线性判别分析（LDA）原理总结

原创 石头 机器学习算法那些事 2019-03-25

前言

线性判别分析（Linear Discriminant Analysis，以下简称LDA）是有监督的降维方法，在模式识别和机器学习领域中常用来降维。PCA是基于最大投影方差或最小投影距离的降维方法，LDA是基于最佳分类方案的降维方法，本文对其原理进行了详细总结。

目录

1. PCA与LDA降维原理对比

1.1 PCA降维原理

PCA是非监督式的降维方法，在降维过程中没有考虑类别的影响，PCA是基于最大投影方差或最小投影距离的降维方法，通俗点说，PCA降维后的样本集最大程度的保留了初始样本信息，常用投影距离来描述投影前后样本的差异信息。

用数学公式来阐述这一思想：

$$d = \| X - \overline{X} \|^2 \qquad (1)$$

其中原始样本集（n个m维数据）：

$$X_{\mathrm{m\times n}} = [x_1, x_2, \ldots, x_n]$$

降维后的样本集（n个k维数据）：

$$\overline{X}_{k\times n} = [\overline{x_1}, \overline{x_2}, \ldots, \overline{x_n}]$$

假设投影变换后的新坐标系（标准正交基）：

$$W_{d\times k} = [w_1, w_2, \ldots, w_k]$$

投影前后的样本关系：

$$\overline{X}_{k\times n} = W^T_{k\times d} X_{\mathrm{d\times n}} \qquad (2)$$

最小化（1）式，并根据条件（2），可求得最佳的投影坐标系W。给定新的输入样本，利用（2）式可求的对应的降维样本。

1.2 LDA降维原理

LDA是有监督的降维方法，在降维过程中考虑了类别的影响，LDA是基于最佳分类效果的降维方法。因此，降维后不同类的样本集具有最大的分类间隔 。

如何描述最大分类间隔，当不同类样本的投影点尽可能远离且相同类样本的投影点尽可能接近，则样本集具有最大分类间隔。我们用类中心间的距离和类的协方差分别表示不同类的距离和相同类的接近程度。

本节只考虑二分类的LDA降维，不同类样本间的投影距离：

$$\| w^T u_1 - w^T u_2 \|^2 \qquad (3)$$

不同类的投影协方差之和：

$$w^T(\Sigma_1 + \Sigma_2)w \qquad (4)$$

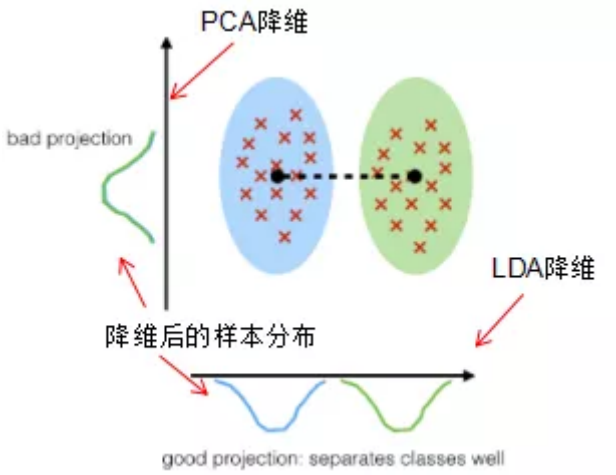
结合(3)(4)式，得到优化目标函数：

$$J(w) = \frac{\| w^T u_1 - w^T u_2 \|^2}{w^T(\Sigma_1 + \Sigma_2)w} \qquad (5)$$

