第三章：线性回归重点

一、线性判别分析 LDA

线性判别分析 (Linear DiscriminantAnalysis, LDA) 的基本思想：给定训练集，设法将样本投影到一条适当选择的直线上，使得同类样本的投影点尽可能接近、异类样例的投影点中心尽可能远离。目标："**投影后类内方差小，类间距离大**"，LDA 的目标是在**保留尽可能多的类区分信息的同时进行降维.**

对 $\boldsymbol{x}$ 中的各个成分作线性组合, 得到 $y=\boldsymbol{w}^T\boldsymbol{x}$, 这样 n 个样本 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_m$ 就产生了 n 个投影结果 $y_1,\cdots,y_n$, 相应的属于集合 Y_0 和 Y_1, 即 $Y_i=\boldsymbol{w}^TX_i\ (i=0,1)$. 如果 $\boldsymbol{\mu}_i$ 为 d 维样本均值为

$$\boldsymbol{\mu}_i=\frac{1}{n_i}\sum_{\boldsymbol{x}\in X_i}\boldsymbol{x},\quad i=0,1,$$

则投影后的点的样本均值为

$$\tilde{\mu}_i=\frac{1}{n_i}\sum_{y\in Y_i}y=\frac{1}{n_i}\sum_{\boldsymbol{x}\in X_i}\boldsymbol{w}^T\boldsymbol{x}=\boldsymbol{w}^T\boldsymbol{\mu}_i$$

也就恰好是原样本均值 $\boldsymbol{\mu}_i$ 的投影.

投影后的点的样本均值之差为

$$|\tilde{\mu}_0-\tilde{\mu}_1|=|\boldsymbol{w}^T(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)|.$$

定义投影后第 i 类的类别 ω_i 的类内散度为

$$\tilde{s}_i^2=\sum_{y\in Y_i}(y-\tilde{\mu}_i)^2,$$

则 $\frac{1}{n}(\tilde{s}_0^2+\tilde{s}_1^2)$ 就是全部数据的总体的方差的估计. $\tilde{s}_0^2+\tilde{s}_1^2$ 称为投影样本的总体类内散度. Fisher 线性可分性准则要求在投影 $y=\boldsymbol{w}^T\boldsymbol{x}$ 下, 准则函数

$$J(\boldsymbol{w})=\frac{|\tilde{\mu}_0-\tilde{\mu}_1|^2}{\tilde{s}_0^2+\tilde{s}_1^2}$$

最大化.

定义**类间散度矩阵** $\boldsymbol{S}_b=(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^T$, 可证明其为对称半正定的. 投影后两类样本均值之差展开为

$$|\tilde{\mu}_0-\tilde{\mu}_1|^2=\boldsymbol{w}^T(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^T\boldsymbol{w}=\boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}.$$

定义原样本空间 X_1 和 X_2 中的**类内散度矩阵**

$$\boldsymbol{\Sigma}_0=\sum_{x\in X_0}(\boldsymbol{x}-\boldsymbol{\mu}_0)(\boldsymbol{x}-\boldsymbol{\mu}_0)^T,$$
$$\boldsymbol{\Sigma}_1=\sum_{x\in X_1}(\boldsymbol{x}-\boldsymbol{\mu}_1)(\boldsymbol{x}-\boldsymbol{\mu}_1)^T.$$

则总类内散度矩阵 $\boldsymbol{S}_w=\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1$, 可证明其是对称半正定的.

投影后第 i 内的类内散度为：

$$\begin{aligned}\tilde{s}_i^2 &= \sum_{y \in Y_i} (y - \tilde{\mu}_i)^2 \\ &= \sum_{y \in Y_i} (\boldsymbol{w}^T \boldsymbol{x} - \boldsymbol{w}^T \boldsymbol{\mu}_i)^2 \\ &= \sum_{\boldsymbol{x} \in X_i} \boldsymbol{w}^T (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{w} = \boldsymbol{w}^T \boldsymbol{\Sigma}_i \boldsymbol{w}.\end{aligned}$$

故散度矩阵总和可写为 $\tilde{s}_0^2 + \tilde{s}_1^2 = \boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}$. 所以

$$J(\boldsymbol{w}) = \frac{|\tilde{\mu}_0 - \tilde{\mu}_1|^2}{\tilde{s}_0^2 + \tilde{s}_1^2} = \frac{\boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}}.$$

求解结果公式（背会！！！）

m维空间到一维空间投影轴的最佳方向

$$\boldsymbol{w}^* = \mathbf{S}_w^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$$ （因$\boldsymbol{w}$与大小无关，只与方向有关）

J(w) 最大值

$$(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}} \mathbf{S}_w^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$$

最佳投影变换为

$$y = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\mathrm{T}} \mathbf{S}_w^{-1} \boldsymbol{x}$$

二、矩阵向量求导

1 向量对向量的偏导计算

- $y = Wx$ **中** $\frac{\partial \vec{y}}{\partial \vec{x}}$

其中$y \in \mathbb{R}^{C \times 1}, W \in \mathbb{R}^{C \times D}, x \in \mathbb{R}^{D \times 1}$

