线性判别分析

cloud

2017.1.10

1 概述

线性判别分析（LDA）是一种经典的线性学习方法，亦称 Fisher 判别分析。LDA 的思想非常朴素，即给定训练集，设法将样本投影到一条直线上，使得同类样本的投影点尽可能接近，异类样本的投影点尽可能的远。在对新样本进行分类时，将其投影到这条直线上，再根据投影点的位置来确定新样本的类别，示意图如下：

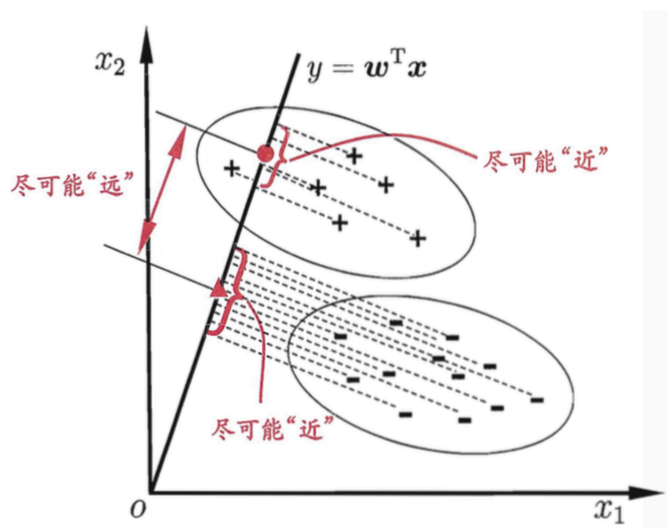

Figure 1: 二分类示意图

上图中 $+, -$ 分别表示正反例，红色实心圆盒三角形分别表示两类样本投影后的中心点。

2 二分类情况

2.1 目标函数

给定数据集 $D=(x^{(i)},y^{(i)}),y^{(i)}\in\{0,1\}$，令 X_i,μ_i,Σ_i 分别表示第 $i\in\{0,1\}$ 类样本的集合，均值向量和协方差矩阵。若将数据投影到直线 $y=\omega^Tx$ 上，则两类样本的中心在直线上的投影分别为 $\omega^T\mu_0$ 和 $\omega^T\mu_1$，两类样本的协方差分别为 $\omega^T\Sigma_0\omega$ 和 $\omega^T\Sigma_1\omega$。由于直线是一维空间，因此投影和协方差都是实数。

前面说到要使同类样本尽可能接近，可以让同类样本点的协方差尽可能小，即 $\omega^T\Sigma_0\omega+\omega^T\Sigma_1\omega$ 尽可能小。而欲要使异类样本的投影点尽可能远离，可以让类中心之间距离尽可能大，即 $(\omega^T\mu_0-\omega^T\mu_1)^2$ 尽可能大，那么可以得到最大化目标：

$$
\begin{aligned}
J&=\frac{(\omega^T\mu_0-\omega^T\mu_1)^2}{\omega^T\Sigma_0\omega+\omega^T\Sigma_1\omega}\\
&=\frac{\omega^T(\mu_0-\mu_1)(\mu_0-\mu_1)^T\omega}{\omega^T(\Sigma_0+\Sigma_1)\omega}
\end{aligned}
$$

定义类内散度矩阵：

$$
\begin{aligned}
S_\omega&=\Sigma_0+\Sigma_1\\
&=\sum_{x\in X_0}(x-\mu_0)(x-\mu_0)^T+\sum_{x\in X_1}(x-\mu_1)(x-\mu_1)^T
\end{aligned}
$$

定义类间散度矩阵：

$$
S_b=(\mu_0-\mu_1)(\mu_0-\mu_1)^T
$$

那么最大化目标可重写为如下的形式，称为 S_b 和 S_ω 的广义瑞利商。

$$
J=\frac{\omega^TS_b\omega}{\omega^TS_\omega\omega}
$$

那么下面要求解 ω，注意到上式的分子和分母都是关于 ω 的二次项，因此上式得解与 ω 的长度无关，只与其方向有关。在此令 $\omega^TS_\omega\omega=1$，则上式等价于：

$$
\begin{aligned}
&\min_\omega -\omega^TS_b\omega \qquad (1)\\
&s.t.\quad \omega^TS_\omega\omega=1
\end{aligned}
$$

