线性判别分析 LDA

Li Liang*

1 线性判别分析

1.1 二类 LDA 原理

给定数据集 $D=\{(\mathbf{x}_i,y_i)\}_{i=1}^m$，$y_i\in\{0,1\}$，令 $\mathbf{X_i},\mu_\mathbf{i},\sum_\mathbf{i}$ 表示第 $i\in\{0,1\}$ 类示例的集合、均值向量和协方差矩阵。若将数据投影到直线 ω 上，则两类样本的中心在直线上的投影分别为 $\omega^T\mu_0$ 和 $\omega^T\mu_1$，两类样本在直线上投影的协方差为 $\omega^T\sum_\mathbf{0}\omega$ 和 $\omega^T\sum_\mathbf{1}\omega$（协方差传播）。

若是同类尽可能近，则协方差尽可能小；异类尽可能远则要求类中心距离大。同时考虑两者，则最大化目标为：

$$\begin{aligned} J &= \frac{\|\omega^T\mu_0-\omega^T\mu_1\|_2^2}{\omega^T\sum_\mathbf{0}\omega+\omega^T\sum_\mathbf{1}\omega} \\ &= \frac{\omega^T(\mu_0-\mu_1)(\mu_0-\mu_1)^T\omega}{\omega^T(\sum_0+\sum_1)\omega} \end{aligned} \tag{1}$$

定义类内散度矩阵和类间散度矩阵为：

$$\begin{aligned} S_w &= \sum_0+\sum_1 \\ &= \sum_{x\in\mathbf{X_0}}(x-\mu_0)(x-\mu_0)^T+\sum_{x\in\mathbf{X_1}}(x-\mu_1)(x-\mu_1)^T \end{aligned} \tag{2}$$

$$S_b = (\mu_0-\mu_1)(\mu_0-\mu_1)^T \tag{3}$$

带入 (1) 式：

$$J=\frac{\omega^TS_b\omega}{\omega^TS_w\omega} \tag{4}$$

*https://github.com/leeliang/

上式就是广义瑞利商，广义瑞利商有个性质为：J 的最大值为矩阵 $S_w^{-1}S_b$ 的最大特征值，而对应的 ω 为 $S_w^{-1}S_b$ 的最大特征值对应的特征向量（特征值最大，意味着在对应的特征向量上的变化最大，分类越大），即：

$$S_w^{-1}S_b\omega = \lambda\omega \tag{5}$$

上式就是 Fisher Linear Discriminantion 公式，推导过程可参见西瓜书。$S_w^{-1}S_b$ 不一定是对称矩阵，在求它的特征向量时不能使用奇异值分解，这样就只能使用普通的求特征向量的方式。注意到 $S_b\omega = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T\omega$，其中 $c = (\mu_0 - \mu_1)^T\omega$ 是标量，是两个类别中心在直线上的距离。(5) 式可以重写为：

$$\omega = \frac{c}{\lambda}S_w^{-1}(\mu_0 - \mu_1) \tag{6}$$

注意到 (4) 式分子和分母都是关于 ω 的二次型，因此 (4) 式的解与 ω 的大小无关，只与方向有关。则上式等价于:

$$\omega = S_w^{-1}(\mu_0 - \mu_1) \tag{7}$$

在实践中，通常对 S_w 进行奇异值分解，求取 ω。

~~LDA 可从贝叶斯决策理论的角度来阐释，并可证明，当两类数据同先验、满足高斯分布且协方差相等时，LDA 可达到最优分类。~~

1.2 多类 LDA 原理

给定数据集 $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$，$y_i \in \{0, 1, ..., k\}$，类似地，目标函数为：

$$J = \frac{W^T S_b W}{W^T S_w W} \tag{8}$$

其中，W 为一个超平面，

$$\begin{aligned} S_b &= \sum_{1\le i,j\le k}^{k} (\mu_i - \mu_j)(\mu_i - \mu_j)^T \\ S_w &= \sum_{j=1}^{k} S_{wj} = \sum_{j=1}^{k}\sum_{x\in X_j}(x - \mu_j)(x - \mu_j)^T \end{aligned} \tag{9}$$

按照上式计算 S_b 较复杂，LDA 采用一个间接的方式。~~等价?~~

首先定义 X 的整体散度矩阵:

$$S_t = S_b + S_w = \sum_{j=1}^{k}(x_j - \mu)(x_j - \mu)^T \tag{10}$$

μ 是所有示例的均值向量。则:

$$S_b = S_t - S_w = \sum_{j=1}^{k} N_j(\mu_j - \mu)(\mu_j - \mu)^T \tag{11}$$

N_j 为第 j 个类别的示例个数。

由于 $W^T S_b W$ 和 $W^T S_w W$ 都是矩阵，无法使用标量的最优化方法，一般来说，常使用的优化目标为:

$$J = \frac{tr(W^T S_b W)}{tr(W^T S_w W)} \tag{12}$$

1.3 LDA 分类

对二分类问题。由于只有两个类别，在经过上面的求解后，最后所有样本将会映射到一维空间中，一般将两个类别的中心点之间中心点作为分类点。

对于多类的情况主要考虑用来数据降维，那么降维后的维度是多少呢? W 是一个利用了样本的类别得到的投影矩阵，是 $S_w^{-1}S_b$ 的特征向量，该特征向量的维数不超过 $S_w^{-1}S_b$ 的秩。

$$\begin{aligned} R(S_w^{-1}S_b) \le \min(R(S_w^{-1}), R(S_b)) \le R(S_b) \\ &= R(\sum_{j=1}^{k} N_j(\mu_j - \mu)(\mu_j - \mu)^T) \\ &\le \sum_{j=1}^{k} N_j R[(\mu_j - \mu)(\mu_j - \mu)^T] \\ &= \sum_{j=1}^{k} R(\mu_j - \mu) \end{aligned} \tag{13}$$

由于存在线性组合 $\sum_{j=1}^{k} N_j(\mu_j - \mu) = 0$，故上式的秩最大为 $k-1$，即特征向量最多有 $k-1$ 个。