线性判别分析（Linear Discriminant Analysis, LDA）

研究生课程深入讲解

赵凯

上海大学通信与信息工程学院

https://kaizhao.net/teaching/

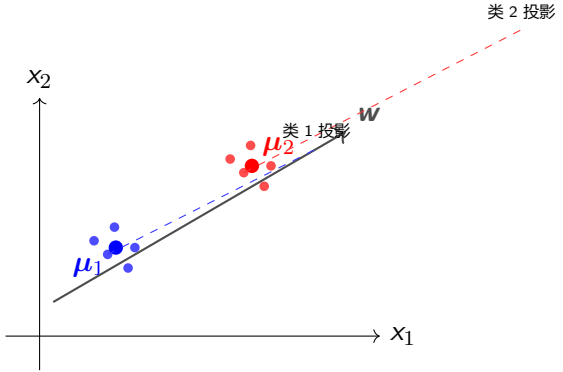
本次课程结构

1. 问题背景与直观理解
2. 数学建模：散度矩阵与 Fisher 准则
3. 广义特征值问题与求解
4. 二分类 LDA 的闭式解与几何解释
5. 概率视角与贝叶斯最优性
6. 与 PCA 的比较及优缺点
7. 小结

LDA 是什么?

- Linear Discriminant Analysis（线性判别分析）是一个经典的**监督学习**方法。
- 核心目标：寻找一个**线性投影**，使得
  - 同一类样本在投影后尽量**聚集**（类内紧凑）
  - 不同类样本在投影后尽量**分开**（类间分离）
- 主要用途：
  - 监督降维（supervised dimensionality reduction）
  - 线性分类器（linear discriminant classifier）
  - 特征提取（feature extraction）
- 与 PCA 的关键区别：PCA 不使用标签信息，LDA 使用类别标签。

几何直观：二维二类投影示意图



- 目标：找到方向 $\boldsymbol{w}$，使得两类投影后的距离尽可能大，类内方差尽可能小。
- 一维投影后，只需要在投影轴上设定一个阈值即可进行分类。

数据与符号设定

- 假设有 C 个类别：

$$\mathcal{D}_c = \{\boldsymbol{x}_i^{(c)} \in \mathbb{R}^d \mid i = 1, \ldots, N_c\}, \quad c = 1, \ldots, C$$

- 总样本数：$N = \sum_{c=1}^{C} N_c$。
- 类均值与整体均值：

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{\boldsymbol{x}_i \in \mathcal{D}_c} \boldsymbol{x}_i, \qquad \boldsymbol{\mu} = \frac{1}{N} \sum_{c=1}^{C} N_c \boldsymbol{\mu}_c.$$

- 希望找到一个投影矩阵：

$$\boldsymbol{W} \in \mathbb{R}^{d \times m}, \quad m \leq C - 1,$$

将高维数据投影到 m 维：

$$\boldsymbol{z} = \boldsymbol{W}^\top \boldsymbol{x}.$$

类内散度矩阵 S_W

定义：Within-class Scatter

$$S_W = \sum_{c=1}^{C} \sum_{\boldsymbol{x}_i \in \mathcal{D}_c} (\boldsymbol{x}_i - \boldsymbol{\mu}_c)(\boldsymbol{x}_i - \boldsymbol{\mu}_c)^\top.$$

- 直观理解：度量“同一类内部”的扩散程度。
- S_W 越小表示类内越紧凑。
- 投影后类内散度：

$$S'_W = \boldsymbol{W}^\top S_W \boldsymbol{W}.$$

类间散度矩阵 S_B

定义：Between-class Scatter

$$S_B = \sum_{c=1}^{C} N_c(\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^\top.$$

- 直观理解：度量“不同类别均值之间”的分离程度。
- S_B 越大表示类间越分散，类别越**可分**。
- 投影后类间散度：

$$S_B' = \boldsymbol{W}^\top S_B \boldsymbol{W}.$$

Fisher 判别准则

- LDA 的核心思想：在投影空间中最大化类间散度、最小化类内散度。
- 对于多维投影矩阵 $\boldsymbol{W}$，常用的目标函数为：

$$J(\boldsymbol{W}) = \frac{\det\left(\boldsymbol{W}^\top S_B \boldsymbol{W}\right)}{\det\left(\boldsymbol{W}^\top S_W \boldsymbol{W}\right)}.$$

- 对于一维投影向量 $\boldsymbol{w}$，简化为：

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top S_B \boldsymbol{w}}{\boldsymbol{w}^\top S_W \boldsymbol{w}}.$$

- 这是一个**广义瑞利商**（generalized Rayleigh quotient）优化问题。

从优化到广义特征值问题

- 最大化

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top S_B \boldsymbol{w}}{\boldsymbol{w}^\top S_W \boldsymbol{w}}$$

  在约束 $\boldsymbol{w}^\top S_W \boldsymbol{w} = 1$ 下。

- 拉格朗日函数：

$$\mathcal{L}(\boldsymbol{w}, \lambda) = \boldsymbol{w}^\top S_B \boldsymbol{w} - \lambda(\boldsymbol{w}^\top S_W \boldsymbol{w} - 1).$$

