LDA的降维原理

0. 参考资料

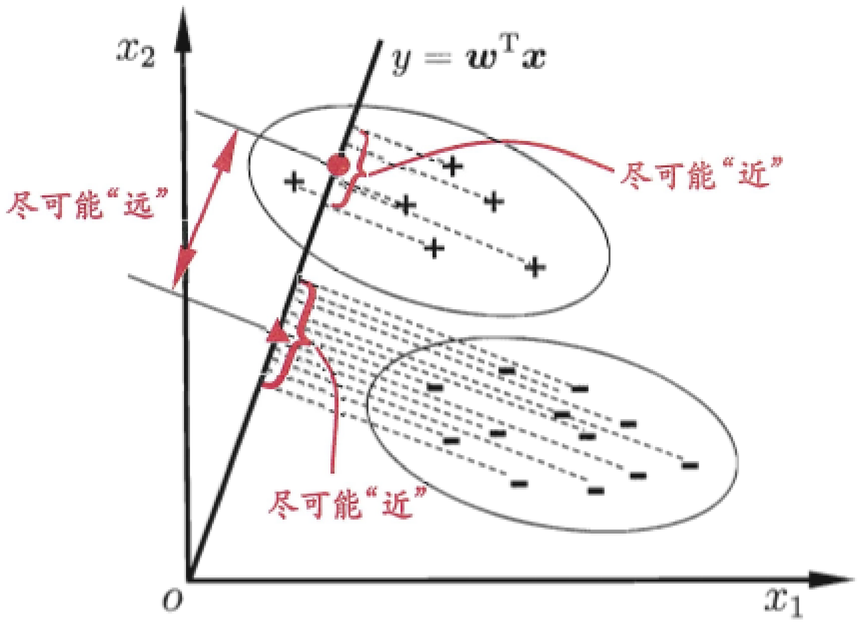
CSDN [线性判别分析LDA原理总结]

CSDN [【机器学习】LDA线性判别分析]

周志华《机器学习》中对应章节

1. LDA的思想

可以从几何角度来理解LDA降维的原理。LDA是一种监督降维技术，其思想非常简单，给定训练样本，设法将样本投影到一条直线（或者超平面）上，使得**同类样本的的投影点尽可能接近、异类样本的投影点尽可能远离。**



用数学语言来描述，就是要求不同类别样本的均值相差尽可能大，同类样本的协方差尽可能小。

2. 二分类LDA原理

我们首先从简单的二分类LDA开始，来分析LDA的原理。

假设数据集$D=\{(x_1,y_1),(x_2,y_2),\ldots,((x_m,y_m))\}$，其中任意样本$x_i$为$n$维向量，$y_i\in\{0,1\}$。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$j$类样本的集合，而$\mu_j(j=0,1)$为第$j$类样本的均值向量，定义$\Sigma_j(j=0,1)$为第$j$类样本的协方差矩阵。

μ_j的表达式为：

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

Σ_j的表达式为：

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

若将数据投影到直线w上，则两个类别的中心点在直线w上的投影为$w^T\mu_0$和$w^T\mu_1$，同类样本点投影的协方差为$w^T\sum_0 w$和$w^T\sum_1 w$。根据LDA的思想，我们要最大化$\|w^T\mu_0 - w^T\mu_1\|_2^2$，同时最小化$w^T\Sigma_0 w + w^T\Sigma_1 w$。于是得到欲最大化的目标为：

$$J(w) = \frac{\|w^T\mu_0 - w^T\mu_1\|_2^2}{w^T\Sigma_0 w + w^T\Sigma_1 w} = \frac{w^T(\mu_0-\mu_1)(\mu_0-\mu_1)^T w}{w^T(\Sigma_0+\Sigma_1)w}$$

我们定义类内散度矩阵S_w为：

$$S_w = \Sigma_0 + \Sigma_1 = \sum_{x \in X_0}(x-\mu_0)(x-\mu_0)^T + \sum_{x \in X_1}(x-\mu_1)(x-\mu_1)^T$$

同时定义类间散度矩阵S_b为：

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

于是优化目标重写为：

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

这就是**广义瑞利商**的形式。因此根据广义瑞利商的性质，$J(w)$最大值为矩阵$S_w^{-\frac{1}{2}} S_b S_w^{-\frac{1}{2}}$的最大特征值，即$S_w^{-1}S_b$的最大特征值。而$w$就是$S_w^{-1}S_b$的最大特征值对应的特征向量，它和$S_w^{-\frac{1}{2}} S_b S_w^{-\frac{1}{2}}$的特征向量$w'$满足关系：

$$w = S_w^{-\frac{1}{2}} w'$$

注意到对于二分类的时候，$S_b w$的方向恒为$\mu_0-\mu_1$，不妨令$S_b w = \lambda(\mu_0-\mu_1)$，将其带入：$\left(S_w^{-1}S_b\right)w = \lambda w$，可得：

