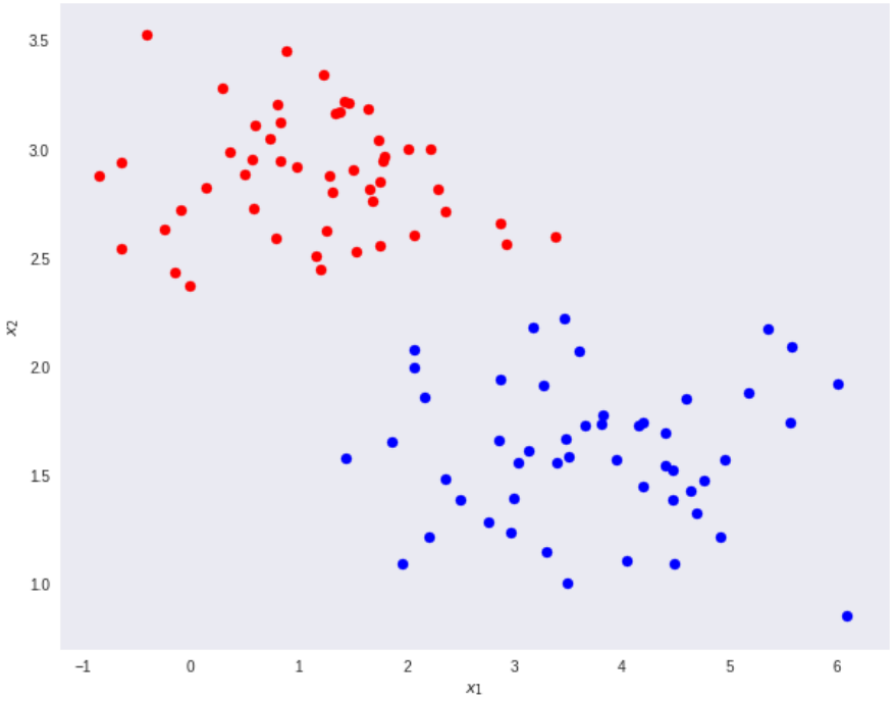
Linear Discriminant Analysis (线性判别分析) 和 Fisher 线性判别

HZAU-BioNLP and DeepSeek

2026 年 4 月 2 日

1 梗概

本讲义旨在系统阐述线性判别分析（LDA）的理论框架，考虑的是一个数据带有标签的分类问题。设样本观测向量为 $\vec{x}$，类别为 $y = 1, 0$。LDA 分类问题是在仅给定观测值 $\vec{x}$ 的情况下，为其找到一个线性的判别分界面。[1]



讲义首先从“Fisher 判别条件”的几何直观出发建立判别条件，随后引入《西瓜书》中的散度矩阵给出通用算法解。最后通过“QDA 判别条件”的似然比对数，分析其在正态分布下的贝叶斯最优性。

