

线性判别分析 (Linear Discriminant Analysis)

JerryLead

csxulijie@gmail.com

1. 问题

之前我们讨论的 PCA、ICA 也好，对样本数据来言，可以是没有类别标签 y 的。回想我们做回归时，如果特征太多，那么会产生不相关特征引入、过度拟合等问题。我们可以使用 PCA 来降维，但 PCA 没有将类别标签考虑进去，属于无监督的。

比如回到上次提出的文档中含有“learn”和“study”的问题，使用 PCA 后，也许可以将这两个特征合并为一个，降了维度。但假设我们的类别标签 y 是判断这篇文章的 topic 是不是有关学习方面的。那么这两个特征对 y 几乎没什么影响，完全可以去除。

再举一个例子，假设我们对一张 $100*100$ 像素的图片做人脸识别，每个像素是一个特征，那么会有 10000 个特征，而对应的类别标签 y 仅仅是 0/1 值，1 代表是人脸。这么多特征不仅训练复杂，而且不必要特征对结果会带来不可预知的影响，但我们想得到降维后的一些最佳特征（与 y 关系最密切的），怎么办呢？

2. 线性判别分析（二类情况）

回顾我们之前的 logistic 回归方法，给定 m 个 n 维特征的训练样例 $x^{(i)}\{x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)}\}$ (i 从 1 到 m)，每个 $x^{(i)}$ 对应一个类标签 $y^{(i)}$ 。我们就是要学习出参数 θ ，使得 $y^{(i)} = g(\theta^T x^{(i)})$ (g 是 sigmoid 函数)。

现在只考虑二值分类情况，也就是 $y=1$ 或者 $y=0$ 。

为了方便表示，我们先换符号重新定义问题，给定特征为 d 维的 N 个样例， $x^{(i)}\{x_1^{(i)}, x_2^{(i)}, \dots, x_d^{(i)}\}$ ，其中有 N_1 个样例属于类别 ω_1 ，另外 N_2 个样例属于类别 ω_2 。

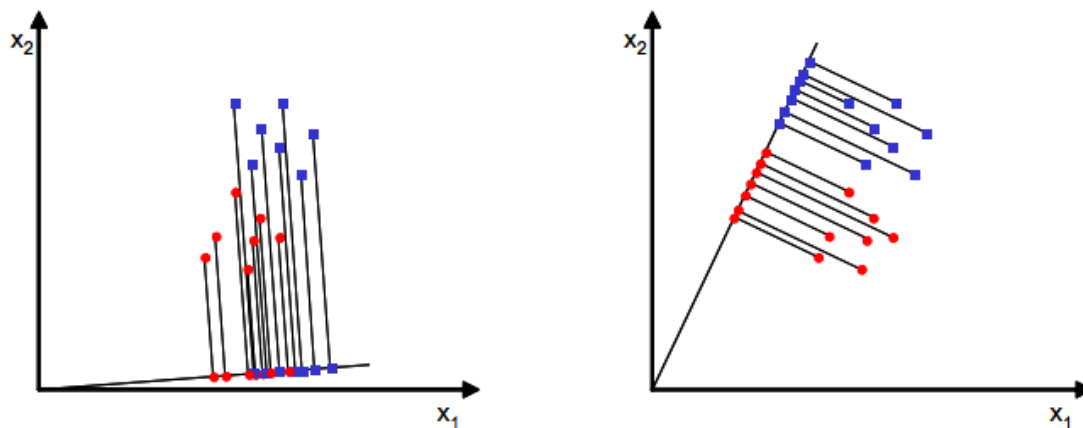
现在我们觉得原始特征数太多，想将 d 维特征降到**只有一维**，而又要保证类别能够“清晰”地反映在低维数据上，也就是这一维就能决定每个样例的类别。

我们将这个最佳的向量称为 w (d 维)，那么样例 x (d 维) 到 w 上的投影可以用下式来计算

$$y = w^T x$$

这里得到的 y 值不是 0/1 值，而是 x 投影到直线上的点到原点的距离。

当 x 是二维的，我们就是要找一条直线（方向为 w ）来做投影，然后寻找最能使样本点分离的直线。如下图：



从直观上来看，右图比较好，可以很好地将不同类别的样本点分离。

接下来我们从定量的角度来找到这个最佳的 w 。

首先我们寻找每类样例的均值（中心点），这里 i 只有两个

$$\mu_i = \frac{1}{N_i} \sum_{x \in \omega_i} x$$

由于 x 到 w 投影后的样本点均值为

$$\tilde{\mu}_i = \frac{1}{N_i} \sum_{y \in \omega_i} y = \frac{1}{N_i} \sum_{x \in \omega_i} w^T x = w^T \mu_i$$