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

3. 多分类LDA原理

假设数据集$D=\{(x_1,y_1),(x_2,y_2),\ldots,((x_m,y_m))\}$，其中任意样本$x_i$为$n$维向量，$y_i \in \{C_1,C_2,\ldots,C_k\}$。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$j$类样本的集合，而$\mu_j(j=1,2\ldots k)$为第$j$类样本的均值向量，定义$\Sigma_j(j=1,2\ldots k)$为第$j$类样本的协方差矩阵。

由于我们是多类向低维投影，则此时投影到的低维空间就不是一条直线，而是一个超平面。假设我们投影到的低维空间的维度为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，它是一个$n \times d$的矩阵。

此时我们的优化目标应该可以变成为:

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

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。
$S_w = \sum_{j=1}^k S_{wj} = \sum_{j=1}^k \sum_{x \in X_j}(x-\mu_j)(x-\mu_j)^T$。

这里的问题是，$W^T S_b W$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优化！也就是说，我们无法直接用二类LDA的优化方法，怎么办呢？一般来说，我们可以用其他的一些替代优化目标来实现。

常见的一个LDA多类优化目标函数定义为：

$$J(W) = \frac{\prod_{diag} W^T S_b W}{\prod_{diag} W^T S_w W}$$

其中$\prod_{diag} A$为A的主对角线元素的乘积，W为$n \times d$的矩阵。

$J(W)$的优化过程可以转化为：

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

上式最右边仍为广义瑞利商的形式。因此，最大值是矩阵$S_w^{-1} S_b$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。

需要注意的是，**LDA降维得到的维度d的最大值是$k-1$**。

这是因为S_b中每个$\mu_j - \mu$的秩为1，$(\mu_j - \mu)(\mu_j - \mu)^T$的秩也为1（可由Sylvester不等式得到），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和μ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。直觉上，也可以理解为，当我们知道了前$k-1$个类别的均值后，也就知道了属于这些类别的具体是哪些样本（因为只有这样才可以算出均值），那么属于剩下最后一个类别的样本就是总样本中剩下的那些样本，从而可以求出最后一个均值。

4. LDA降维算法流程

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'。

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$S_w^{-1} S_b$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

5）对样本集中的每一个样本特征x_i，转化为新的样本$z_i = W^T x_i$

6） 得到输出样本集$D' = \{(z_1, y_1), (z_2, y_2), \ldots, ((z_m, y_m))\}$

5. LDA vs PCA

LDA用于降维，和PCA有很多相同，也有很多不同的地方，因此值得好好的比较一下两者的降维异同点。

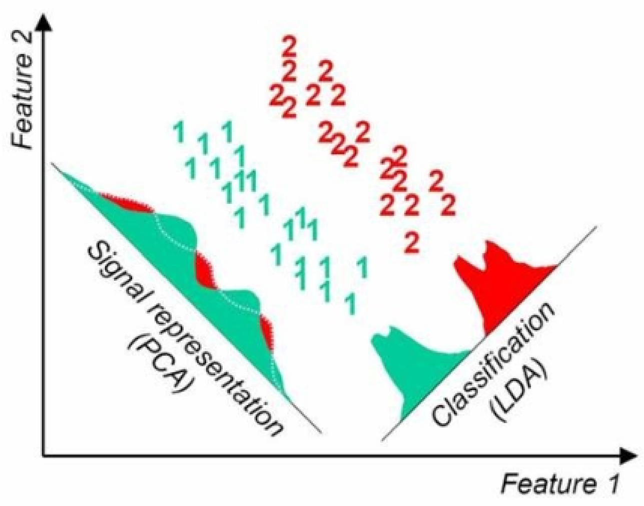
首先我们看看相同点：

- 两者均可以对数据进行降维。
- 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。

我们接着看看不同点：

- 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没有这个限制。
- LDA除了可以用于降维，还可以用于分类。
- 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。

  因此，LDA在样本分类信息依赖均值而不是方差的时候（即均值差别较大时），比PCA之类的算法较优。



LDA在样本分类信息依赖方差而不是均值的时候（即均值相差不大时），降维效果不好。