QDA 和 Fisher 是两个不同的线性判别条件，在其下均都能导出 LDA 寻求的投影向量 w，使数据投影到 w 后，正类和负类能够分离。

两者的思想略有不同，前者通过似然比，后者通过类间和类内的方差比。

为了推导出相同的结果。QDA 判别条件下还需要追加假设数据服从正态分布。而 Fisher 线性判别不需要这个假设。

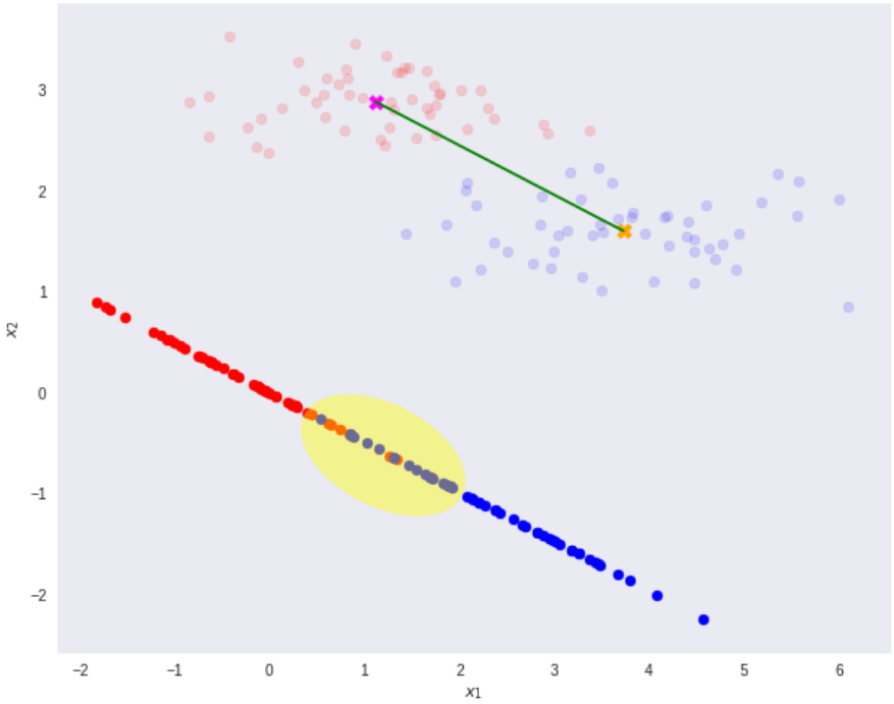
另外，QDA 还需追加假设：正、负样本的协方差是相同的 ($\Sigma_0 = \Sigma_1$)。而 Fisher 线性判别则不需要这个假设。

[1]这不是 PCA 所应对的无监督学习场景，这里考虑的是一个数据带有标签的分类问题。(但这些算法都考虑投影，这是提醒各位注意的地方——建议大家回忆《PCA》一课中的投影相关计算。)

2 Fisher 线性判别下的 LDA

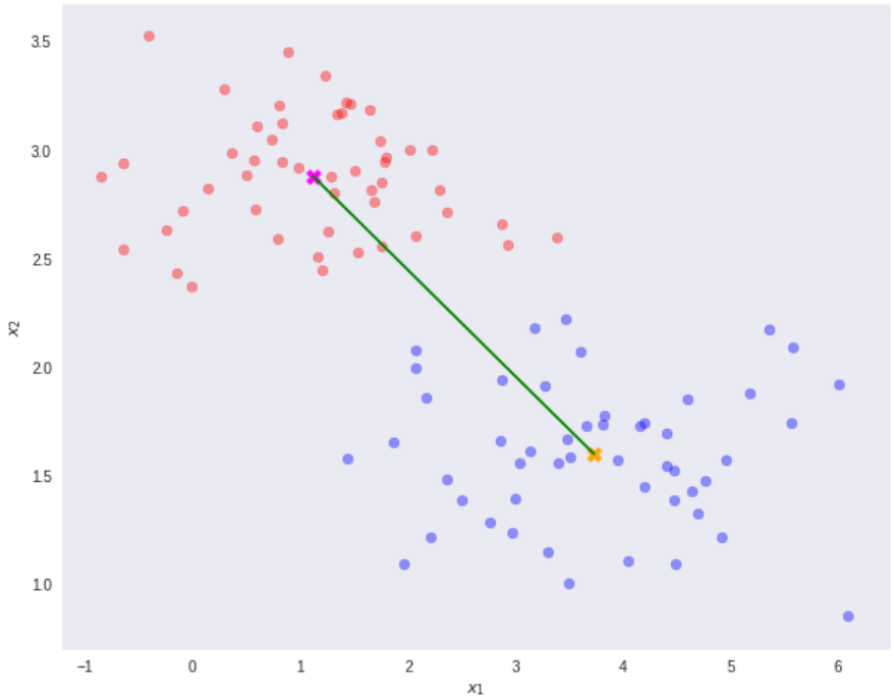
2.1 几何直观 (The Geometric Intuition)

正类和负类数据的均值与协方差参数分别为 $(\vec{\mu}_0, \Sigma_0)$ 和 $(\vec{\mu}_1, \Sigma_1)$。注意，此时数据服从的分布是不定的。



上图下方示意的是 w 投影向量，以及投影在 w 上后，可观测到的坐标聚集和分离的情形。

2.2 Fisher 线性判别函数



Fisher 线性判别的思想非常朴素：给定训练样例集，设法将样例投影到一条直线上。其核心目标是寻找一个投影方向 $\vec{w}$，使得：

- **类内尽可能近**：同类样例的投影点尽可能接近，即投影后的协方差尽可能小。
- **类间尽可能远**：异类样例的投影点尽可能远离，即类中心之间的距离尽可能大。

