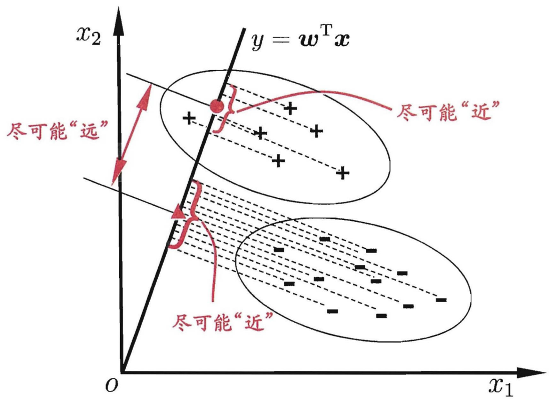

图 1: 教材图 3.3

(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: $\mathbf{S}_\mathrm{w} = \sum_{i=1}^N \mathbf{S}_{\mathrm{w}i}$. 对于第 i 个类别的类内散度矩阵定义如下: $\mathbf{S}_{\mathrm{w}i} = \sum_{\boldsymbol{x}\in X_i}(\boldsymbol{x}-\boldsymbol{\mu}_i)(\boldsymbol{x}-\boldsymbol{\mu}_i)^\mathsf{T}$. 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: $\mathbf{S}_\mathrm{b} = \sum_{i=1}^N \mathbf{S}_{\mathrm{b}i}$. 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: $\mathbf{S}_{\mathrm{b}i} = m_i(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^\mathsf{T}$.

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.

$$
\begin{aligned}
\max_{\mathbf{W}} \quad & J(\mathbf{W}) = \frac{\operatorname{tr}\left(\mathbf{W}^\mathsf{T}\mathbf{S}_\mathrm{b}\mathbf{W}\right)}{\operatorname{tr}\left(\mathbf{W}^\mathsf{T}\mathbf{S}_\mathrm{w}\mathbf{W}\right)} = \frac{\operatorname{tr}\left(\mathbf{W}^\mathsf{T}\left(\sum_{i=1}^N \mathbf{S}_{\mathrm{b}i}\right)\mathbf{W}\right)}{\operatorname{tr}\left(\mathbf{W}^\mathsf{T}\left(\sum_{i=1}^N \mathbf{S}_{\mathrm{w}i}\right)\mathbf{W}\right)} = \frac{\frac{1}{N}\sum_{i=1}^N \operatorname{tr}\left(\mathbf{W}^\mathsf{T}\mathbf{S}_{\mathrm{b}i}\mathbf{W}\right)}{\frac{1}{N}\sum_{i=1}^N \operatorname{tr}\left(\mathbf{W}^\mathsf{T}\mathbf{S}_{\mathrm{w}i}\mathbf{W}\right)} \\
\text{s.t.} \quad & \mathbf{W}^\mathsf{T}\mathbf{W} = \mathbf{I}_{d'}
\end{aligned}
\tag{4.1}
$$

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:

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

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 $\mathbf{A} \succcurlyeq \mathbf{B}$ 或 $\mathbf{B} \preccurlyeq \mathbf{A}$. 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. 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:

$$\begin{aligned} \max_{\mathbf{W},r} \quad & r \\ \text{s.t.} \quad & r \cdot \operatorname{tr}(\mathbf{S}_{\mathrm{w}}\mathbf{M}) - \operatorname{tr}(\mathbf{S}_{\mathrm{b}}\mathbf{M}) \leqslant 0 \\ & -r \leqslant 0 \\ & \mathbf{O} \preccurlyeq \mathbf{M} \preccurlyeq \mathbf{I}_{d'} \\ & \operatorname{tr}(\mathbf{M}) = d' \end{aligned} \tag{4.3}$$

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).
本题表述存在问题, 形如 $x \cdot y - z \leqslant 0$ 或者 $x \cdot y - z = 0$ 的约束不构成凸集合, 正确的表述应为证明 (4.3) 除 $r \cdot \operatorname{tr}(\mathbf{S}_{\mathrm{w}}\mathbf{M}) - \operatorname{tr}(\mathbf{S}_{\mathrm{b}}\mathbf{M}) \leqslant 0$ 以外构成凸优化问题.

(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 $r \cdot \operatorname{tr}(\mathbf{S}_{\mathrm{w}}\mathbf{M}) - \operatorname{tr}(\mathbf{S}_{\mathrm{b}}\mathbf{M}) = 0$ 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, 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) 大致如下图 2.

(2) PCA 不包含标签信息, 最大化类别无关的全局散度.

(3) 求解 $\mathbf{W}$ 依赖于 i, j, 求解 i, j 依赖于 $\mathbf{W}$.