由于每个 y 中的元素对每个 x 中的元素都要求偏导，因此结果肯定是个二维雅可比矩阵：

$$\begin{bmatrix} \frac{\partial \vec{y}_1}{\partial \vec{x}_1} & \frac{\partial \vec{y}_1}{\partial \vec{x}_2} & \frac{\partial \vec{y}_1}{\partial \vec{x}_3} & \cdots & \frac{\partial \vec{y}_1}{\partial \vec{x}_D} \\ \frac{\partial \vec{y}_2}{\partial \vec{x}_1} & \frac{\partial \vec{y}_2}{\partial \vec{x}_2} & \frac{\partial \vec{y}_2}{\partial \vec{x}_3} & \cdots & \frac{\partial \vec{y}_2}{\partial \vec{x}_D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \vec{y}_C}{\partial \vec{x}_1} & \frac{\partial \vec{y}_C}{\partial \vec{x}_2} & \frac{\partial \vec{y}_C}{\partial \vec{x}_3} & \cdots & \frac{\partial \vec{y}_C}{\partial \vec{x}_D} \end{bmatrix}$$

对！由上面的规律（你可以再求个 $\vec{y}_1$ 对 $\vec{x}_3$ 再试试），容易得到：

$$\frac{\partial \vec{y}_i}{\partial \vec{x}_j} = W_{i,j}$$

因此：

$$\frac{d\vec{y}}{d\vec{x}} = W$$

2 向量对矩阵的偏导计算

- $y = Wx$ **中** $\dfrac{\partial \vec{y}}{\partial \vec{W}}$

首先要明确，一维的向量对二维的矩阵的偏导其结果必然是一个三维的矩阵 $\mathbb{R}^{C\times C\times D}$,你想一下每一个 $\vec{y}_i$ 都要对 $W_{j,k}$ 求偏导，那么将会得到 $C \times C \times D$ 个偏导值，我们先不讨论怎么排列，只关注每个位置怎么求。

我们求 $\vec{y}_3$ ，容易得到：

$$\vec{y}_3 = \vec{x}_1 W_{1,3} + \vec{x}_2 W_{2,3} + \cdots + \vec{x}_D W_{D,3}$$

注意下 x 和 W 的下标，是不是和上面一样有规律可循？

对了！能够发现式子 $\vec{y}_3$ 的等式中，x 的列维度和 W 的行维度是一样的（符合矩阵的运算规律）， W 的列维度等于3,容易推出：

$$\frac{\partial \vec{y}_j}{\partial W_{i,j}} = \vec{x}_i$$

为了更好地表示，我们使用 F 表示 y 对 W 的三维偏导，其中：

$$F_{i,j,k} = \frac{\partial \vec{y}_i}{\partial W_{j,k}}$$

注意到只有 $i == k$ 时，偏导等于 $\vec{x}_j$ ，其他都为0，因此：

$$F_{i,j,i} = \partial \vec{x}_i$$

因此， F 中实际的有效信息只有2维!这个可以手动验证。

3 矩阵对矩阵的偏导计算

- $Y = XW$ 中 $\frac{\partial Y}{\partial X}$

其中$Y \in \mathbb{R}^{N\times C}, W \in \mathbb{R}^{D\times C}, \boldsymbol{x} \in \mathbb{R}^{N\times D}$

现在3个元素都是矩阵，同样还是利用矩阵的计算方法，容易得到：

$$Y_{i,j} = \Sigma_{k=1}^{D} X_{i,k} W_{k,j}$$

容易看出对 $\frac{\partial Y_{a,b}}{\partial X_{c,d}}$ 来说只有a==c时，其偏导数的值不为0,即：

$$\frac{\partial Y_{i,j}}{\partial X_{i,k}} = W_{k,j}$$

如果我们只考虑Y的第i行和X的第i行，那么可以得到：

$$\frac{\partial Y_{i,:}}{\partial X_{i,:}} = W^T$$

是不是就是公式(2)！

三、对数几率

- **几率(odds)定义**：几率是指该事件发生的概率与该事件不发生的概率的比值。即如果事件发生概率是 p，那么该事件的几率为 $\frac{p}{1-p}$ 。(可以想象，当几率大于1时，说明该事件发生的概率大，几率小于 1 时，说明该事件发生的概率小；几率变化范围为 (0, +∞))
- 从这个几率的概念推广一个概念叫**对数几率**(log odds)或logit函数：$\log\frac{p}{1-p}$ (可以想象，当对数几率大于0时，说明该事件发生的概率大；对数几率小于0时，说明该事件发生的概率小)

- 对数几率（log odds）
  - 样本作为正例的相对可能性的对数

将 y 视为样本 $\boldsymbol{x}$ 作为正例的可能性
(y表示为 x 被分到正例的概率)

称比值 $\frac{y}{1-y}$ 为几率

$$\ln\frac{y}{1-y} = \boldsymbol{w}^T\boldsymbol{x} + b$$

1 - y 是看作样本为反例的概率

对数几率回归优点：无需事先假设数据分布；可得到“类别”的近似概率预测；可直接应用现有数值优化算法求取最优解