2.3 类间方差和类内方差

这是 Fisher 线性判别关注的类间和类内方差：

- 类间方差 [2]：$\sigma^2_{\text{between}} = (\vec{w}\cdot\vec{\mu}_1 - \vec{w}\cdot\vec{\mu}_0)^2$
- 类内方差 [3]：$\sigma^2_{\text{within}} = \vec{w}^T\Sigma_1\vec{w} + \vec{w}^T\Sigma_0\vec{w}$

将两项代入目标函数 $J(\vec{w})$：

$$J(\vec{w}) = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{(\vec{w}^T\vec{\mu}_1 - \vec{w}^T\vec{\mu}_0)^2}{\vec{w}^T\Sigma_1\vec{w} + \vec{w}^T\Sigma_0\vec{w}} \tag{1}$$

为了利用矩阵运算求解，我们引入类间散度矩阵：

$$S_b = (\vec{\mu}_1 - \vec{\mu}_0)(\vec{\mu}_1 - \vec{\mu}_0)^T$$

和类内散度矩阵：

$$S_w = \Sigma_0 + \Sigma_1$$

此时目标函数变为：

$$J(\vec{w}) = \frac{\vec{w}^T S_b\vec{w}}{\vec{w}^T S_w\vec{w}}$$

由于 $J(\vec{w})$ 的分子和分母都是关于 $\vec{w}$ 的二次项，其解与 $\vec{w}$ 的长度无关，只与其方向有关。因此，我们可以设定一个约束条件 $\vec{w}^T S_w\vec{w} = 1$ 。

此时问题转化为：

$$\min_{\vec{w}} -\vec{w}^T S_b\vec{w} \quad \text{s.t. } \vec{w}^T S_w\vec{w} = 1 \tag{2}$$

引入拉格朗日乘子 λ，构造函数：

$$L(\vec{w},\lambda) = -\vec{w}^T S_b\vec{w} + \lambda(\vec{w}^T S_w\vec{w} - 1)$$

对 $\vec{w}$ 求导并令其为零：

$$\frac{\partial L}{\partial \vec{w}} = -2S_b\vec{w} + 2\lambda S_w\vec{w} = 0$$

由此得到广义特征值方程：

$$S_b\vec{w} = \lambda S_w\vec{w}$$

注意到 $S_b\vec{w} = (\vec{\mu}_1 - \vec{\mu}_0)(\vec{\mu}_1 - \vec{\mu}_0)^T\vec{w}$ 。其中 $(\vec{\mu}_1 - \vec{\mu}_0)^T\vec{w}$ 是一个标量，设为 k。

因此 $S_b\vec{w}$ 的方向恒为 $(\vec{\mu}_1 - \vec{\mu}_0)$ 。所以有

$$\lambda S_w\vec{w} = k(\vec{\mu}_1 - \vec{\mu}_0)$$

[2]请回忆《PCA》讲授部分的投影坐标结果。

[3]请回忆《PCA》讲授部分（这里另有一个数据 zero-mean 的假设）。

$$\vec{w} = \frac{k}{\lambda} S_w^{-1}(\vec{\mu}_1 - \vec{\mu}_0)$$

由于在 LDA 中，我们只关心投影向量 $\vec{w}$ 的方向，而不关心它的绝对长度（因为缩放 $\vec{w}$ 不会改变投影后的相对分布和分类结果），所以我们可以忽略前面的常数系数 $\frac{k}{\lambda}$，直接取其方向解。最终得出 $\vec{w}$ 的具体求解公式：

$$\vec{w} = S_w^{-1}(\vec{\mu}_1 - \vec{\mu}_0) \tag{3}$$

3 二次判别分析 (QDA) 追加假设情形下的 LDA

3.1 二次判别分析 (QDA)

QDA 假设条件概率密度函数 $p(\vec{x}|y=0)$ 和 $p(\vec{x}|y=1)$ 均为正态分布。在此假设下，贝叶斯最优解决方案是当似然比的对数大于某个阈值 T 时，预测为第二类（$y=1$）。因此，判别为第二类的条件可以列为下式：[4]

$$\begin{aligned} \log \frac{p(\vec{x}|y=1)}{p(\vec{x}|y=0)} &= \frac{1}{2}(\vec{x}-\vec{\mu}_0)^T\Sigma_0^{-1}(\vec{x}-\vec{\mu}_0) + \frac{1}{2}\ln|\Sigma_0| - \frac{1}{2}(\vec{x}-\vec{\mu}_1)^T\Sigma_1^{-1}(\vec{x}-\vec{\mu}_1) - \frac{1}{2}\ln|\Sigma_1| \\ &> T \end{aligned} \tag{4}$$