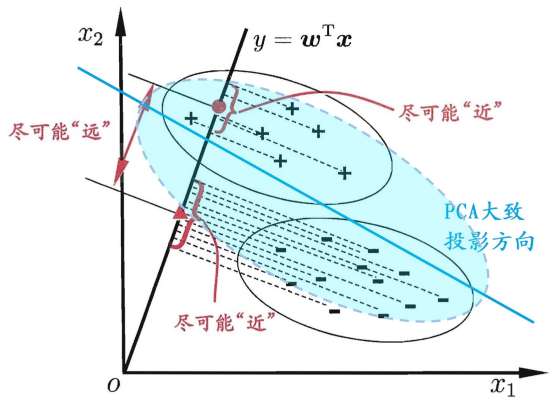

图 2: 教材图 3.3, PCA 的大致投影方向

(4) 省略约束条件 $-r \leqslant 0$, $\mathbf{O} \preccurlyeq \mathbf{M} \preccurlyeq \mathbf{I}_d$, $\operatorname{tr}(\mathbf{M}) = d'$, 如下三种形式均可. 证明要点: **半正定锥 (Positive Semi-definite Cone) 凸**; 凸函数小于等于一定值的解集凸; $\operatorname{tr}(\cdot)$, $\max$, $-\min$ 和线性运算保凸.

$$
\begin{aligned}
\max_{\mathbf{M},r,a,b}\quad & r \\
\text{s. t.}\quad & r \cdot \max_i \operatorname{tr}(\mathbf{S}_{\mathrm{w}i}\mathbf{M}) - \min_j \operatorname{tr}\left(\mathbf{S}_{\mathrm{b}j}\mathbf{M}\right) \leqslant 0 \\
\text{s. t.}\quad & r \cdot \operatorname{tr}(\mathbf{S}_{\mathrm{w}i}\mathbf{M}) - \operatorname{tr}\left(\mathbf{S}_{\mathrm{b}j}\mathbf{M}\right) \leqslant 0 && \forall i, j \in [N] \\
\text{s. t.}\quad & r \cdot a - b \leqslant 0 \\
& \operatorname{tr}(\mathbf{S}_{\mathrm{w}i}\mathbf{M}) - a \leqslant 0 && \forall i \in [N] \\
& b - \operatorname{tr}\left(\mathbf{S}_{\mathrm{b}j}\mathbf{M}\right) \leqslant 0 && \forall j \in [N]
\end{aligned}
$$

(5-a) $\operatorname{tr}(\mathbf{ABC}) = \operatorname{tr}(\mathbf{BCA}) = \operatorname{tr}(\mathbf{CBA})$; $\mathbf{M} = \mathbf{W}\mathbf{W}^{\mathrm{T}}$, $\mathbf{M}$ 无法表达正交性约束.

(5-b) 注意 $\{\mathbf{M}\}$ 和 $\{\mathbf{W}\mathbf{W}^T\}$ 的正交相似变换不变性. 幂集记作 $\mathcal{P}$, 考虑集合函数

$$
\sigma : \mathcal{P}(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^{d\times d}), \qquad \sigma(\{\boldsymbol{\lambda}\}) = \{\mathbf{Q}^{-1}\operatorname{diag}(\boldsymbol{\lambda})\mathbf{Q} \mid \mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}\}.
$$

注意到 $\mathbf{M} = \sum_{i=1}^{d} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^{\mathrm{T}}$, $\mathbf{W}\mathbf{W}^{\mathrm{T}} = \sum_{i=1}^{d'} \boldsymbol{v}_i \boldsymbol{v}_i^{\mathrm{T}}$. 因此 $\{\mathbf{M}\}$ 由 $\{\boldsymbol{\lambda} \mid \mathbf{0} \preccurlyeq \boldsymbol{\lambda} \preccurlyeq \mathbf{1}, \mathbf{1}^{\mathrm{T}}\boldsymbol{\lambda} = d'\}$ 生成, $\{\mathbf{W}\mathbf{W}^{\mathrm{T}}\}$ 由 $\{\boldsymbol{\lambda} \mid \forall i \in \{1, \cdots, d\}, \boldsymbol{\lambda}_i \in \{0, 1\}, \mathbf{1}^{\mathrm{T}}\boldsymbol{\lambda} = d'\}$ (d 维 “d'-hot” 向量) 生成.
不难发现 $\sigma^{-1}(\{\mathbf{M}\})$ 是一个闭合多面体/凸集 (Polygon / Convex Set), 而 $\sigma^{-1}(\{\mathbf{W}\mathbf{W}^{\mathrm{T}}\})$ 恰好构成其顶点/极点 (Vertex / Extreme Point), 注意到 σ 具有保凸性, 从而凸包得证.
注 1: 在正交相似变换 (Orthogonal Similar Transformation), $\tilde{\mathbf{M}} = \mathbf{Q}^{-1}\mathbf{M}\mathbf{Q}$, $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$.
注 2: 极点是不能用凸集内其他点组合表示的点, 凸集中所有点可表示成极点的组合.
注 3: 初等证明等价于 d 维 “d'-hot” 向量对应的 Toeplitz Matrix 的非奇异性, 非常困难.

