LDA

类似于无监督聚类方法PCA，LDA引入已知的数据标签。

- PCA基本思路是寻求最优的变换矩阵，使得变换后数据各轴上投影方差和最大(max型目标函数，函数式中不掺杂数据标签)
- LDA基本思路是寻求最优变换矩阵，使得变换后同类数据尽量集聚，异类数据尽量分散，即可形式化为最小化类内样本方差及最大化类间样本方差(min/max型目标函数，且函数式中掺杂了数据标签)

详细推导

原始样本为：D={ (x_1, y_1) ,..., (x_n, y_n) }
xi为 n_1 维向量， y_i 属于 $1, 2, ..., m$ (m类样本)

目标：求最佳转换矩阵，降维至 n_2 维，且同类数据尽量集聚，异类数据尽量分散

降维前后第k类中某一样本分别记为： x_i^k / z_i^k
设降维变换阵为P， z_i^k = P * x_i^k
这里的 P 即为降维后所在低维空间的基向量行组

(A)类内样本分散程度度量

(1)自然地，从代数角度看，我们容易联想到用协方差矩阵表示多维变量的分散程度。
降维后第 k 类样本类内方差为

$$S_k = \frac{1}{N_k} \sum_{i=1}^{N_k} (z_i^k - \mu_k^-)(z_i^k - \mu_k^-)^T$$

$$\mu_k^- = \frac{1}{N_k} \sum_{i=1}^{N_k} z_i^k$$

则有

$$\sum_{1 \le i \le m} S_k = \frac{1}{N_k} \sum_{k=1}^{m} \sum_{i=1}^{N_k} (z_i^k - \mu_k^-)(z_i^k - \mu_k^-)^T = P S_w P^T$$

S_w 称为类内散布矩阵

$$S_w = \frac{1}{N_k} \sum_{k=1}^{m} \sum_{i=1}^{N_k} (x_i^k - \mu_k)(x_i^k - \mu_k)^T$$

其中

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N_k} x_i^k$$

(2)从几何上看，我们也可利用某类数据样本中各样本点与均值点的欧式距离之和来表示该类的分散程度(代数上即为向量二范数的和)
对于第k类样本，该度量标准可写为

$$g_k = \frac{1}{N_k}\sum_{i=1}^{N_k}||(z_i^k - \mu_k^-)||_2^2 = \frac{1}{N_k}\sum_{i=1}^{N_k} tr(P(x_i^k - \mu_k)(x_i^k - \mu_k)^T P^T)$$

代入有

$$g_k = \frac{1}{N_k}\sum_{i=1}^{N_k}\sum_{j=1}^{n_2} p_j^T (x_i^k - \mu_k)(x_i^k - \mu_k)^T p_j$$

此处 p_j 为低维空间基向量(列向量)

考虑各类样本总和，可写为

$$G = \sum_{k=1}^{m} g_k = \sum_{j=1}^{n_2} p_j^T S_w p_j$$

(B)类间样本分散程度度量

可用降维后各类样本的均值向量组的分散程度去度量
(1)自然地，从代数角度，可联想起用 m 个均值向量的协方差矩阵表示分散程度
类似于(A)中推导，定义类间散布矩阵如下

$$S_b = \frac{1}{m}\sum_{i=1}^{m}(\mu_k^- - \mu^-)(\mu_k^- - \mu^-)^T$$

μ^- 表示降维后所有样本的均值

继续代入有

$$\frac{1}{m}\sum_{i=1}^{m}(\mu_k^- - \mu^-)(\mu_k^- - \mu^-)^T = PS_bP^T$$

$$S_b = \frac{1}{m}\sum_{i=1}^{m}(\mu_k - \mu)(\mu_k - \mu)^T$$

S_b 称为类间散布矩阵， μ 表示降维前所有样本均值

(2)从几何上看，类似于A中思路，可以用到中心点的欧氏距离之和表示均值向量组的分散程度

类似A中推导

$$B = \frac{1}{m}\sum_{k=1}^{m}\sum_{j=1}^{n_2} p_j^T(\mu_k - \mu)(\mu_k - \mu)^T p_j = \sum_{j=1}^{n_2} p_j^T S_b p_j$$

(C)构造目标函数

最终目标是最大化类间样本分散程度，并最小化类内样本分散程度，考虑将两个目标合二为一

与上文呼应，从代数角度构造如下目标函数

$$J(P)=\frac{PS_bP^T}{PS_wP^T}$$

由(A)(B)推导可知 S_b , S_w 均为对称半正定阵，但由于数据满足高斯分布的假设，可认为 S_w 可逆，可即等价为广义瑞利商函数;
特殊地，当低维空间仅为一维时，J(P)的分子分母均为标量，可直接求解;但一般情形下，分子分母均为矩阵，无法直接相除!
此时呼应(A)(B)中的几何思路，构造如下目标函数

$$J(P)=\frac{B}{G}=(\sum_{j=1}^{n_2}p_j^TS_bp_j)/(\sum_{j=1}^{n_2}p_j^TS_wp_j)$$

此时分子分母均为标量，可以继续求解
事实上，如上两种目标函数的转换关系恰是呼应(A)(B)中所描述的代数思路与几何思路的转换关系
完成了目标函数的初步构造，但进一步求解时发现分子分母都是和式的形式，为了后续方便求最优解，将和式改为乘式，继续转化为最终形态的目标函数，如下

$$J(P)=\prod_{j=1}^{n_2}(p_j^TS_bp_j)/(p_j^TS_wp_j)$$

该函数即为瑞利商型函数的连乘，利用(D)中结论即可对每一项求最优的基向量 p_j ;
此外易看出，当问题较简单，降维子空间为一维时，上述三类目标函数 $J(P)$ 恰好相等，化为最简形式，此时无需过多考虑上述三类目标函数间的变换，可针对第一类J(P)，根据(D)中结论直接上手求解