这个判别式的左侧表达式在代数上是一个关于 $\vec{x}$ 的二次型（Quadratic Form），因此，得到的分类器被称为 **二次判别分析 (QDA)** 分类器。

因为判别函数保留了二次项，所以决策边界（即 $L(\vec{x}) = T$ 的轨迹）在几何上不再是一条直线或一个平面，而是二次曲面。在二维空间中，边界可能是椭圆、抛物线或双曲线。在高维空间中，它是一个超二次曲面。

3.2 “协方差相同”假设下的 LDA

LDA 做出额外的简化同方差假设（$\Sigma_0 = \Sigma_1 = \Sigma$），此时决策标准简化为点积的阈值：

$$\text{其中：} \begin{cases} \vec{w}^T\vec{x} > c \\ \vec{w} = \Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0), \\ c = T + \frac{1}{2}\vec{w}^T(\vec{\mu}_1 + \vec{\mu}_0). \end{cases} \tag{5}$$

此时，$\vec{w}$ 的物理角色既是投影向量，也是决策平面的法向量。[5] 同时，还请将该结果与 (eq.3) 对比。

[4] 请参考附录 A 部分的结果，予以计算。

[5] 问题：请同学们绘图解释这个原因。

4 算法思考

4.1 矩阵不可逆带来的数值稳定性

在实际计算 S_w^{-1} 时，矩阵奇异（不可逆）是常见的，这给计算带来的不稳定性 [6]，通常对 S_w 进行奇异值分解 (SVD) 。通过将 S_w 分解为 $U\Sigma V^T$，我们可以计算其伪逆 $S_w^{-1} = V\Sigma^{-1}U^T$ 。

正则化手段：除了 SVD，实践中也常采用在 S_w 的对角线上加上一个微小的扰动项（如 ϵI），强行使其满秩，从而保证计算过程不会崩溃。

4.2 多分类扩展

当类别数量 $N > 2$ 时，我们不能再简单地投影到一条直线上，而是需要投影到一个低维的超平面。多分类 LDA 将样本投影到 $N-1$ 维空间。例如，如果有 3 个类别，它们在空间中构成的中心点最多能确定一个平面（2 维），因此投影到 2 维是最优的。有关它的讨论另辟章节。

参考文献

[1] Wikipedia: Linear Discriminant Analysis

[2] 周志华《西瓜书》p61-62.

[3] `https://sthalles.github.io/fisher-linear-discriminant/`

[6]首先，请回顾我们在讲解最小二乘线性回归时遇到的场景，样本数量远小于 Feature 维度的所谓“维度灾难”现象是很常见的，这直接导致协方差矩阵不可逆。
另外，此处 Σ_0 和 Σ_1 是各类的协方差矩阵。这里还请留意协方差矩阵的性质：任何协方差矩阵 Σ 都是半正定的。这意味着对于任意非零向量 $\vec{v}$，都有 $\vec{v}^T\Sigma\vec{v} \geq 0$。实际上，$\vec{v}^T\Sigma\vec{v}$ 代表了数据在 $\vec{v}$ 方向上的方差，而方差永远不可能为负数。

A 多元高斯分布的似然计算

* 高维高斯分布 / 多元高斯分布

R.V. $X=\begin{pmatrix}X_1\\X_2\\\vdots\\X_n\end{pmatrix}\sim N(\mu,\Sigma)$ ← 协方差矩阵；← 均值

$\mu=E[X]=\begin{pmatrix}E[X_1]\\E[X_2]\\\vdots\\E[X_n]\end{pmatrix}$,

$\Sigma_{ij}=cov(X_i,X_j)=E[(X_i-\mu_i)(X_j-\mu_j)]$ $\Rightarrow \Sigma=E[(X-\mu)(X-\mu)^T]$

pdf. $f(X_1,\cdots,X_n)=\dfrac{\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)}{\sqrt{(2\pi)^n|\Sigma|}}$

$f(x)$

① 回顾1维 pdf. $(n=1)$

$\mu=E[X]$ 为常数，Σ 退化 σ^2

$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

② $\sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}$ 称为 x 与 μ 的 Mahalanobis distance（马氏距离）

或称 X 与分布 P 之间的马氏距离（P 不需必须为 Gaussian. 可为任意分布.）（但 Σ 需正定）

③ 协方差矩阵 Σ 永为半正定.

