

LDA 算法入门

一. LDA 算法概述:

线性判别式分析 (Linear Discriminant Analysis, LDA), 也叫做 Fisher 线性判别 (Fisher Linear Discriminant, FLD), 是模式识别的经典算法, 它是在 1996 年由 Belhumeur 引入模式识别和人工智能领域的。性鉴别分析的基本思想是将高维的模式样本投影到最佳鉴别矢量空间, 以达到抽取分类信息和压缩特征空间维数的效果, 投影后保证模式样本在新的子空间有最大的类间距离和最小的类内距离, 即模式在该空间中有最佳的可分离性。因此, 它是一种有效的特征抽取方法。使用这种方法能够使投影后模式样本的类间散布矩阵最大, 并且同时类内散布矩阵最小。就是说, 它能够保证投影后模式样本在新的空间中有最小的类内距离和最大的类间距离, 即模式在该空间中有最佳的可分离性。

二. LDA 假设以及符号说明:

假设对于一个 R^n 空间有 m 个样本分别为 x_1, x_2, \dots, x_m 即 每个 x 是一个 n 行的矩阵, 其中 n_i 表示属于 i 类的样本个数, 假设有一个有 c 个类, 则 $n_1 + n_2 + \dots + n_i + \dots + n_c = m$ 。

S_b 类间离散度矩阵

S_w 类内离散度矩阵

n_i 属于 i 类的样本个数

x_i 第 i 个样本

u 所有样本的均值

u_i 类 i 的样本均值

三. 公式推导, 算法形式化描述

根据符号说明可得类 i 的样本均值为:

$$u_i = \frac{1}{n_i} \sum_{x \in \text{class } i} x \quad (1)$$

同理我们也可以得到总体样本均值:

$$u = \frac{1}{m} \sum_{i=1}^m x_i \dots\dots\dots (2)$$

根据类间离散度矩阵和类内离散度矩阵定义，可以得到如下式子：

$$S_b = \sum_{i=1}^c n_i (u_i - u)(u_i - u)^T \dots\dots\dots (3)$$

$$S_w = \sum_{i=1}^c \sum_{x_k \in \text{class } i} (u_i - x_k)(u_i - x_k)^T \dots\dots\dots (4)$$

当然还有另一种类间类内的离散度矩阵表达方式：

$$S_b = \sum_{i=1}^c P(i) (u_i - u)(u_i - u)^T$$

$$S_w = \sum_{i=1}^c \frac{P(i)}{n_i} \sum_{x_k \in \text{class } i} (u_i - x_k)(u_i - x_k)^T = \sum_{i=1}^c P(i) E\{(u_i - x)(u_i - x)^T | x \in \text{class } i\}$$

其中 $P(i)$ 是指 i 类样本的先验概率，即样本中属于 i 类的概率 ($P(i) = \frac{n_i}{m}$)，把 $P(i)$ 代入第

二组式子中，我们可以发现第一组式子只是比第二组式子都少乘了 $1/m$ ，我们将在稍后进行

讨论，其实对于乘不乘该 $1/m$ ，对于算法本身并没有影响，现在我们分析一下算法的思想，

我们可以知道矩阵 $(u_i - u)(u_i - u)^T$ 的实际意义是一个[协方差矩阵](#)，这个矩阵所刻画的是该类与样本总体之间的关系，其中该矩阵对角线上的函数所代表的是该类相对样本总体的方差（即分散度），而非对角线上的元素所代表是该类样本总体均值的协方差（即该类和总体样本的相关联度或称冗余度），所以根据公式（3）可知（3）式即把所有样本中各个样本根据自己所属的类计算出样本与总体的协方差矩阵的总和，这从宏观上描述了所有类和总体之间的离散冗余程度。同理可以得出（4）式中为分类内各个样本和所属类之间的协方差矩阵之和，它所刻画的是从总体来看类内各个样本与类之间（这里所刻画的类特性是由是类内各个样本的平均值矩阵构成）离散度，其实从中可以看出不管是类内的样本期望矩阵还是总体样本期望矩阵，它们都只是充当一个媒介作用，不管是类内还是类间离散度矩阵都是从宏观上刻画出类与类之间的样本的离散度和类内样本和样本之间的离散度。

LDA 做为一个分类的算法，我们当然希望它所分的类之间耦合度低，类内的聚合度高，即类内离散度矩阵中的数值要小，而类间离散度矩阵中的数值要大，这样的分类的效果才好。这里我们引入 Fisher 鉴别准则表达式：

$$J_{fisher}(\varphi) = \frac{\varphi^T S_b \varphi}{\varphi^T S_w \varphi} \dots\dots\dots (5)$$