- 对 $\boldsymbol{w}$ 求导并置零：

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = 2S_B \boldsymbol{w} - 2\lambda S_W \boldsymbol{w} = 0 \quad \Rightarrow \quad S_B \boldsymbol{w} = \lambda S_W \boldsymbol{w}.$$

- 得到**广义特征值问题**：

$$S_B \boldsymbol{w} = \lambda S_W \boldsymbol{w}.$$

多维情形与降维维度

- 对于多维投影 $\boldsymbol{W} \in \mathbb{R}^{d \times m}$，可以将问题扩展为：

$$\max_{\boldsymbol{W}} \quad J(\boldsymbol{W}) = \frac{\det(\boldsymbol{W}^\top S_B \boldsymbol{W})}{\det(\boldsymbol{W}^\top S_W \boldsymbol{W})}.$$

- 结论：最优的 $\boldsymbol{W}$ 由广义特征值问题

$$S_B \boldsymbol{w}_k = \lambda_k S_W \boldsymbol{w}_k$$

  的前 m 个最大特征值对应的特征向量组成。
- S_B 的秩最多为 $C-1$，因此 LDA 最多将维度降到

$$m_{\max} = C - 1.$$

- 典型应用：多类问题的可视化（2D 或 3D LDA 子空间）。

二分类情形：闭式解

- 对于 $C = 2$ 的情况：

$$S_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top.$$

- 可以证明最优解为：

$$\boldsymbol{w}^\star = S_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

- 直观理解：
  - $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$：指向两类均值连线的方向。
  - S_W^{-1}：对类内散度进行“白化”和重加权，减弱方差大的方向。
- 投影后在一维轴上选择阈值 t，得到决策规则：

$$\boldsymbol{w}^{\star\top}\boldsymbol{x} \underset{y=2}{\overset{y=1}{\gtrless}} t.$$

高斯判别模型与 LDA

- 统计假设：

$$\boldsymbol{x} \mid y = c \sim \mathcal{N}(\boldsymbol{\mu}_c, \Sigma),$$

即所有类别共享同一个协方差矩阵 Σ。

- 根据贝叶斯规则，可以得到判别函数：

$$\delta_c(\boldsymbol{x}) = \boldsymbol{x}^\top \Sigma^{-1} \boldsymbol{\mu}_c - \frac{1}{2} \boldsymbol{\mu}_c^\top \Sigma^{-1} \boldsymbol{\mu}_c + \log \pi_c,$$

其中 π_c 为先验概率。

- 决策规则：选择使得 $\delta_c(\boldsymbol{x})$ 最大的类别 c。
- 由于判别函数关于 $\boldsymbol{x}$ 是**线性**的，因此决策边界是线性的，这与 LDA 的线性判别边界一致。

LDA vs. PCA

PCA

- 无监督：不使用类别标签。
- 目标：最大化整体数据方差。
- 求解：S_T 的特征分解，其中 S_T 为总散度矩阵。
- 得到的方向不一定有利于分类。

LDA

- 有监督：显式利用类别标签。
- 目标：最大化类间可分性，最小化类内散度。
- 求解：广义特征值问题 $S_B\boldsymbol{w} = \lambda S_W\boldsymbol{w}$。
- 更适合用于分类前的降维和特征提取。

LDA 的优缺点与常见扩展

优点：

- 理论清晰，数学形式优雅。
- 与高斯判别模型、贝叶斯最优分类器有紧密联系。
- 计算成本相对较低，可作为很多复杂方法的基线。

局限性：

- 依赖于“各类协方差相同”的假设。
- 仅适用于线性可分或近似线性的情形。
- 高维小样本（$d \gg N$）时 S_W 奇异，需要正则化。
- 降维维度上限为 $C-1$。

一些常见变体与改进

- 正则化 LDA（Regularized LDA）：

$$S_W \leftarrow S_W + \lambda I,$$

  用于缓解 S_W 奇异问题。
- 核 LDA（Kernel LDA）：
  - 在高维特征空间中执行 LDA 实现非线性判别。
- 半监督 LDA（Semi-supervised LDA）：
  - 利用少量标注样本 + 大量未标注样本。
- 深度 LDA（Deep LDA-like loss）：
  - 将类间/类内散度思想融入深度网络的损失函数中。

本课小结

- LDA 的核心思想：
  - 最大化类间散度 S_B，最小化类内散度 S_W。
  - 通过广义特征值分解获得最优投影方向。
- 数学上：
  - Fisher 判别准则 $J(\boldsymbol{w}) = \dfrac{\boldsymbol{w}^\top S_B \boldsymbol{w}}{\boldsymbol{w}^\top S_W \boldsymbol{w}}$。
  - 求解 $S_B \boldsymbol{w} = \lambda S_W \boldsymbol{w}$。
- 统计视角：
  - 等协方差高斯判别模型下的贝叶斯最优线性分类器。
- 实践中：
  - 常用于监督降维、人脸识别、医学影像分类、工业质检特征压缩等。

课后思考题

1. 推导二分类 LDA 的闭式解 $\boldsymbol{w}^\star = S_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$。
2. 在高维小样本情形下，如何设计合适的正则化策略来稳定 LDA？
3. 思考：如果各类别协方差矩阵不同，会发生什么？为什么需要 QDA？
4. 将 LDA 与 logistic regression 的决策边界进行比较，它们在什么条件下会接近？

谢谢大家！

Questions?