证：$\forall y\in\mathbb{R}^n$, $y^T\Sigma y=y^TE[(x-\mu)(x-\mu)^T]y$

$=E[y^T(x-\mu)(x-\mu)^Ty]=E[(y^T(x-\mu))^2]\geq 0$

右注解

$\therefore\Sigma$ 半正定.

注：$X_{n\times n}$, $a_{n\times 1}$

$E[X]\cdot a=E[Xa]$

证明之.

*MLE 最大似然估计

MLE的计算可训练到以下部分，一是似然函数的计算，可是矩阵偏分的若干计算技巧。

· likelihood

$$\text{证据}\ p(\mathcal{O}) = \frac{\overbrace{p(\mathcal{O}|\theta)}^{\text{似然}} \cdot \overbrace{p(\theta)}^{\text{先验}}}{\underbrace{p(\theta|\mathcal{O})}_{\text{后验}}}$$

$\mathcal{O}$ 为 observation.

θ 为模型参数.

这是Bayes统计的基本概念.

取 $\mathcal{O} = \{X^{(1)}, \cdots, X^{(i)}, \cdots, X^{(m)}\}$

$\theta = \{\mu, \Sigma\}$

log-likelihood $L(\mathcal{O}, \theta) =$

$$\log \prod_{i=1}^{m} f(X^{(i)} | \mu, \Sigma)$$

$$= \log \prod_{i=1}^{m} \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(X^{(i)}-\mu)^T \Sigma^{-1} (X^{(i)}-\mu)\right)$$

$$= \sum_{i=1}^{m} \left[-\frac{n}{2} \cdot \log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(X^{(i)}-\mu)^T \Sigma^{-1} (X^{(i)}-\mu)\right]$$

$$= -\frac{nm}{2}\log 2\pi - \frac{m}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{m}\left[(X^{(i)}-\mu)^T \Sigma^{-1} (X^{(i)}-\mu)\right]$$

B 《西瓜书》关于 LDA 的求解步骤

$$J = \frac{\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w}} \,. \tag{3.35}$$

这就是 LDA 欲最大化的目标, 即 $\mathbf{S}_b$ 与 $\mathbf{S}_w$ 的“广义瑞利商”(generalized Rayleigh quotient).

若 $\boldsymbol{w}$ 是一个解, 则对于任意常数 α, $\alpha\boldsymbol{w}$ 也是式(3.35)的解.

如何确定 $\boldsymbol{w}$ 呢? 注意到式(3.35)的分子和分母都是关于 $\boldsymbol{w}$ 的二次项, 因此式(3.35)的解与 $\boldsymbol{w}$ 的长度无关, 只与其方向有关. 不失一般性, 令 $\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w} = 1$, 则式(3.35)等价于

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & -\boldsymbol{w}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{w} \\ \text{s.t.} \quad & \boldsymbol{w}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{w} = 1 \,. \end{aligned} \tag{3.36}$$

拉格朗日乘子法参见附录 B.1.

由拉格朗日乘子法, 上式等价于

$$\mathbf{S}_b\boldsymbol{w} = \lambda\mathbf{S}_w\boldsymbol{w} \,, \tag{3.37}$$

其中 λ 是拉格朗日乘子. 注意到 $\mathbf{S}_b\boldsymbol{w}$ 的方向恒为 $\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$, 不妨令

$$\mathbf{S}_b\boldsymbol{w} = \lambda(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \,, \tag{3.38}$$

代入式(3.37)即得

$$\boldsymbol{w} = \mathbf{S}_w^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \,. \tag{3.39}$$

奇异值分解参见附录 A.3.

考虑到数值解的稳定性, 在实践中通常是对 $\mathbf{S}_w$ 进行奇异值分解, 即 $\mathbf{S}_w = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$, 这里 $\mathbf{\Sigma}$ 是一个实对角矩阵, 其对角线上的元素是 $\mathbf{S}_w$ 的奇异值, 然后再由 $\mathbf{S}_w^{-1} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^{\mathrm{T}}$ 得到 $\mathbf{S}_w^{-1}$.