其中 φ 为任一 n 维列矢量。Fisher 线性鉴别分析就是选取使得 $J_{fisher}(\varphi)$ 达到最大值的矢量 φ 作为投影方向，其物理意义就是投影后的样本具有最大的类间离散度和最小的类内离散度。

我们把公式 (4) 和公式 (3) 代入公式 (5) 得到：

$$J_{fisher}(\varphi) = \frac{\sum_{i=1}^c n_i \varphi^T (u_i - u)(u_i - u)^T \varphi}{\sum_{i=1}^c \sum_{x_k \in class i} \varphi^T (u_i - x_k)(u_i - x_k)^T \varphi}$$

我们可以设矩阵 $R = \varphi^T (u_i - u)$ ；其中 φ 可以看成是一个空间，也就是说 $\varphi^T (u_i - u)$ 是 $(u_i - u)$ 构成的低维空间（超平面）的投影。 $\varphi^T (u_i - u)(u_i - u)^T \varphi$ 也可表示为 RR^T ，而当

样本为列向量时， RR^T 即表示 $(u_i - u)$ 在 φ 空间的几何距离的平方。所以可以推出 Fisher 线性鉴别分析表达式的分子即为样本在投影 φ 空间下的类间几何距离的平方和，同理也可推出分母为样本在投影 φ 空间下的类内几何距离的平方和，所以分类问题就转化到找一个低维空间使得样本投影到该空间下时，投影下来的类间距离平方和与类内距离平方和之比最大，即最佳分类效果。

所以根据上述思想，即通过最优化下面的准则函数找到有一组最优鉴别矢量构成的投影矩阵 W_{opt} （这里我们也可以看出 $1/m$ 可以通过分子分母约掉，所以前面所提到的第一组公式和第二组公式所表达的效果是一样的）。

$$W_{opt} = \arg \max \frac{|W^T S_b W|}{|W^T S_w W|} = [w_1, w_2, \dots, w_n] \dots\dots\dots (6)$$

可以证明，当 S_w 为非奇异（一般在实现 LDA 算法时，都会对样本做一次 PCA 算法的降维，消除样本的冗余度，从而保证 S_w 是非奇异阵，当然即使 S_w 为奇异阵也是可以解的，可以把 S_w 或 S_b 对角化，这里不做讨论，假设都是非奇异的情况）时，最有投影矩阵 W_{opt} 的列向量恰为下列广义特征方程

$$S_b \varphi = \lambda S_w \varphi \dots\dots\dots (7)$$

的 d 个最大的特征值所对应的特征向量(矩阵 $S_w^{-1} S_b$ 的特征向量), 且最优投影轴的个数 $d \leq c-1$.

$$\text{根据 (7) 式可以推出 } S_b \varphi_i = \lambda_i S_w \varphi_i \dots\dots\dots (8)$$

$$\text{又由于 } W = [\varphi_1, \varphi_2, \dots, \varphi_d]$$

下面给出验证: 把 (7) 式代入 (6) 式可得:

$$\begin{aligned} S_b \varphi_1 &= \lambda_1 S_w \varphi_1 \\ \vdots &= \vdots \Rightarrow \\ S_b \varphi_d &= \lambda_d S_w \varphi_d \\ \varphi_1^T S_b \varphi_1 &= \varphi_1^T \lambda_1 S_w \varphi_1 \\ \vdots &= \vdots \Rightarrow \end{aligned}$$

$$\varphi_d^T S_b \varphi_d = \varphi_d^T \lambda_d S_w \varphi_d$$

$$\text{又 } \because W = [\varphi_1, \varphi_2, \dots, \varphi_d]$$

\therefore 根据公式 (6) 分解可知公式 (6)

$$\begin{aligned} \text{求的是方程 } \max &= \frac{|\varphi_i^T S_b \varphi_i|}{|\varphi_i^T S_w \varphi_i|} = \frac{|\varphi_i^T \lambda_i S_w \varphi_i|}{|\varphi_i^T S_w \varphi_i|} \\ &= \frac{\lambda_i |\varphi_i^T S_w \varphi_i|}{|\varphi_i^T S_w \varphi_i|} = \lambda_i \end{aligned}$$

\therefore 根据公式意义来看 要使得 \max 最大则只要取

λ_i 即可, 所以根据公式 (6) 可得出结论:

投影矩阵 W_{opt} 的列向量为 d (自取) 个最大特征值所对应的特征向量, 其中 $d \leq c-1$.

