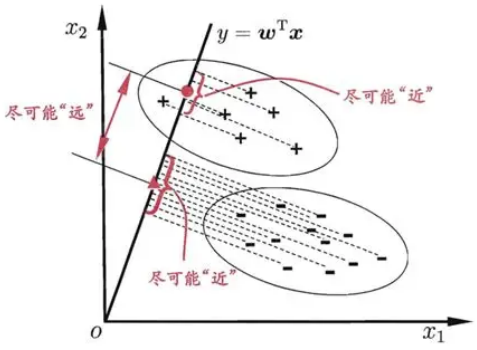
线性判别分析（LDA）

这部分内容在西瓜书中也存在，事实上是吴恩达老师没有讲LDA，而我在啃西瓜书的时候看到了，所以这部分就简单的介绍一下。

DeepSeek Prompt：你现在是一个人工智能大学教授，你要给你的本科学生讲述线性判别分析，需要使用Latex给出线性判别分析LDA的详细推导流程

好的，同学们。



今天我们来深入探讨一种经典的监督降维与分类方法——***线性判别分析（Linear Discriminant Analysis, LDA）***。它的核心思想非常直观：为数据找到一个投影方向，使得在这个方向上，***不同类别的数据点尽可能分开（类间方差大），而同类的数据点尽可能聚集（类内方差小）***。

我们将从目标函数出发，一步步推导出LDA的最优解。请大家准备好纸笔，跟上我的思路。

1. 问题设定与符号定义

假设我们的数据集有 C 个类别。定义以下符号：

- N：总样本数
- N_c：第 c 类中的样本数，满足 $\sum_{c=1}^{C} N_c = N$
- $\mathbf{x}_i^{(c)}$：第 c 类中的第 i 个样本（d 维列向量）
- $\boldsymbol{\mu}_c$：第 c 类的均值向量（$d \times 1$）

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \mathbf{x}_i^{(c)}$$

- $\boldsymbol{\mu}$：全体数据的均值向量（$d \times 1$）

$$\boldsymbol{\mu} = \frac{1}{N} \sum_{c=1}^{C} \sum_{i=1}^{N_c} \mathbf{x}_i^{(c)} = \frac{1}{N} \sum_{c=1}^{C} N_c \boldsymbol{\mu}_c$$

- $\mathbf{w}$：我们想要寻找的投影方向（$d \times 1$ 列向量）。投影后的新数据为 $y = \mathbf{w}^T \mathbf{x}$。

我们的目标是找到一个最优的 $\mathbf{w}$。

2. 散度矩阵 (Scatter Matrices)

为了衡量“类内聚集”和“类间分开”的程度，我们定义两个关键的散度矩阵。

a) 类内散度矩阵 (Within-Class Scatter Matrix) S_W

S_W 衡量了每个类别内部样本分布的分散程度。它是所有类别的协方差矩阵之和。

$$S_W = \sum_{c=1}^{C} S_c = \sum_{c=1}^{C} \sum_{i=1}^{N_c} (\mathbf{x}_i^{(c)} - \boldsymbol{\mu}_c)(\mathbf{x}_i^{(c)} - \boldsymbol{\mu}_c)^T$$

b) 类间散度矩阵 (Between-Class Scatter Matrix) S_B

S_B 衡量了不同类别的均值向量之间的分散程度。它是由每个类别的均值与总体均值的差向量外积后加权（乘以类别样本数）求和得到。

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

c) 总散度矩阵 (Total Scatter Matrix) S_T

可以证明，总散度矩阵（所有样本与全局均值的散度）等于类内散度与类间散度之和：
$S_T = S_W + S_B$ （南瓜书上有证明过程）。

$$S_T = \sum_{c=1}^{C} \sum_{i=1}^{N_c} (\mathbf{x}_i^{(c)} - \boldsymbol{\mu})(\mathbf{x}_i^{(c)} - \boldsymbol{\mu})^T$$

3. 目标函数：广义瑞利商 (Generalized Rayleigh Quotient)

现在，我们将“类间方差大，类内方差小”的思想转化为数学目标函数。

一个样本 $\mathbf{x}$ 投影后的值为 $y = \mathbf{w}^T \mathbf{x}$。

- 投影后，全体的均值标量为 $\mathbf{w}^T \boldsymbol{\mu}$
- 投影后，第 c 类的均值标量为 $\mathbf{w}^T \boldsymbol{\mu}_c$

那么，投影后的数据可以定义：

- **类间方差 (Between-Class Variance)**:

$$\sigma_B^2 = \sum_{c=1}^{C} N_c (\mathbf{w}^T \boldsymbol{\mu}_c - \mathbf{w}^T \boldsymbol{\mu})^2 = \mathbf{w}^T \left[\sum_{c=1}^{C} N_c (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T \right] \mathbf{w} = \mathbf{w}^T S_B \mathbf{w}$$

- **类内方差 (Within-Class Variance)**:

$$\sigma_W^2 = \sum_{c=1}^{C} \sum_{i=1}^{N_c} (\mathbf{w}^T \mathbf{x}_i^{(c)} - \mathbf{w}^T \boldsymbol{\mu}_c)^2 = \mathbf{w}^T \left[\sum_{c=1}^{C} \sum_{i=1}^{N_c} (\mathbf{x}_i^{(c)} - \boldsymbol{\mu}_c)(\mathbf{x}_i^{(c)} - \boldsymbol{\mu}_c)^T \right] \mathbf{w} = \mathbf{w}^T S_W \mathbf{w}$$