最大化（5）式，得到投影向量w，其中 u_1 和 u_2 分别是两个类样本的中心点，Σ_1 和 Σ_2 分别是两个类的协方差。

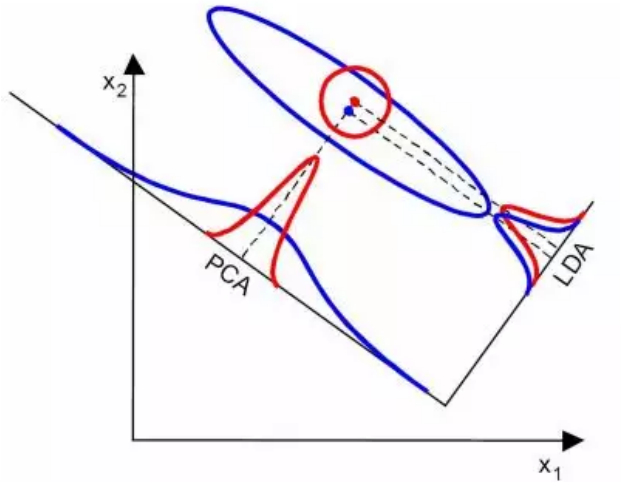
1.3 PCA与LDA降维应用场景对比

若训练样本集两类的均值有明显的差异，LDA降维的效果较优，如下图：



由上图可知，LDA降维后的二分类样本集具有明显差异的样本分布。

若训练样本集两类的均值无明显的差异，但协方差差异很大，PCA降维的效果较优，如下图：



由上图可知，PCA降维后的二分类样本分布较LDA有明显的差异。

2. 二类LDA算法推导

假设二类数据集$D_{m\times n}=\{(x_1,y_1),(x_1,y_1),...,(x_n,y_n)\}$，其中xi为m维列向量，我们定义两类为C1和C2，即$y_i\in\{C1,C2\}$，对应的样本集个数分别为N_1和N_2。

根据上一节的LDA的优化目标函数推导投影向量，即最大化目标函数：

$$J(w)=\frac{\|w^Tu_1-w^Tu_2\|^2}{w^T(\Sigma_1+\Sigma_2)w} \qquad (5)$$

其中u_1和u_2为二类的均值向量：

$$u_1=\frac{1}{N_1}\sum_{i\in C1}x_i$$

$$u_2=\frac{1}{N_2}\sum_{i\in C2}x_i$$

Σ_1和Σ_2为二类的协方差矩阵：

$$\Sigma_1=\frac{1}{N_1}\sum_{i\in C1}(x_i-u_1)(x_i-u_1)^T$$

$$\Sigma_2=\frac{1}{N_2}\sum_{i\in C2}(x_i-u_2)(x_i-u_2)^T$$

目标函数转化为：

$$J(w)=\frac{w^T(u_1-u_2)((u_1-u_2)^Tw}{w^T(\Sigma_1+\Sigma_2)w} \qquad (6)$$

定义类内散度矩阵S_w和类间散度矩阵S_b：

$$S_w=\Sigma_1+\Sigma_2$$

$$S_b=(u_1-u_2)(u_1-u_2)^T$$

则(6)式等价于：

$$J(w)=\frac{w^T S_b w}{w^T S_w w} \qquad (7)$$

我们对（7）式的分母进行标准化，则（7）式等价于：

$$J(w)=w^T S_b w$$
$$s.t \quad w^T S_w w=1$$

引用拉格朗日乘子法，得：

$$J(w)=w^T S_b w-\lambda(w^T S_w w-1)$$

由$\frac{\partial J(w)}{\partial w}=0$，得：

$$\frac{\partial J(w)}{\partial Jw}=2S_b w-2\lambda S_w w=0$$
$$\Rightarrow S_b w=\lambda S_w w \qquad (8)$$
$$\because \quad S_b w=(u_1-u_2)(u_1-u_2)^T \mathrm{w}，且(u_1-u_2)^T \mathrm{w}为标量$$
$$令 \quad S_b w=\lambda(u_1-u_2)，代入(8)式得：$$
$$\mathrm{w}=S_w^{-1}(u_1-u_2)$$

因此，只要求出原始二类样本的均值和协方差就可以确定最佳的投影方向w了。

3. 多类LDA算法推导

假设k类数据集$D_{m\times n}=\{(x_1,y_1),(x_1,y_1),...,(x_n,y_n)\}$，其中xi为m维列向量，我们定义k类为$\{C_1,C_2,...,C_k\}$，对应的样本集个数分别为$\{N_1,N_2,...,N_k\}$。二类样本数据集通过投影向量w降到一维空间，多类样本数据集降到低维空间是一个超平面，假设投影到低维空间的维度为d，对应的基向量矩阵$W_{m\times d}=\{w_1,w_2,...,w_d\}$。

