线性判别分析（Linear Discriminant Analysis， LDA）

LDA是一种经典的降维方法，和主成份分析PCA不考虑样本类别输出的无监督降维技术有所区别，LDA是一种监督学习的将为技术，数据集的每个样本有类别输出。

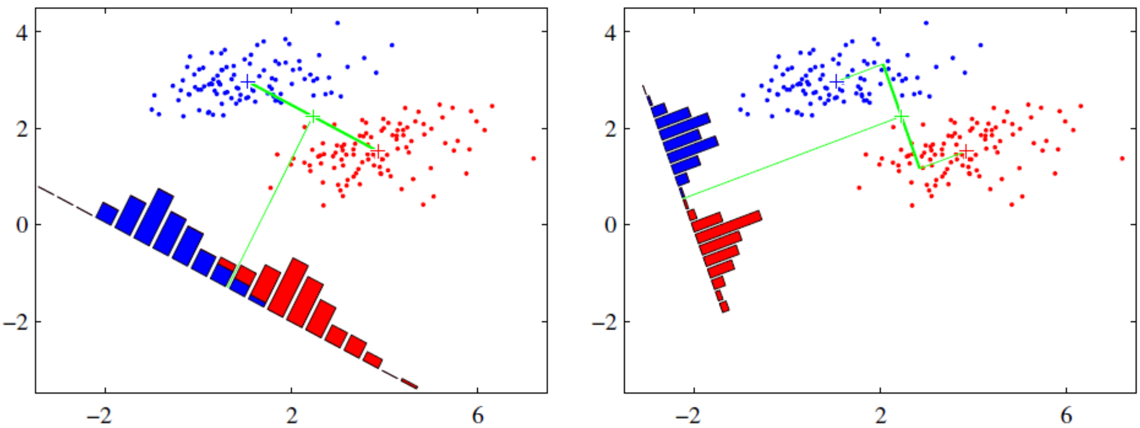
1. LDA思想总结

“投影后类内方差最小，类间方差最大”

1. LDA将多维空间中的数据投影到一条直线上，将d维数据转化成以为数据进行处理

```
2. 将多维训练数据投影到一条直线上，同类数据的投影点尽量接近，异类数据点尽量远离。
```

3. 对数据分类时，将其投影到同样的这条直线上，再根据投影点位置确定样本类别。



左图：让不同类别的平均点距离最远

右图：让同类别数据相距最近

2. 二类LDA算法原理

输入数据集：$D=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_2,\boldsymbol{y}_2),\ldots,(\boldsymbol{x}_m,\boldsymbol{y}_m)\}$，其中样本 $\boldsymbol{x}_i$ 是n维向量，$\boldsymbol{y}_i \epsilon \{0,1\}$，降维后的目标维度 d。定义：
$N_j(j=0,1)$ 为第 j 类样本个数；
$X_j(j=0,1)$ 为第 j 类样本的集合；
$u_j(j=0,1)$ 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。
其中：

$$u_j = \frac{1}{N_j}\sum_{\boldsymbol{x}\epsilon X_j}\boldsymbol{x}(j=0,1),\ \sum_j = \sum_{\boldsymbol{x}\epsilon X_j}(\boldsymbol{x}-u_j)(\boldsymbol{x}-u_j)^T(j=0,1) \tag{1}$$

假设投影直线是向量 $\boldsymbol{w}$，对任意样本 $\boldsymbol{x}_i$，它在直线 w 上的投影为 $\boldsymbol{w}^T x_i$，两个类别的中心点 u_0, u_1 在直线 w 的投影分别为 $\boldsymbol{w}^T u_0$ 、$\boldsymbol{w}^T u_1$。

LDA的目标是让两类别的数据中心间的距离 $\|\boldsymbol{w}^T u_0 - \boldsymbol{w}^T u_1\|_2^2$ 尽量大，与此同时，希望同类样本投影点的协方差 $\boldsymbol{w}^T \sum_0 \boldsymbol{w}$、$\boldsymbol{w}^T \sum_1 \boldsymbol{w}$ 尽量小，最小化 $\boldsymbol{w}^T \sum_0 \boldsymbol{w} - \boldsymbol{w}^T \sum_1 \boldsymbol{w}$ 。

可以定义**类散度矩阵**：

$$S_w = \sum_0 + \sum_1 = \sum_{\boldsymbol{x}\epsilon X_0}(\boldsymbol{x}-u_0)(\boldsymbol{x}-u_0)^T + \sum_{\boldsymbol{x}\epsilon X_1}(\boldsymbol{x}-u_1)(\boldsymbol{x}-u_1)^T \tag{2}$$