(D)瑞利商与广义瑞利商

瑞利商定义为

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

其中 A 为Hermite矩阵 瑞利商的定义与矩阵2范数的形式类似，根据矩阵论相关定理，可得如下性质

$$\lambda_1\le\frac{x^HAx}{x^Hx}\le\lambda_2$$

λ_1 / λ_2 为最小/大特征值，且 x 取对应特征向量时取得最大/最小值

进而有广义瑞利商定义:

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

其中 A , B 均为Hermite矩阵,且 B 为正定矩阵

通过变量代换

$$z=B^{\frac{1}{2}}x$$

代入

$$x^HBx=z^H(B^{\frac{-1}{2}})^HBB^{\frac{-1}{2}}z=z^HB^{\frac{-1}{2}}BB^{\frac{-1}{2}}z=z^Hz$$

$$x^H Ax = z^H B^{\frac{-1}{2}} AB^{\frac{-1}{2}} z$$

因此广义瑞利商转化为

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

知广义瑞利商最大值为 $B^{\frac{-1}{2}} AB^{\frac{-1}{2}}$ 的特征值
设 $B^{-1}A$ 的最大特征值为 λ_1 ，对应特征向量为 v_1
$B^{\frac{-1}{2}} AB^{\frac{-1}{2}}$ 最大特征值为 λ_2 ,对应特征向量为 v_2
可导出

$$\lambda_1 = \lambda_2$$

$$v_2 = B^{\frac{1}{2}} v_1$$

即对于广义瑞利商 $R(A,B,z)$ ， $z = v_2$ 时取最大值 λ_2 ;
对于 $R(A,B,x)$, $x = B^{\frac{-1}{2}} v_2 = v_1$ 时取最大值 λ_1

至此可知
$R(A,x)$ 的最大值及对应自变量为 A 的最大特征值及对应特征向量
$R(A,B,x)$ 的最大值及对应自变量为 $B^{-1}A$ 的最大特征值及对应特征向量

(E)LDA目标函数求最优解

至此，可基于(D)中结论，对(C)中最终形态的目标函数(如下)求最优解

$$J(P) = \prod_{j=1}^{n_2} (p_j^T S_b p_j)/(p_j^T S_w p_j)$$

可看出, $J(P)$ 每个乘积项均为广义瑞利商函数，可对第 j 项求最优的 p_j ，合并后即得降维子空间的基向量组
根据(D)中最后的结论，当满足如下条件

$$p_j = eig_j(S_w^{-1} S_b)$$

其中，降维至 n_2 维，则选择降序排列后前 n_2 个特征值 $\lambda_{n_2} \leq ... \leq \lambda_1$ ($j = 1, ..., n_2$)， $eig_j()$ 表示矩阵特征值 λ_j 对应的特征向量；
则

$$J(P) = \prod_{j=1}^{n_2} \lambda_j$$

因此降维子空间的基向量组应取为

$$P = (p_1 ... p_{n_2})$$

附记:

$$S_b = \frac{1}{m} \sum_{i=1}^{m} (\mu_k^- - \mu^-)(\mu_k^- - \mu^-)^T$$

由于

$$rank(\mu_k^- - \mu^-) = 1$$

且有定理

$$rank(AB) \le min(rank(A), rank(B))$$

$$rank(A+B) \le rank(A) + rank(B)$$

则

$$rank(S_b) \le m$$

又

$$\sum_{i=1}^{m} \frac{N_i}{N}(\mu_i^- - \mu^-) = 0$$

则

$$rank(S_b) \le m-1$$

$$rank(S_w^{-1}S_b) \le min(n_1, rank(S_b)) \le m-1$$

可知 $S_w^{-1}S_b$ 特征值最多为 $m-1$ 个，即说明降维子空间维度最高不超过 $m-1$，且 $n_2 \le m-1$

(F)算法提炼

总结(A)(B)(C)(D)，提炼最终算法如下
输入 $D = ((x_1, y_1), ..., (x_k, y_k)), x_i \in R^n$
输出 $D' = ((z_1, y_1), ..., (z_k, y_k)), z_i \in R^m$

- 计算 S_w
- 计算 S_b
- 计算 $S_w^{-1}S_b$ 的最大的 m 个特征值及对应的特征向量 $(p_1, p_2, ..., p_m)$，得到投影矩阵P
- 对所有原始样本变换得 $z = P^T x$,输出降维样本集