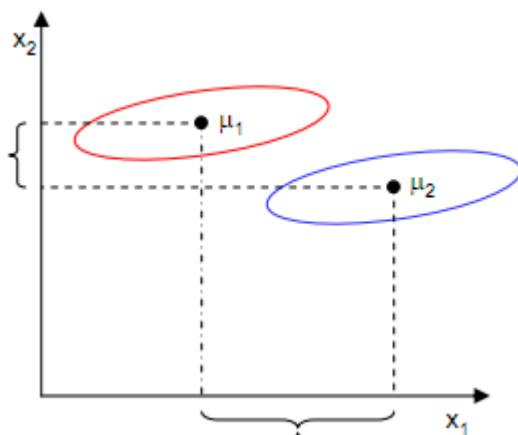
由此可知，投影后的均值也就是样本中心点的投影。

什么是最佳的直线（ w ）呢？我们首先发现，能够使投影后的两类样本中心点尽量分离的直线是好的直线，定量表示就是：

$$J(w) = |\tilde{\mu}_1 - \tilde{\mu}_2| = |w^T(\mu_1 - \mu_2)|$$

$J(w)$ 越大越好。

但是只考虑 $J(w)$ 行不行呢？不行，看下图



样本点均匀分布在椭圆里，投影到横轴 x_1 上时能够获得更大的中心点间距 $J(w)$ ，但是由于有重叠， x_1 不能分离样本点。投影到纵轴 x_2 上，虽然 $J(w)$ 较小，但是能够分离样本点。因此我们还需要考虑样本点之间的方差，方差越大，样本点越难以分离。

我们使用另外一个度量值，称作散列值（scatter），对投影后的类求散列值，如下

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2$$

从公式中可以看出，只是少除以样本数量的方差值，散列值的几何意义是样本点的密集程度，值越大，越分散，反之，越集中。

而我们想要的投影后的样本点的样子是：不同类别的样本点越分开越好，同类的越聚集越好，也就是均值差越大越好，散列值越小越好。正好，我们可以使用 $J(w)$ 和 S 来度量，最终的度量公式是

$$J(w) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

接下来的事就比较明显了，我们只需寻找使 $J(w)$ 最大的 w 即可。

先把散列值公式展开

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2 = \sum_{x \in \omega_i} (w^T x - w^T \mu_i)^2 = \sum_{x \in \omega_i} w^T (x - \mu_i)(x - \mu_i)^T w$$

我们定义上式中中间那部分

$$S_i = \sum_{x \in \omega_i} (x - \mu_i)(x - \mu_i)^T$$

这个公式的样子不就是少除以样例数的协方差矩阵么，称为散列矩阵（scatter matrices）

我们继续定义

$$S_w = S_1 + S_2$$

S_w 称为 **Within-class scatter matrix**。

那么回到上面 \tilde{s}_i^2 的公式，使用 S_i 替换中间部分，得

$$\tilde{s}_i^2 = w^T S_i w$$

$$\tilde{s}_1^2 + \tilde{s}_2^2 = w^T S_w w$$

然后，我们展开分子

$$(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (w^T \mu_1 - w^T \mu_2)^2 = w^T \underbrace{(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T}_{S_B} w = w^T S_B w$$

S_B 称为 **Between-class scatter**，是两个向量的外积，虽然是个矩阵，但秩为 1。

那么 $J(w)$ 最终可以表示为

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

在我们求导之前，需要对分母进行归一化，因为不做归一的话， w 扩大任何倍，都成立，我们就无法确定 w 。因此我们打算令 $\|w^T S_W w\| = 1$ ，那么加入拉格朗日乘子后，求导

$$c(w) = w^T S_B w - \lambda(w^T S_W w - 1)$$

$$\Rightarrow \frac{dc}{dw} = 2S_B w - 2\lambda S_W w = 0$$

$$\Rightarrow S_B w = \lambda S_W w$$

其中用到了矩阵微积分，求导时可以简单地把 $w^T S_W w$ 当做 $S_W w^2$ 看待。如果 S_W 可逆，那么将求导后的结果两边都乘以 S_W^{-1} ，得

$$S_W^{-1} S_B w = \lambda w$$

这个可喜的结果就是 w 就是矩阵 $S_W^{-1} S_B$ 的特征向量了。
这个公式称为 Fisher linear discrimination。

等等，让我们再观察一下，发现前面 S_B 的公式

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

那么

$$S_B w = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w = (\mu_1 - \mu_2) * \lambda_w$$

