

线性判别分析

机器学习笔记 create by siwanghu v1.0

LDA (线性判别分析), 又称 Fisher 线性判别。将高维的样本投影到最佳鉴别空间, 以达到抽取分类信息和压缩特征空间维数的效果。用于分类或数据降低维数, LDA 与 PCA 相比最大的特点是 LDA 是有监督的数据降维, PCA 是无监督的数据降维。

在 R^n 空间中样本 $X = \{x_1, x_2, x_3, \dots, x_m\}$, 也就是说样本总共有 m 个, 每个样本是 n 维的, 代表每个样本有 n 个特征。样本总共分为 c 个类别, 每个类别的数目大小用 N_i , $0 \leq i \leq c$, 也就是说 $N_1 + N_2 + \dots + N_c = m$ 。

LDA 要做这样一件事情, 它将样本投影到空间 R^l 中, $l \leq n$, 代表将样本数据特征降维。在投影的空间中, 样本类别很好分离。LDA 算法的思想是, 投影到另一个空间中时让不同类别的样本尽量离的远, 同一个类别的样本尽量离的尽。

为了方便叙述算法, 定义如下符号:

S_b 类间离散度矩阵

S_w 类内离散度矩阵

N_i 属于第 i 类样本的个数

N 样本的总数

x_i 代表第 i 个样本, 是一个多维的列向量, 存储了样本的每个特征

u 所有样本的均值 $u = \frac{\sum_{i=1}^m x_i}{m}$

u_i 代表第 i 类样本的均值 $u_i = \frac{\sum_{i \in \text{第 } i \text{ 类}} x_i}{n_i}$

LDA 算法的关键步骤是计算类间离散度矩阵和类内离散度矩阵

$$S_b = \sum_{i=1}^c \frac{N_i}{N} (u_i - u)(u_i - u)^T$$

$$S_w = \sum_{i=1}^c \sum_{x_k \in \text{第} i \text{ 类}} \frac{1}{N} (u_i - x_k)(u_i - x_k)^T$$

如果只有两类 $S_b = (u_2 - u_1)(u_2 - u_1)^T$

然后求出矩阵 $S_w^{-1}S_b$ 的特征值，取前最大 l 个特征值对应的特征向量 (w_1, w_2, \dots, w_l) 组成矩阵 $W_{n,l}$

然后取出样本 X 中的每个样本，经过投影矩阵 W ，求出在投影特征空间中的对应点即可，计算方式为：

$$Y_{l,1} = W_{n,l}^T * X_{n,1}$$

LDA 与 PCA 的区别

相同点：

1. 两者均可以对数据进行降维
2. 两者在降维时均使用了矩阵特征分解的思想
3. 两者都假设数据符合高斯分布

不同点：

1. LDA 是有监督的降维方法，而 PCA 是无监督的降维方法
2. LDA 降维最多降到类别数 $k-1$ 的维数，而 PCA 没有这个限制
3. LDA 除了可以用于降维，还可以用于分类
4. LDA 选择分类性能最好的投影方向，而 PCA 选择样本点投影具有最大方差的方向