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] `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, `y` 是形状为 `(m,)` 的向量, 其元素为 32 位整数; 在题 (2) 中, `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 $\{0, \cdots, 19\}$.

(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 $\boldsymbol{x}_+$, 约定 $y_+ = 1$, 对于负类样本 $\boldsymbol{x}_-$, 约定 $y_- = -1$, 对于训练得到的分类器 f 和测试样例 $\boldsymbol{x}$, 如果 $f(\boldsymbol{x}) \geqslant 0$ 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 ($f(\boldsymbol{x})$ 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 λ 变化的折线图**. 注意: 图像横轴为 λ; 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)

(1,3) NumPy 代码与矩阵运算一一对应.

(2) 特点: 训练数据完全可分, 泛化能力很差.

(4) `RidgeN` 递归调用 `Ridge2`.

(5) 特点: 训练数据误差为零, 泛化能力随正则化力度增大而增强.

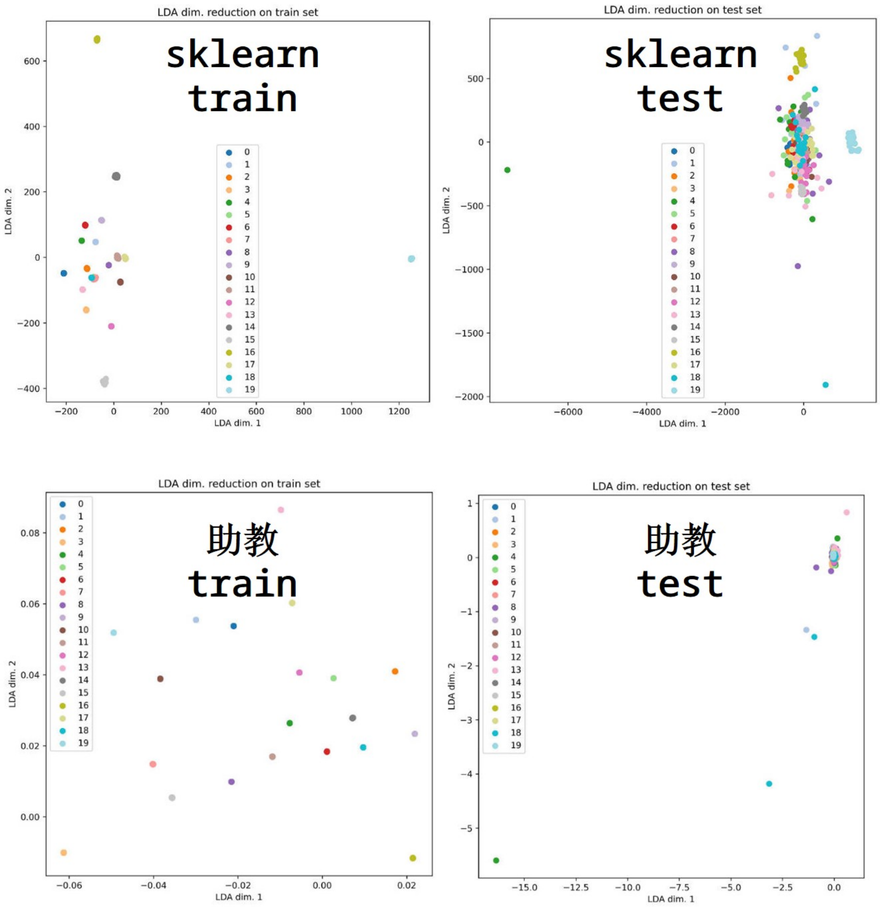

图 3: LDA

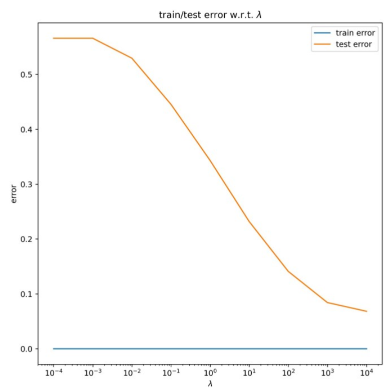

图 4: RidgeN