代入最后的特征值公式得

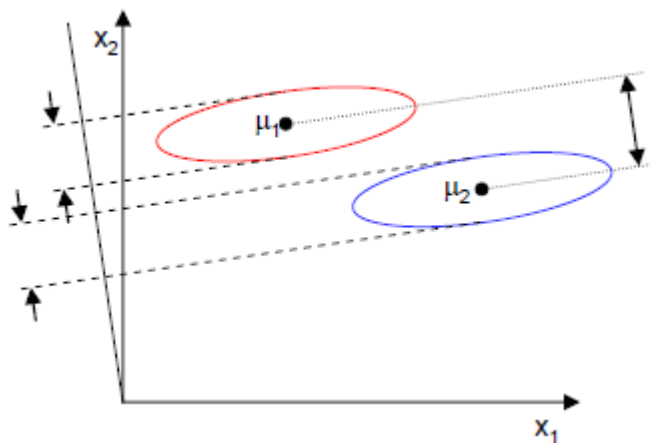
$$S_W^{-1} S_B w = S_W^{-1} (\mu_1 - \mu_2) * \lambda_w = \lambda w$$

由于对 w 扩大缩小任何倍不影响结果，因此可以约去两边的未知常数 λ 和 λ_w ，得到

$$w = S_W^{-1} (\mu_1 - \mu_2)$$

至此，我们只需要求出原始样本的均值和方差就可以求出最佳的方向 w ，这就是 Fisher 于 1936 年提出的线性判别分析。

看上面二维样本的投影结果图：



3. 线性判别分析（多类情况）

前面是针对只有两个类的情况，假设类别变成多个了，那么要怎么改变，才能保证投影后类别能够分离呢？

我们之前讨论的是如何将 d 维降到一维，现在类别多了，一维可能已经不能满足要求。假设我们有 C 个类别，需要 K 维向量（或者叫做基向量）来做投影。

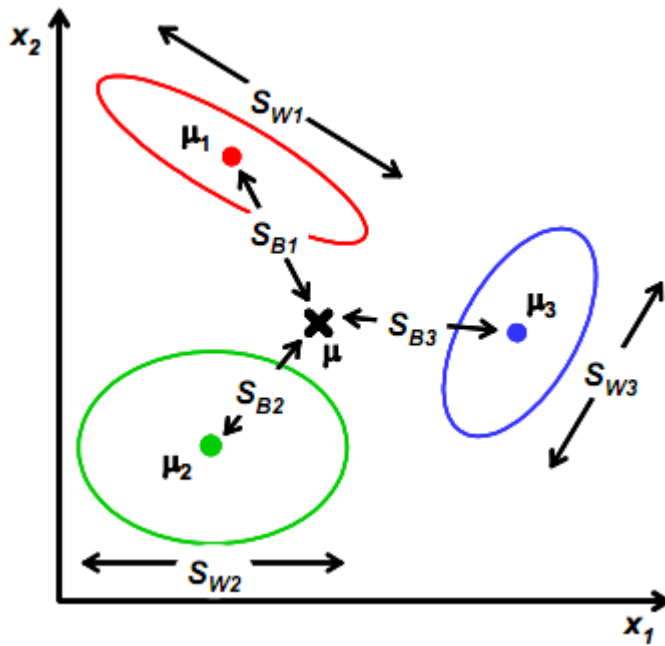
将这 K 维向量表示为 $W = [w_1 | w_2 | \dots | w_K]$ 。

我们将样本点在这 K 维向量投影后结果表示为 $[y_1, y_2, \dots, y_K]$ ，有以下公式成立

$$y_i = w_i^T x$$

$$y = W^T x$$

为了像上节一样度量 $J(w)$ ，我们打算仍然从类间散列度和类内散列度来考虑。当样本是二维时，我们从几何意义上考虑：



其中 μ_i 和 S_w 与上节的意义一样， S_{w1} 是类别 1 里的样本点相对于该类中心点 μ_1 的散列程度。 S_{B1} 变成类别 1 中心点相对于样本中心点 μ 的协方差矩阵，即类 1 相对于 μ 的散列程度。