2.2 求解

利用拉格朗日乘子法，即 (1) 式等价于求解 $L(\omega,\lambda)=-\omega^TS_b\omega+\lambda(\omega^TS_\omega\omega-1)$，分别对 ω 和 λ 求导取 0 可得：

$$
S_b\omega=\lambda S_\omega\omega \qquad (2)
$$

又因为 $S_b\omega$ 的方向恒为 $\mu_0 - \mu_1$，不妨令：

$$S_b\omega = \lambda(\mu_0 - \mu_1) \tag{3}$$

将 (3) 式代入 (2) 式得

$$\omega = S_\omega^{-1}(\mu_0 - \mu_1)$$

考虑到数值的稳定性，在实践通常是对 S_ω 进行奇异值分解，即 $S_\omega = U\Sigma V^T$。这里 Σ 是一个实对角矩阵，其对角线上的元素是 S_ω 的奇异值。

3 多分类情况

假设存在 M 个类，样本总数是 N 且第 i 类样本数为 N_i，可以将上节的式子推广到多分类的情况，图示如下：

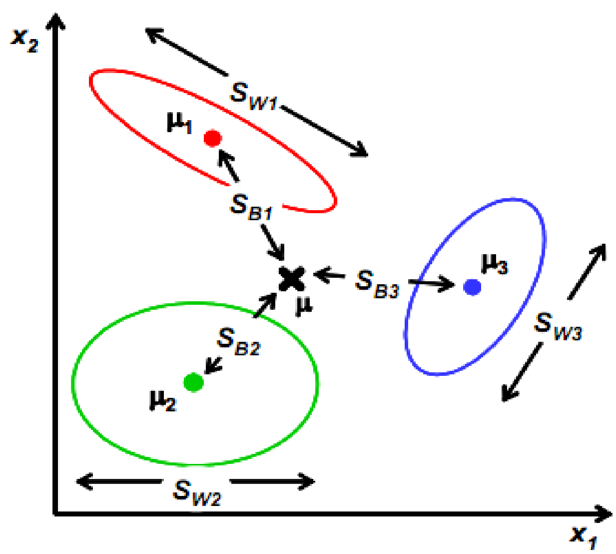

Figure 2: 多分类示意图

对于类内散度矩阵二分类和多分类情况写法一致，所以可以写成：

$$S_\omega = \sum_{i=1}^{M} S_{\omega_i}$$

其中 S_{ω_i} 如下，μ_i 代表每个类样本的均值。

$$S_{\omega_i} = \sum_{x \in X_i} (x - \mu_i)(x - \mu_i)^T$$

对于类间散度矩阵，原来的是计算两个均值点的散列情况，现在度量的是每类均值点相对于样本中心的散列情况。类间散度矩阵如下，其中 μ 是所有样本的均值向量。

$$S_b = \sum_{i=1}^{M} N_i(\mu_i - \mu)(\mu_i - \mu)^T$$

上式中可以认为每类样本的权重是 $\frac{N_i}{N}$，由于 J 对样本总数不敏感，所以可以直接写成上式的形式。

上节是针对只有二分类的情况，如果在多分类情况下，要怎么改变才能保证投影后类别能够分离呢？二分类下将 d 维特征的样本投影到一条一维的直线上，那么在 M 个类别下，优化目标可以写成：

$$\max_{W} \frac{tr(W^T S_b W)}{tr(W^T S_\omega W)}$$

其中 $W \in R^{d\times(M-1)}$，$tr(\cdot)$ 表示矩阵的迹，上式可以通过如下广义特征值求解：

$$S_b W = \lambda S_\omega W$$

W 的闭式解则是 $S_\omega^{-1} S_b$ 的 $M-1$ 个最大广义特征值所对应的特征向量的矩阵。若将 W 视为一个投影矩阵，则多分类 LDA 将样本投影到 $M-1$ 维空间，$M-1$ 通常远小于样本特征数，这样就通过了投影减少样本点的维数，因此 LDA 常被视为一种经典的降维技术。

4 参考文献

1. 周志华的机器学习

2.http://www.cnblogs.com/jerrylead/archive/2011/04/21/2024384.html