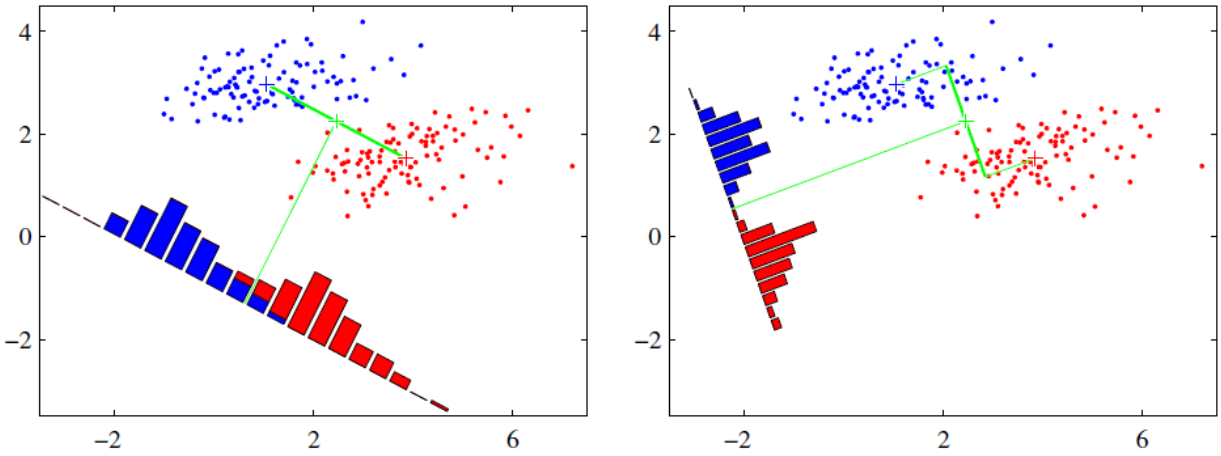
1.LDA简介

线性判别分析(Linear Discriminant Analysis, LDA)是一种监督学习的降维方法，也就是说数据集的每个样本是有类别输出。和之前介绍的机器学习降维之主成分分析(PCA)方法不同，PCA是不考虑样本类别输出的无监督学习方法。LDA的原理简单来说就是将带上标签的数据（点），通过投影的方法，投影到维度更低的空间中，使得投影后的点会形成按类别区分。而我们的目标就是使得投影后的数据，**类间方差最大，类内方差最小**。

以下图为例，假设有两类数据，分别为红色和蓝色。现在我们希望，将这些数据投影到一维的直线上，让每一种类别数据的投影点尽可能的接近，而红色和蓝色数据中心之间的距离尽可能的大。



从上图的两种投影方式能够看出，右图能够更好的满足我们的目标，即**类间方差最大，类内方差最小**。下面我们来看看LDA内部原理，如何达到我们所希望的目标。

2.瑞利商和广义瑞利商

介绍LDA原理之前，我们先了解一些数学知识，即**瑞利商(Rayleigh quotient)**与**广义瑞利商(genralized Rayleigh quotient)**。首先来看看瑞利商的函数R(A,x)

$$R(A,x)=\frac{x^HAx}{x^Hx}$$

其中x为非零向量，而A为n*n的Hermitan矩阵。Hermitan矩阵是指满足共轭转置矩阵和自己相等的矩阵，即$A^H=A$。如果矩阵A是实矩阵的话，如果满足$A^T=A$，那么就是Hermitan矩阵。

瑞利商R(A,x)有一个非常重要的性质，即它的最大值等于矩阵A的最大特征值，而最小值等于矩阵A的最小特征值，即满足

$$\lambda_{min}\leq\frac{x^HAx}{x^Hx}\leq\lambda_{max}$$

以上就是瑞利商的内容，现在看看广义瑞利商内容，广义瑞利商函数R(A,B,x)

$$R(A,B,x)=\frac{x^HAx}{x^HBx}$$

其中x为非零向量，而A,B为n*n的Hermitan矩阵，B是正定矩阵。那么R(A,B,x)的最大值和最小值是什么呢?