S_w 为

$$S_w = \sum_{i=1}^c S_{wi}$$

S_{wi} 的计算公式不变，仍然类似于类内部样本点的协方差矩阵

$$S_{wi} = \sum_{x \in \omega_i} (x - \mu_i)(x - \mu_i)^T$$

S_B 需要变，原来度量的是两个均值点的散列情况，现在度量的是每类均值点相对于样本中心的散列情况。类似于将 μ_i 看作样本点， μ 是均值的协方差矩阵，如果某类里面的样本点较多，那么其权重稍大，权重用 N_i/N 表示，但由于 $J(w)$ 对倍数不敏感，因此使用 N_i 。

$$S_B = \sum_{i=1}^c N_i (\mu_i - \mu)(\mu_i - \mu)^T$$

其中

$$\mu = \frac{1}{N} \sum_{\forall x} x = \frac{1}{N} \sum_{x \in \omega_i} N_i \mu_i$$

μ 是所有样本的均值。

上面讨论的都是投影前的公式变化，但真正的 $J(w)$ 的分子分母都是在投影后计算的。下面我们看样本点投影后的公式改变：

这两个是第 i 类样本点在某基向量上投影后的均值计算公式。

$$\tilde{\mu}_i = \frac{1}{N_i} \sum_{y \in \omega_i} y$$

$$\tilde{\mu} = \frac{1}{N} \sum_{\forall y} y$$

下面两个是在某基向量上投影后的 S_w 和 S_B

$$\widetilde{S}_w = \sum_{i=1}^c \sum_{y \in \omega_i} (y - \tilde{\mu}_i)(y - \tilde{\mu}_i)^T$$

$$\widetilde{S}_B = \sum_{i=1}^c N_i (\tilde{\mu}_i - \tilde{\mu})(\tilde{\mu}_i - \tilde{\mu})^T$$

其实就是将 μ 换成了 $\tilde{\mu}$ 。

综合各个投影向量（ w ）上的 \widetilde{S}_w 和 \widetilde{S}_B ，更新这两个参数，得到

$$\begin{aligned}\widetilde{S}_w &= W^T S_w W \\ \widetilde{S}_B &= W^T S_B W\end{aligned}$$

W 是基向量矩阵， \widetilde{S}_w 是投影后的各个类内部的散列矩阵之和， \widetilde{S}_B 是投影后各个类中心相对于全样本中心投影的散列矩阵之和。

回想我们上节的公式 $J(w)$ ，分子是两类中心距，分母是每个类自己的散列度。现在投影方向是多维了（好几条直线），分子需要做一些改变，我们不是求两两样本中心距之和（这个对描述类别间的分散程度没有用），而是求每类中心相对于全样本中心的散列度之和。

然而，最后的 $J(w)$ 的形式是

$$J(w) = \frac{|\widetilde{S}_B|}{|\widetilde{S}_w|} = \frac{|W^T S_B W|}{|W^T S_w W|}$$

由于我们得到的分子分母都是散列矩阵，要将矩阵变成实数，需要取行列式。又因为行列式的值实际上是矩阵特征值的积，一个特征值可以表示在该特征向量上的发散程度。因此我们使用行列式来计算（此处我感觉有点牵强，道理不是那么有说服力）。