类间散度矩阵：

$$S_b = (u_0 - u_1)(u_0 - u_1)^T \tag{3}$$

则优化目标为：

$$\arg\max_{\boldsymbol{w}} J(\boldsymbol{w}) = \frac{\|\boldsymbol{w}^T u_0 - \boldsymbol{w}^T u_1\|_2^2}{\boldsymbol{w}^T \sum_0 \boldsymbol{w} + \boldsymbol{w}^T \sum_1 \boldsymbol{w}} = \frac{\boldsymbol{w}^T (u_0 - u_1)(u_0 - u_1)^T \boldsymbol{w}}{\boldsymbol{w}^T (\sum_0 + \sum_1)\boldsymbol{w}} = \frac{\boldsymbol{w}^T S_b \boldsymbol{w}}{\boldsymbol{w}^T S_w \boldsymbol{w}} \tag{4}$$

根据广义瑞丽商的性质，矩阵 $S_w^{-1} S_b$ 的最大特征值为 $J(\boldsymbol{w})$ 的最大值，矩阵 $S_w^{-1} S_b$ 的最大特征值对应的特征向量即为 $\boldsymbol{w}$。

> 注：瑞丽商（Rayleigh quotient）与广义瑞丽商（generalized Rauleigh quotient)
>
> 定义：称函数 $R(A,x)$ 为瑞丽商,若 $R(A,x)$ 满足：
>
> $$R(A,x) = \frac{x^H A x}{x^H x} \tag{5}$$
>
> 其中 x 为非零向量，而 A 是 $n\times n$ 的Hermitan矩阵（Hermitan矩阵就是满足共轭转置矩阵　　与自身相等的矩阵），若 A 是实矩阵，则满足 $A^T = A$ 的矩阵即为Hermitan矩阵。
>
> 性质：$R(A,x)$ 最大值等于 A 的最大特征值，最小值等于 A 的最小特征值，即：

> $$\lambda_{\min} \leq \frac{x^H A x}{x^H x} \leq \lambda_{max} \tag{6}$$
>
> 当x是标准正交基时，即满足：$x^H x = 1$时，瑞丽商退化为：$R(A,x) = x^H A x$
>
> 定义：称函数$R(A,B,x)$为广义瑞丽商，若$R(A,B,x)$满足：
>
> $$R(A,x) = \frac{x^H A x}{x^H B x} \tag{7}$$
>
> 其中x为非零向量，而A，B是$n \times n$的Hermitan矩阵，且B是正定矩阵。
>
> 广义瑞丽商函数的最大、最小值推导：
>
> > 思路：将其通过标准化转为瑞丽商的形式
> >
> > 令$x' = B^{-\frac{1}{2}} x$，则分母化为：
> >
> > $$x^H B x = x^H \left(B^{-\frac{1}{2}}\right)^H B B^{-\frac{1}{2}} x' = x^H B^{-\frac{1}{2}} B B^{-\frac{1}{2}} x' = x^H x' \tag{8}$$
> >
> > 分子化为：
> >
> > $$x^H A x = x'^H B^{-\frac{1}{2}} A B^{-\frac{1}{2}} x' \tag{9}$$
> >
> > 此时将$R(A,B,x)$转化为$R(A,B,x')$:
> >
> > $$R\left(A,B,x'\right) = \frac{x'^H B^{-\frac{1}{2}} A B^{-\frac{1}{2}} x'}{x'^H x'} \tag{10}$$
> >
> > 则$R(A,B,x)$的最大值为矩阵$B^{-\frac{1}{2}} A B^{-\frac{1}{2}}$的最大特征值，最小值为矩阵$B^{-\frac{1}{2}} A B^{-\frac{1}{2}}$的最小特征值。

3. LDA算法流程

LDA算法降维流程如下：

输入：

数据集 $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$，其中样本 x_i 是n维向量，$y_i \epsilon \{C_1, C_2, \ldots, C_k\}$，降维后的目标维度 d 。

输出：

降维后的数据集 $\overline{D}$ 。

步骤：

1. 计算类内散度矩阵 S_w。
2. 计算类间散度矩阵 S_b 。
3. 计算矩阵 $S_w^{-1}S_b$ 。
4. 计算矩阵 $S_w^{-1}S_b$ 的最大的 d 个特征值。
5. 计算 d 个特征值对应的 d 个特征向量，记投影矩阵为 W 。
6. 转化样本集的每个样本，得到新样本 $P_i = W^T x_i$ 。
7. 输出新样本集 $\overline{D} = \{(p_1, y_1), (p_2, y_2), \ldots, (p_m, y_m)\}$

4.LDA与PCA对比

异同点	**LDA**	**PCA**
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	有监督的降维方法；	无监督的降维方法；
	降维最多降到k-1维；	降维多少没有限制；
	可以用于降维，还可以用于分类；	只用于降维；
	选择分类性能最好的投影方向；	选择样本点投影具有最大方差的方向；
	更明确，更能反映样本间差异；	目的较为模糊；

5. LDA优缺点

	简要说明
优点	1. 可以使用类别的先验知识； 2. 以标签、类别衡量差异性的有监督降维方式，相对于PCA的模糊性，其目的更明确，更能反映样本间的差异；
缺点	1. LDA不适合对非高斯分布样本进行降维； 2. LDA降维最多降到分类数k-1维； 3. LDA在样本分类信息依赖方差而不是均值时，降维效果不好； 4. LDA可能过度拟合数据。