我们的目标是最大化它们的比值，从而得到目标函数 $J(\mathbf{w})$：

$$J(\mathbf{w}) = \frac{\sigma_B^2}{\sigma_W^2} = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$$

这个函数被称为***广义瑞利商 (Generalized Rayleigh Quotient)***。我们的任务就是找到使 $J(\mathbf{w})$ 最大的 $\mathbf{w}$。

4. 求解最优投影向量 $\mathbf{w}$

直接对 $J(\mathbf{w})$ 求导并令其为零比较困难。我们使用一个经典的技巧：固定分母，最大化分子。即，我们规定类内散度在投影后为一个常数，比如 $\mathbf{w}^T S_W \mathbf{w} = 1$。在这个约束条件下，去最大化分子 $\mathbf{w}^T S_B \mathbf{w}$。

这变成了一个带约束的优化问题：

$$\begin{aligned} \underset{\mathbf{w}}{\text{maximize}} \quad & \mathbf{w}^T S_B \mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^T S_W \mathbf{w} = 1 \end{aligned}$$

我们使用 ***拉格朗日乘数法 (Lagrange Multiplier)*** 来求解。构造拉格朗日函数 L：

$$L(\mathbf{w}, \lambda) = \mathbf{w}^T S_B \mathbf{w} - \lambda(\mathbf{w}^T S_W \mathbf{w} - 1)$$

其中 λ 是拉格朗日乘子。现在，我们对 $\mathbf{w}$ 求导并令其为零：

$$\frac{\partial L(\mathbf{w}, \lambda)}{\partial \mathbf{w}} = 2S_B \mathbf{w} - 2\lambda S_W \mathbf{w} = 0$$

整理得到：

$$S_B \mathbf{w} = \lambda S_W \mathbf{w} \quad (1)$$

这个形式非常重要！它是一个***广义特征值问题 (Generalized Eigenvalue Problem)***。

5. 进一步简化与求解

我们注意到 S_B 矩阵的秩最多为 $C-1$（因为它由 C 个向量外积求和得到，但这些向量线性相关，$\sum N_c(\boldsymbol{\mu}_c - \boldsymbol{\mu}) = 0$）。对于二分类问题（$C=2$），$S_B$ 的秩为 1。

观察 S_B 的定义：$S_B = \sum_{c=1}^{C} N_c(\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T$。对于二分类问题，我们可以将其写为：

$$S_B = N_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu})(\boldsymbol{\mu}_1 - \boldsymbol{\mu})^T + N_2(\boldsymbol{\mu}_2 - \boldsymbol{\mu})(\boldsymbol{\mu}_2 - \boldsymbol{\mu})^T$$

并且注意到 $\boldsymbol{\mu} = \frac{N_1\boldsymbol{\mu}_1 + N_2\boldsymbol{\mu}_2}{N_1 + N_2}$，经过推导可以发现 $S_B\mathbf{w}$ 的方向始终与 $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ 相同。即：

$$S_B \mathbf{w} \propto (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

将这个关系代入之前的广义特征方程 (1)：

$$S_B \mathbf{w} = \lambda S_W \mathbf{w} \propto (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

我们可以立即得到***对于二分类问题，LDA的最优投影方向为***：

$$\mathbf{w} \propto S_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

结论： 最优投影方向 $\mathbf{w}$ 是矩阵 $S_W^{-1}S_B$ 的特征向量。对于二分类，它直接由 $S_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ 给出，我们无需显式地求解特征值。

6. 多分类推广

对于类别数 $C > 2$ 的情况，我们希望将数据投影到一个 $C-1$ 维的子空间（因为类间散度 S_B 的秩最大为 $C-1$）。

广义特征方程 $S_B\mathbf{w} = \lambda S_W\mathbf{w}$ 会有 $C-1$ 个非零广义特征值 $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{C-1} > 0$ 及其对应的广义特征向量 $\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{C-1}$。

这些特征向量构成了我们想要的投影矩阵 W：

$$W = \begin{bmatrix} | & | & & | \\ \mathbf{w}_1 & \mathbf{w}_2 & \cdots & \mathbf{w}_{C-1} \\ | & | & & | \end{bmatrix}_{d \times (C-1)}$$

数据 $\mathbf{x}$ 投影到这个新的子空间后变为：

$$\mathbf{y} = W^T\mathbf{x}$$

总结推导流程

1. **计算各类均值** $\boldsymbol{\mu}_c$ 和全局均值 **$\boldsymbol{\mu}$**。
2. **计算散度矩阵** S_W 和 S_B。
3. **求解广义特征值问题** $S_B\mathbf{w} = \lambda S_W\mathbf{w}$。
   - 对于二分类，解为 $\mathbf{w}^* = S_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$。
   - 对于多分类，解为 $S_W^{-1}S_B$ 的前 $C-1$ 个最大特征值对应的特征向量。
4. **组成投影矩阵** W 并对数据进行降维：$\mathbf{y} = W^T\mathbf{x}$。

LDA不仅是一个优秀的降维工具，投影后的空间也通常能带来更好的分类性能。希望这次的推导能帮助大家真正理解LDA背后的数学原理。下次课我们将讨论它的实际应用以及与相关方法的比较。

有任何问题，欢迎在 office hour 来找我讨论。