整个问题又回归为求 $J(w)$ 的最大值了，我们固定分母为 1，然后求导，得出最后结果（我翻查了很多讲义和文章，没有找到求导的过程）

$$S_B w_i = \lambda S_w w_i$$

与上节得出的结论一样

$$S_w^{-1} S_B w_i = \lambda w_i$$

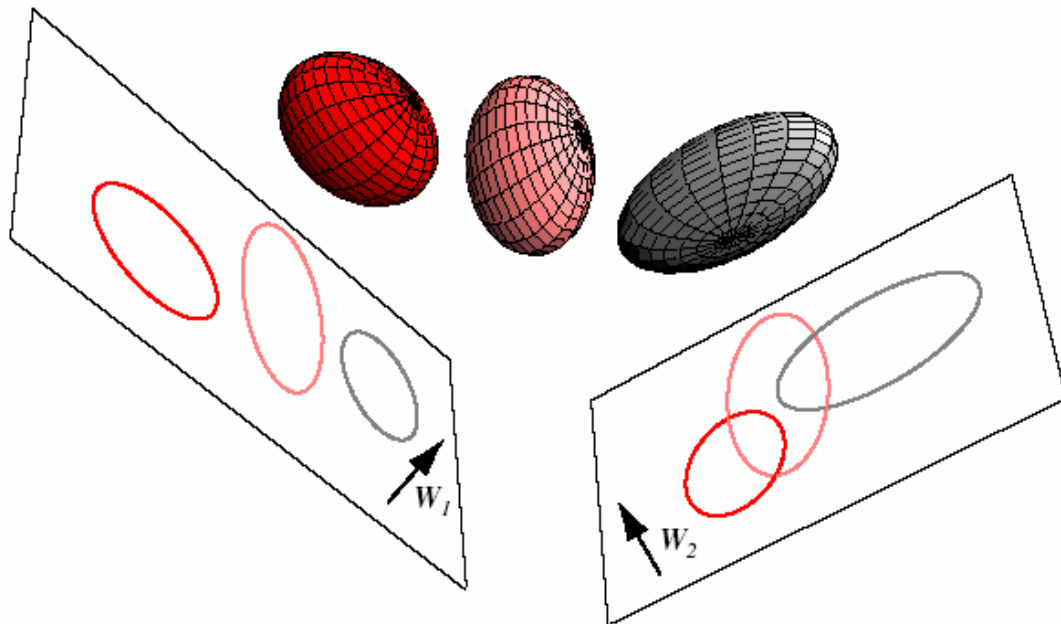
最后还归结到了求矩阵的特征值上来了。首先求出 $S_w^{-1} S_B$ 的特征值，然后取前 K 个特征向量组成 W 矩阵即可。

注意：由于 S_B 中的 $(\mu_i - \mu)$ 秩为 1，因此 S_B 的秩至多为 C （矩阵的秩小于等于各个相加矩阵的秩的和）。由于知道了前 $C-1$ 个 μ_i 后，最后一个 μ_C 可以有前面的 μ_i 来线性表示，因此 S_B 的秩至多为 $C-1$ 。那么 K 最大为 $C-1$ ，即特征向量最多有 $C-1$ 个。特征值大的对应的特征向量分割性能最好。

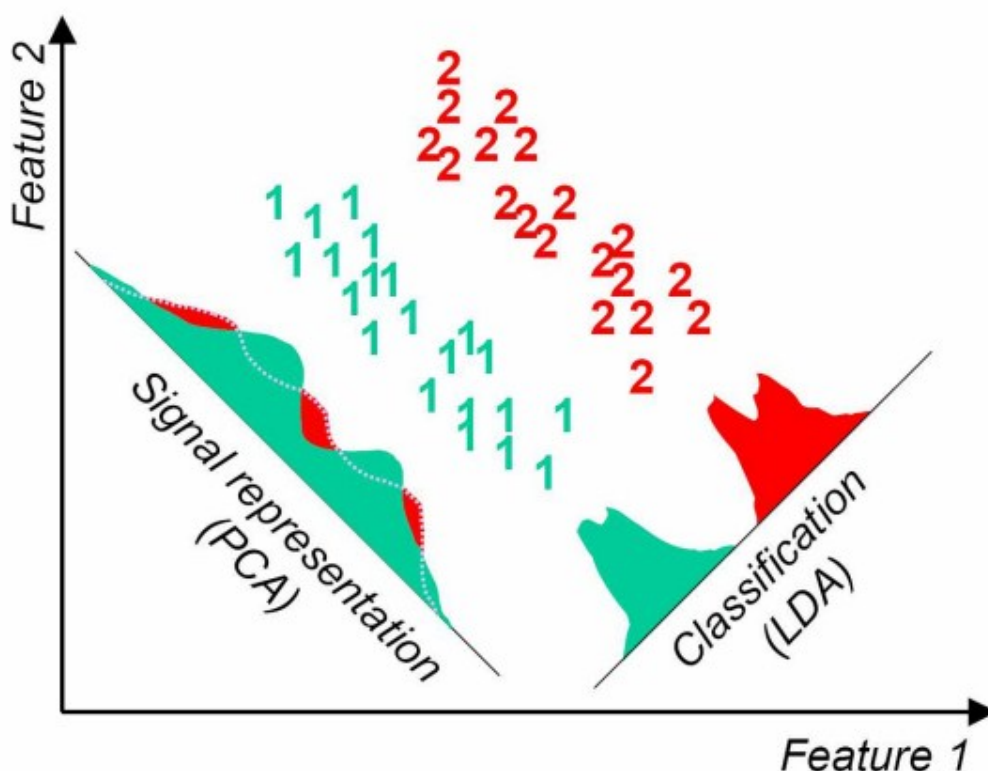
由于 $S_w^{-1}S_B$ 不一定是对称阵，因此得到的 K 个特征向量不一定正交，这也是与 PCA 不同的地方。