首先我们先将广义瑞利商转化为瑞利商的情况，令$x = B^{-1/2}x'$。则其分母变为

$$\begin{aligned} x^H Bx &= x'^H (B^{-1/2})^H BB^{-1/2}x' \\ &= x'^H B^{-1/2} BB^{-1/2}x' \\ &= x'^H x' \end{aligned}$$

分子转化为

$$x^H Ax = x'^H B^{-1/2} AB^{-1/2}x'$$

此时R(A,B,x)转变为R(A,B,x')

$$R(A,B,x') = \frac{x'^H B^{-1/2} AB^{-1/2}x'}{x'^H x'}$$

利用前面的瑞利商性质，我们可以知道，R(A,B,x)的最大值为矩阵$B^{-1/2}AB^{-1/2}$的最大特征值，或者说矩阵$B^{-1}A$的最大特征值，最小值为$B^{-1}A$的最小特征值。

3.二类LDA原理

假如我们数据集为$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$，其中任意样本$x_i$为n维向量，$y_i \in \{0, 1\}$。我们定义$N_j(j = 0, 1)$为第j类样本的个数，$X_j(j = 0, 1)$为第j类样本的结合，$\mu_j(j = 0, 1)$为第j类样本的均值向量，$\sum_j(j = 0, 1)$为第j类样本的协方差矩阵(严格来说是缺少分母部分的协方差矩阵)。其中

$$\mu_j = \frac{1}{N_j}\sum_{x \in X_j} x(j = 0, 1)$$

$$\sum_j = \sum_{x \in X_j}(x - \mu_j)(x - \mu_j)^T \ (j = 0, 1)$$

由于是两类数据，因此我们只需要将数据投影到一条直线上即可。假设我们的投影直线向量为w，则对于任意一个样本x_i，它在直线w的投影为$w^T x_i$。对于我们两个类别的中心点μ_0, μ_1来说，在直线w的投影为$w^T\mu_0, w^T\mu_1$。

由于LDA需要让不同类别数据的中心之间距离尽可能的大，也就是要最大化$||w^T\mu_0 - w^T\mu_1||$。同时需要让同一类别数据的投影点尽可能的接近，也就是要最小化$w^T\sum_0 w + w^T\sum_1 w$。因此，我们的优化目标变为

$$\underbrace{arg\max}_{w} J(w) = \frac{||w^T\mu_0 - w^T\mu_1||_2^2}{w^T\sum_0 w + w^T\sum_1 w} = \frac{w^T(\mu_0 - \mu_1)(\mu_0 - \mu_1)^T w}{w^T(\sum_0 + \sum_1)w}$$

同时，定义类内散度矩阵S_w为

$$S_w = \sum_0 + \sum_1 = \sum_{x \in X_0}(x - \mu_0)(x - \mu_0)^T + \sum_{x \in X_1}(x - \mu_1)(x - \mu_1)^T$$

定义类间散度矩阵S_b为

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

这样我们的优化目标变为

$$\underbrace{arg\max}_{w} J(w) = \frac{w^T S_b w}{w^T S_w w}$$

上述形式便是我们第二节介绍到的广义瑞利商，利用第二节介绍的广义瑞利商的性质，能够得到$J(w)$的最大值为$S_w^{-1/2} S_b S_w^{-1/2}$的最大特征值，而对应的$w$为$S_w^{-1/2} S_b S_w^{-1/2}$的最大特征值所对应的特征向量。

$S_w^{-1} S_b$的特征值和$S_w^{-1/2} S_b S_w^{-1/2}$的特征值相同，$S_w^{-1} S_b$的特征向量$w'$和$S_w^{-1/2} S_b S_w^{-1/2}$的特征向量满足$w' = S_w^{-1/2} w$的关系。注意到对于二类的时候

$$S_b w' = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T w' = (\mu_0 - \mu_1)(w'^T(\mu_0 - \mu_1))^T$$

即$S_b w'$的方向恒为$\mu_0 - \mu_1$，因此令$S_b w' = \lambda(\mu_0 - \mu_1)$，将其带入$(S_w^{-1} S_b) w' = \lambda w'$

$$(S_w^{-1} S_b) w' = S_w^{-1} \lambda(\mu_0 - \mu_1) = \lambda w'$$

$$w' = S_w^{-1}(\mu_0 - \mu_1)$$

也就是说，我们只要求出原始二类样本的均值和方差就可以确定最佳的投影方向w了。

4.多类LDA原理

假如我们数据集为$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$，其中任意样本$x_i$为n维向量，$y_i \in \{C_1, C_2, \ldots, C_k\}$。我们定义$N_j(j = 1, 2, \ldots, k)$为第j类样本的个数，$X_j(j = 1, 2, \ldots, k)$为第j类样本的结合，$\mu_j(j = 1, 2, \ldots, k)$为第j类样本的均值向量，$\sum_j(j = 1, 2, \ldots, k)$为第j类样本的协方差矩阵。

由于是多维向低维投影，此时得到的低位空间就不是一条直线，而是一个超平面。假设投影后得到的低维空间维度为d，对应的基向量为$(w_1, w_2, \ldots, w_d)$，基向量组成的矩阵为W，是一个n*d的矩阵。