四. 算法的物理意义和思考

4.1 用一个例子阐述 LDA 算法在空间上的意义

下面我们利用 LDA 进行一个分类的问题: 假设一个产品有两个参数来衡量它是否合格, 我们假设两个参数分别为:

参数 A	参数 B	是否合格
2.95	6.63	合格
2.53	7.79	合格
3.57	5.65	合格
3.16	5.47	合格
2.58	4.46	不合格
2.16	6.22	不合格
3.27	3.52	不合格

实验数据来源: <http://people-revoledu.com/kardi/tutorial/LDA/Numerical%20Example.html>

所以我们可以根据上图表格把样本分为两类，一类是合格的，一类是不合格的，所以我们可以创建两个数据集类：

`cls1_data =`

```
2.9500    6.6300
2.5300    7.7900
3.5700    5.6500
3.1600    5.4700
```

`cls2_data =`

```
2.5800    4.4600
2.1600    6.2200
3.2700    3.5200
```

其中 `cls1_data` 为合格样本，`cls2_data` 为不合格的样本，我们根据公式（1），（2）可以算出合格的样本的期望值，不合格类样本的合格的值，以及总样本期望：

`E_cls1 =`

```
3.0525    6.3850
```

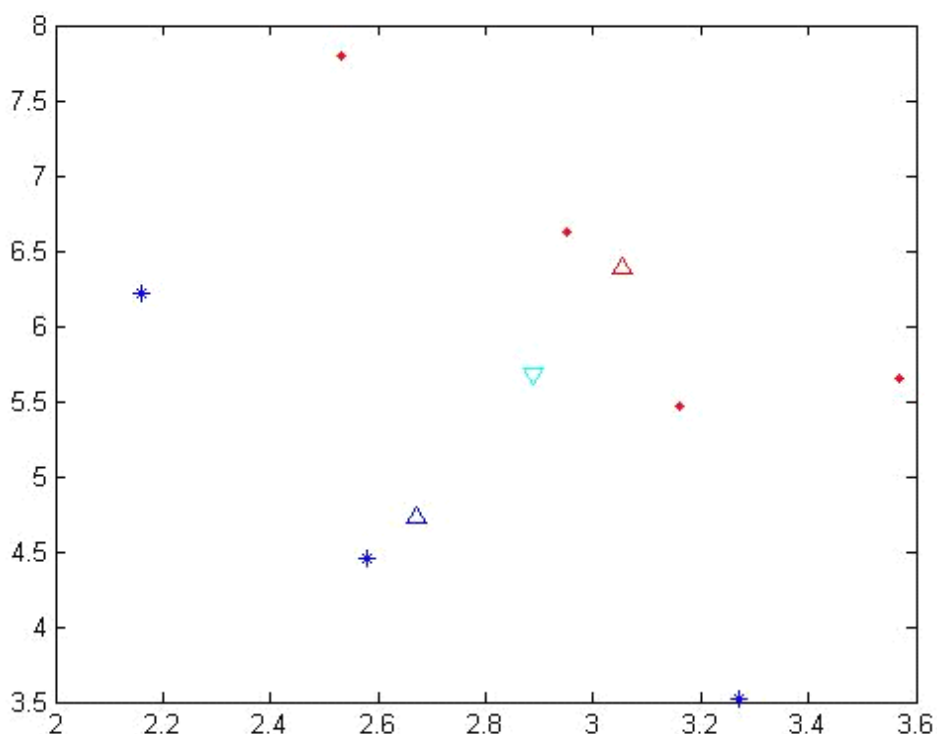
`E_cls2 =`

```
2.6700    4.7333
```

`E_all =`

```
2.8886    5.6771
```

我们可以做出现在各个样本点的位置：



图一

其中蓝色“*”的点代表不合格的样本，而红色实点代表合格的样本，天蓝色的倒三角是代表总期望，蓝色三角形代表不合格样本的期望，红色三角形代表合格样本的期望。从 x, y 轴的坐标方向上可以看出，合格和不合格样本区分度不佳。

我们可以根据表达式 (3), (4) 可以计算出类间离散度矩阵和类内离散度矩阵：

$S_b =$

$$\begin{bmatrix} 0.0358 & 0.1547 \\ 0.1547 & 0.6681 \end{bmatrix}$$

$S_w =$

$$\begin{bmatrix} 0.5909 & -1.3338 \\ -1.3338 & 3.5596 \end{bmatrix}$$

我们可以根据公式 (7), (8) 算出 $S_w^{-1} S_b$ 特征值以及对应的特征向量：

$L =$

$$\begin{bmatrix} 0.0000 & 0 \\ 0 & 2.8837 \end{bmatrix}$$

对角线上为特征值，第一个特征值太小被计算机约为 0 了

与他对应的特征向量为

$V =$

$$\begin{bmatrix} -0.9742 & -0.9230 \\ 0.2256 & -0.3848 \end{bmatrix}$$

根据取最大特征值对应的特征向量：(-0.9230, -0.3848)，该向量即为我们要求的子空间，我们可以把原来样本投影到该向量后 得到新的空间（2 维投影到 1 维，应该为一个数字）

