三、6 线性判别分析

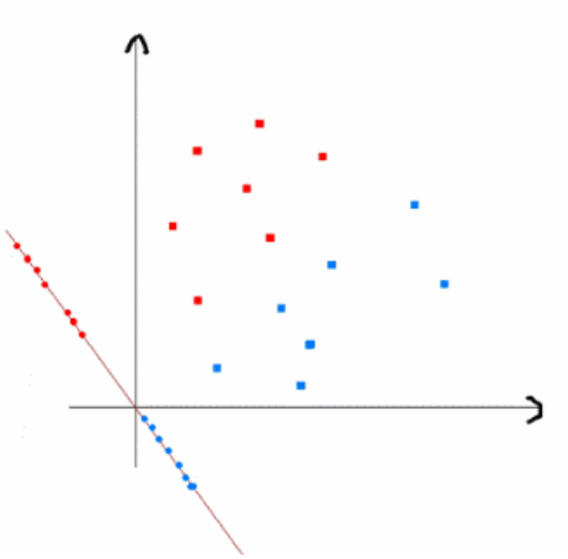
2016年2月14日 星期日　　下午3:56

线性判别分析，LDA，linear discriminant analysis。
二分类问题上的线性判别分析，最早是由Fisher提出的，称作Fisher判别分析；另外还有多重判别分析，对应多分类情况，详见模式分类P99。这里主要介绍二分类的判别分析。

思想：

他是一个有监督的分类器，分类器需要将同类样本尽量靠近，不同类样本尽量分开。LDA的想法非常自然，在训练集在n维空间上的样本点，投影到一条直线上，让同类样本的投足尽量靠近，不同类的尽量分开。这里衡量的方法是，让类均值之间距离尽量大，让类间样本方差尽量小。均值和方差用“除法”结合在一起，作为优化目标。我们优化的结果是，选出一条合适的直线（投影方向）。

LDA是一种非参数的方法（nonparametric procedure）（模式分类P68）



优化目标：

类1的样本点记为x1，类2的记为x2，则经过投影之后分别是$w^T x_1$，$w^T x_2$。投影之后各类的均值分别是$w^T \mu_1$，$w^T \mu_2$，方差是$w^T \Sigma_1 w$，$w^T \Sigma_2 w$，这些都是实数。（Σ代表协方差矩阵）

w, w^T Σ2 w，这些都是实数。（Σ代表协方差矩阵）。

优化目标为J = ||w^T μ1 - w^T μ2||^2 / （w^T Σ1 w + w^T Σ2 w）

为了更好的理解，

定义每一类的类内散度矩阵S_i和总类内散度矩阵S_W

$$\mathbf{S}_i = \sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t$$

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

定义类间散度矩阵S_B

S_B = （μ0 - μ1）（μ0 - μ1）^T

经过化简，最初的损失函数变为

$$J(\mathbf{w}) = \frac{\mathbf{w}^t \mathbf{S}_B \mathbf{w}}{\mathbf{w}^t \mathbf{S}_W \mathbf{w}}$$

这个表达式在数学物理中是经常使用的，通常成为广义的瑞利商。

让J取最大值的w，就是我们要找的投影方向，按照这个方向把样本映射到投影上就ok了。

优化方法：

上面式子J，等价于

min - w^T S_B w s.t. w^T S_W w = 1

可以用拉格朗日乘子法解，详细过程见模式分类和机器学习-周志华

而且这里J对于w是凸函数，有解析解（闭式解）。

结果是：w = S_W^-1（μ0 - μ1）

在实际过程中，通常对S_W进行奇异值分解。

特点：

从被压缩决策理论的角度可以证明当两数据同先验、都满足高斯分布、协方差相等时，LDA可以达到最优分类。（LDA的假设太强了，往往使得LDA并不实用）

对于其他算法对比：

与PCA（模式分类 P96，博客）

与PCA（模式分类P96+博客）

共同点：

A 把高维样本压缩至低维。（主要）

B 在求解过程中，都用到奇异值分解对特征进行降维。（只是凑巧相同）

区别：

A LDA寻找有效的分类方向；PCA寻找有效的主轴方向，通过协方差，找到最大方差的方向。

B LDA是监督学习；PCA是非监督学习。

C 降维后可用维度数量不同。LDA降维后最多可生成C-1维子空间（分类标签数-1），因此LDA与原始维度数量无关，只有数据标签分类数量有关；而PCA最多有n维度可用，即最大可以选择全部可用维度。

D LDA基于贝叶斯决策论，PCA仅仅是数值变换、矩阵变换。

与线性回归（模式分类P198）

通过化简可以得到，两者的判别函数是一致的（虽然原理上没有交集）。

都是 y = w^T x

线性回归通过化简可以得到，（公式里m1、m2是类1和2的均值）

$$\mathbf{w} = \alpha n \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

过程详见模式分类 P 198

推广到多分类，详见机器学习-周志华、模式分类

（LDA部分参考了机器学习-周志华、模式分类+http://www.dataivy.cn/blog/%E7%BA%BF%E6%80%A7%E5%88%A4%E5%88%AB%E5%88%86%E6%9E%90linear-discriminant-analysis_lda/）