此时我们的优化目标变为

$$\frac{W^T S_b W}{W^T S_w W}$$

其中$S_b = \sum_{j=1}^{k} N_j(\mu_j - \mu)(\mu_j - \mu)^T$，$\mu$为所有样本的均值向量。
$S_w = \sum_{j=1}^{k} S_{wj} = \sum_{j=1}^{k} \sum_{x \in X_j} (x - \mu_j)(x - \mu_j)^T$。但是有一个问题，$W^T S_b W$和$W^T S_w W$都是矩阵，不是标量，无法作为一个标量函数来进行优化，怎么办呢。

常见的一个LDA多类优化目标函数定义如下所示，其中$\prod_{diag} A$为A的主对角线元素的乘积，W为n*d的矩阵。

$$\underbrace{arg\max}_{w} J(w) = \frac{\prod_{diag} W^T S_b W}{\prod_{diag} W^T S_w W}$$

J(W)的优化过程可以转化为

$$J(W) = \frac{\prod_{i=1}^{d} w_i^T S_b w_i}{\prod_{i=1}^{d} w_i^T S_w w_i} = \prod_{i=1}^{d} \frac{w_i^T S_b w_i}{w_i^T S_w w_i}$$

上述便是我们前面所介绍的广义瑞利商，最大值便是矩阵$S_w^{-1}S_b$的最大特征值，最大的d个值的乘积就是矩阵$S_w^{-1}S_b$的最大的d个特征值的乘积，此时对应的矩阵W为这最大的d个特征值对应的特征向量张成的矩阵。

由于W是利用了样本的类别得到的投影矩阵，因此它的降维到的维度d最大值为k-1。为什么最大维数不是类别数k呢？因为S_b中每个$\mu_j - \mu$的秩为1，因此协方差矩阵相加后最大的秩为k(矩阵的秩小于等于各个相加矩阵的秩的和)，但是由于如果我们知道前k-1个μ_j后，最后一个μ_k可以由前k-1个μ_j线性表示，因此S_b的秩最大为k-1，即特征向量最多有k-1个。

5.LDA算法流程

输入：数据集$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$，其中任意样本$x_i$为n维向量，$y_i \in \{C_1, C_2, \ldots, C_k\}$，降维到的维度d。

输出：降维后的样本集D'

算法过程：

- 计算类内散度矩阵S_w。
- 计算类内散度矩阵S_b。
- 计算矩阵$S_w^{-1}S_b$。
- 计算$S_w^{-1}S_b$的最大的d个特征值和对应的d个特征向量$w_1, w_2, \ldots, w_d$，得到投影矩阵W。
- 对样本集中每一个样本特征x_i，计算得到新样本$z_i = W^T x_i$。
- 得到输出样本集$D' = \{(z_1, y_1), (z_2, y_2), \ldots, (z_m, y_m)\}$。