$\text{new_cls1_data} =$

-5.2741

-5.3328

-5.4693

-5.0216

为合格样本投影后的样本值

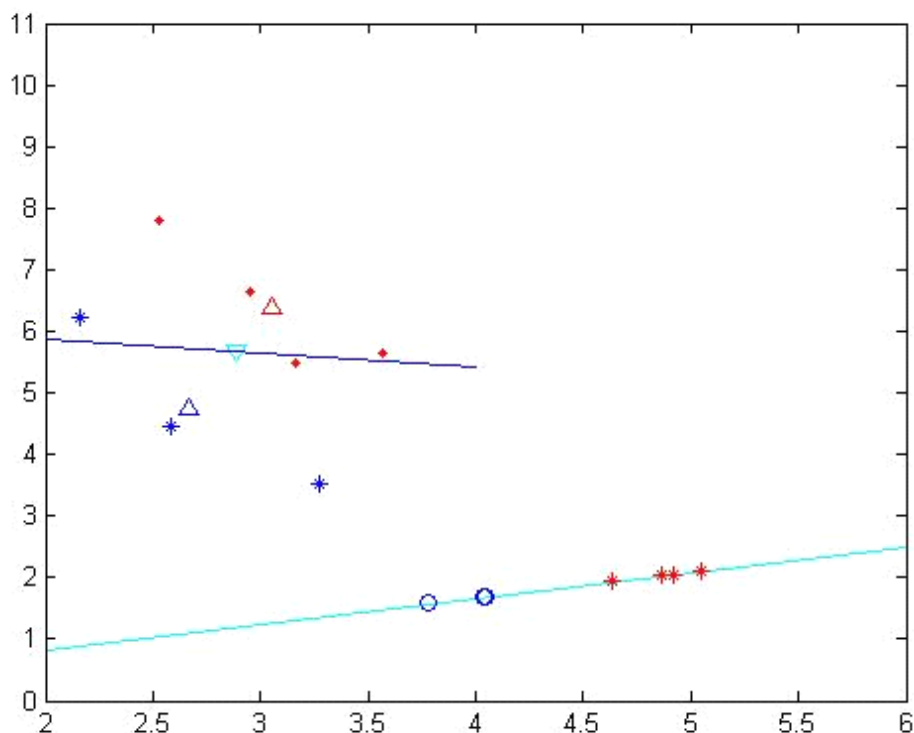
new_cls2_data =

-4.0976

-4.3872

-4.3727

为不合格样本投影后的样本值，我们发现投影后，分类效果比较明显，类和类之间聚合度很高，我们再次作图以便更直观看分类效果



图二

蓝色的线为特征值较小所对应的特征向量，天蓝色的为特征值较大的特征向量，其中蓝色的圈点为不合格样本在该特征向量投影下来的位置，二红色的“*”符号的合格样本投影后的数据集，从中个可以看出分类效果比较好（当然由于 x , y 轴单位的问题投影不那么直观）。我们再利用所得到的特征向量，来对其他样本进行判断看看它所属的类型，我们取样本点 $(2.81, 5.46)$ ，

我们把它投影到特征向量后得到：result = -4.6947 所以它应该属于不合格样本。

4.2 LDA 算法与 PCA 算法

在传统特征脸方法的基础上，研究者注意到特征值打的特征向量（即特征脸）并一定是分类性能最好的方向，而且对 K-L 变换而言，外在因素带来的图像的差异和人脸本身带来

的差异是无法区分的，特征连在很大程度上反映了光照等的差异。研究表明，特征脸，特征脸方法随着光线，角度和人脸尺寸等因素的引入，识别率急剧下降，因此特征脸方法用于人脸识别还存在理论的缺陷。线性判别式分析提取的特征向量集，强调的是不同人脸的差异而不是人脸表情、照明条件等条件的变化，从而有助于提高识别效果。

（说明：以上是本人参阅文献以及自己独立思考理解所作，由于水平有限难免有纰漏以及错误之处，希望读者谅解，欢迎通过邮件跟我交流，邮箱在页眉处）

五. 参考文献

- 【1】《人脸识别的子空间方法研究与实现》 郑自明
- 【2】《基于图像分块的特征提取算法的改进及应用》 王宏玉
- 【3】《人脸识别算法研究和实现》 马驰
- 【4】《基于统计的人脸识别方法》 余冰
- 【5】一种改进的线性判别分析法在人脸识别中的应用 《计算机工程》 2006 覃志祥
- 【6】Linear Discriminant analysis – a brief tutorial S· Balakrishnama, A· Ganapathiraju
- 【7】Eigenfaces vs· Fisherfaces: Recognition Using Class Specific Linear Projection
《PAMI》 Peter N· Belhumeur, Joao P· Hespanha, and David J· Kriegman
- 【8】PCA versus LDA 《PAMI》 Aleix M· MartíÃnez