因此，多类LDA算法的优化目标函数为：

$$J(w)=\frac{W^T S_b W}{W^T S_w W} \qquad (8)$$

其中类内散度矩阵S_w和类间散度矩阵S_b：

$$S_w=\Sigma_1+\Sigma_2+...+\Sigma_k$$
$$=\sum_{j=1}^{k}\sum_{i\in Cj}(x_i-u_j)(x_i-u_j)^T$$

$$S_b = \sum_{j=1}^{k} N_j (u_j - u)(u_j - u)^T$$

u_j 为第j类样本的均值向量，u为所有样本的均值向量：

$$u = \frac{1}{n} \sum_{i=1}^{n} x_i$$

因为(8)式分子分母都是矩阵，常见的一种实现是取矩阵的迹，优化目标函数转化为：

$$J(w) = \frac{tr(W^T S_b W)}{tr(W^T S_w W)} \qquad (8)$$

优化过程如下：

$$J(w) = \frac{\sum_{i=1}^{d} w_i^T S_b w_i}{\sum_{i=1}^{d} w_i^T S_w w_i} = \sum_{i=1}^{d} \frac{w_i^T S_b w_i}{w_i^T S_w w_i} \qquad (9)$$

参考二类LDA算法，利用拉格朗日乘子法，得：

$$S_b w_i = \lambda S_w w_i$$

两边左乘 S_w^{-1} ：

$$(S_w^{-1} S_b) w_i = \lambda w_i$$

由上式可得LDA的最优投影空间是矩阵 $S_w^{-1} S_b$ 最大d个特征值对应的特征向量所组成的。

4. LDA算法流程

前两节推导了LDA算法，现在对LDA算法流程进行总结，理清一下思路。

假设k类数据集 $D_{m \times n} = \{(x_1, y_1), (x_1, y_1), ..., (x_n, y_n)\}$ ，其中xi为m维列向量，我们定义k类为 $\{C_1, C_2, ..., C_k\}$ ，降维后的维度是d。

1）计算每个类样本的均值向量 u_j 和所有数据集的均值向量 u

2）计算散度矩阵，包括类内散度矩阵 S_w 和类间散度矩阵 S_b

3）计算 $S_w^{-1} S_b$ 的特征向量 $(e_1, e_2, ..., e_m)$ 和对应的特征值 $(\lambda_1, \lambda_2, ..., \lambda_m)$

4）选择d个最大特征值对应的矩阵 $W_{m \times k}$ ，矩阵的每一列表示特征向量

5）对数据集D进行降维，得到对应的降维数据集 $Y_{k \times n}$ ，其中 $Y = X \times W$ 。

5. 正态性假设

LDA算法对数据集进行了如下假设：

1）数据集是服从正态分布的；

2）特征间是相互独立的；

3）每个类的协方差矩阵是相同的；

但是如果不满足了这三个假设，LDA算法也能用来分类和降维，因此LDA算法对数据集的分布有较好的鲁棒性。

6. LDA分类算法

前面我们重点分析了LDA算法在降维的应用，LDA算法也能用于分类 。LDA假设各类的样本数据集符合正态分布，LDA对各类的样本数据进行降维后，我们可以通过最大似然估计去计算各类别投影数据的均值和方差，如下式：

$$\bar{u}_j = \frac{1}{N_j} \sum_{i \in Cj} \overline{x_i}$$

$$\bar{\sigma}_j = \frac{1}{N_j - 1} \sum_{i \in Cj} (\overline{x_i} - \bar{u}_j)^2$$

进而得到各个类样本的概率密度函数：

$$f(x_i)_{i \in Cj} = \frac{1}{\sqrt{2\pi}\bar{\sigma}_j} \exp(-\frac{(\overline{x_i} - \bar{u}_j)^2}{2\bar{\sigma}_j^2})$$

其中$\overline{x_i}$为降维后的样本。

因此对一个未标记的输入样本进行LDA分类的步骤：

1） LDA对该输入样本进行降维；

2）根据概率密度函数，计算该降维样本属于每一个类的概率；

3）最大的概率对应的类别即为预测类别。

7. LDA小结

PCA是基于最大投影方差的降维方法，LDA是基于最优分类的降维方法，当两类的均值有明显差异时，LDA降维方法较优；当两类的协方差有明显差异时，PCA降维方法较优。在实际应用中也常结合LDA和PCA一起使用，先用PCA降维去消除噪声，再用LDA降维。

参考

http://sebastianraschka.com/Articles/2014_python_lda.html#introduction

https://www.cnblogs.com/pinard/p/6244265.html

推荐阅读

PCA算法原理总结

SVD算法原理总结

LLE算法原理总结