6.LDA vs PCA

LDA和PCA有很多相同点和不同点，我们来对比看看两者的区别。

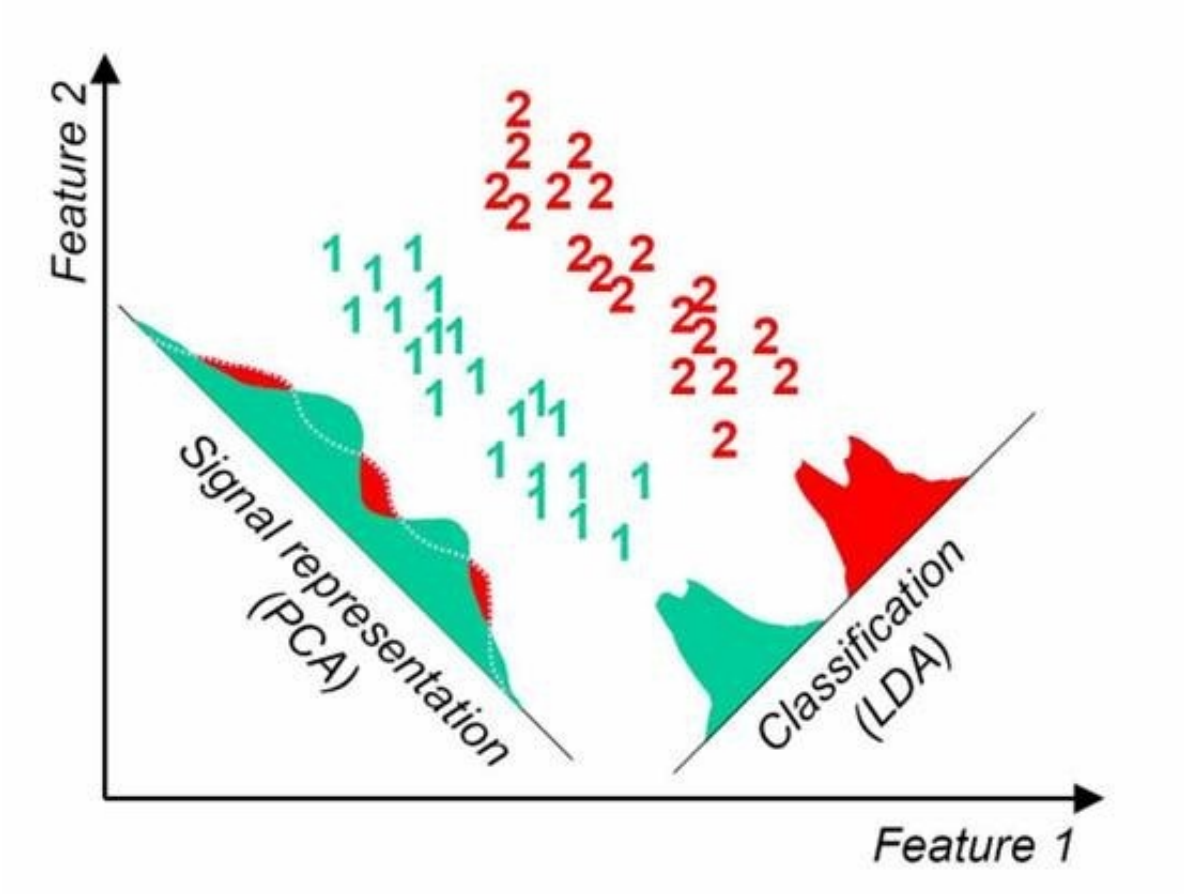
相同点

- 两者均可对数据进行降维。
- 两者在降维时均使用了矩阵特征分解的思想。
- 两者都假设数据符合高斯分布。

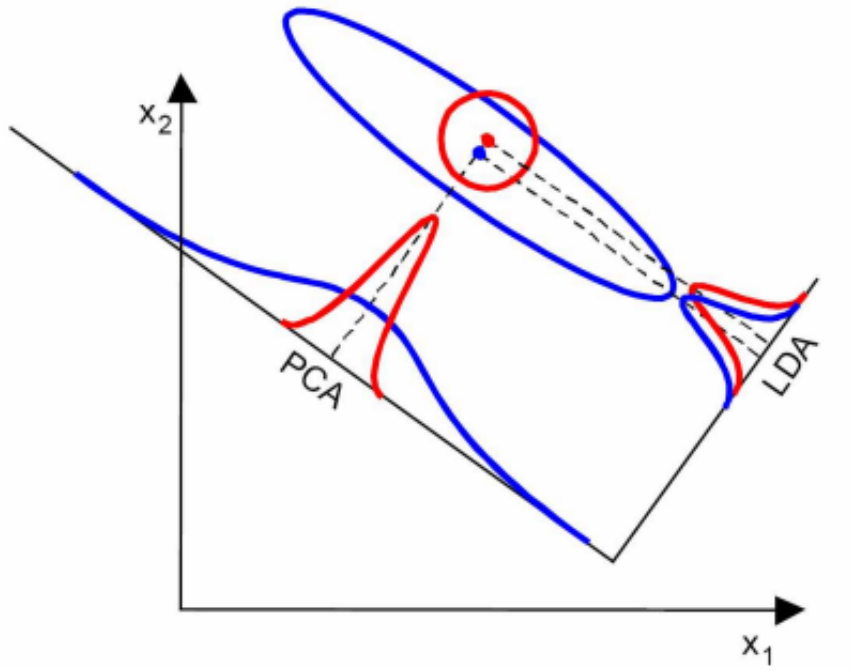
不同点

- LDA是有监督的降维方法，而PCA是无监督的降维方法。
- LDA降维最多降到类别数k-1的维数，而PCA无此限制。
- 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。

不同数据情况下，LDA和PCA降维方法各有优劣。例如某些数据情况下LDA比PCA方法更好



某些数据情况下PCA比LDA方法降维更好



7.LDA算法总结

LDA优点

- 在降维过程中可以使用类别的先验知识经验，而PCA这种无监督学习则无法使用类别先验知识。

LDA缺点

- LDA可能过度拟合数据。
- LDA不适合对非高斯分布样本进行降维，PCA也有这个问题。
- LDA降维最多降到类别数k-1的维数，如果我们降维的维数大于k-1，则不能使用LDA。

8.推广

更多内容请关注公众号**谓之小一**，若有疑问可在公众号后台提问，随时回答，欢迎关注，内容转载请注明出处。

由锤子便签发送 via Smartisan Notes