4. 实例

将 3 维空间上的球体样本点投影到二维上， w_1 相比 w_2 能够获得更好的分离效果。



PCA 与 LDA 的降维对比：



PCA 选择样本点投影具有最大方差的方向，LDA 选择分类性能最好的方向。

LDA 既然叫做线性判别分析，应该具有一定的预测功能，比如新来一个样例 x ，如何确定其类别？

拿二值分类来说，我们可以将其投影到直线上，得到 y ，然后看看 y 是否在超过某个阈值 y_0 ，超过是某一类，否则是另一类。而怎么寻找这个 y_0 呢？

看

$$y = w^T x$$

根据中心极限定理，独立同分布的随机变量和符合高斯分布，然后利用极大似然估计求

$$P(y|C_i)$$

然后用决策理论里的公式来寻找最佳的 y_0 ，详情请参阅 PRML。

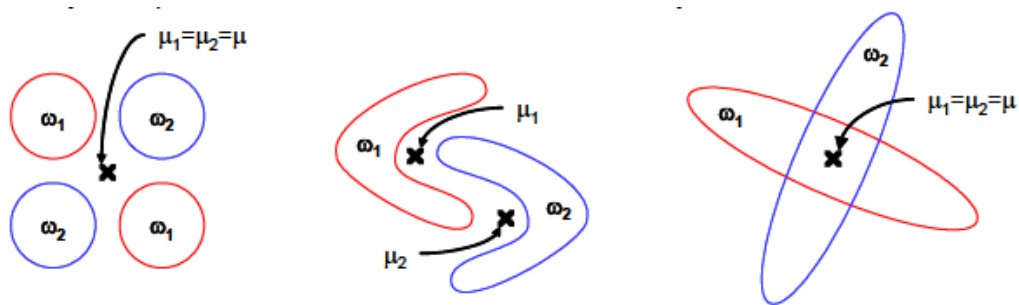
这是一种可行但比较繁琐的选取方法，可以看第 7 节（一些问题）来得到简单的答案。

5. 使用 LDA 的一些限制

1、LDA 至多可生成 $C-1$ 维子空间

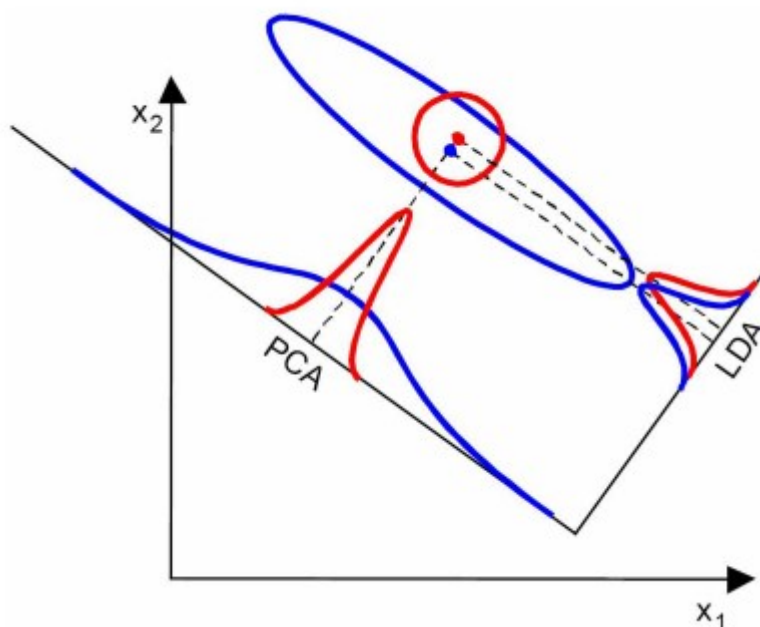
LDA 降维后的维度区间在 $[1, C-1]$ ，与原始特征数 n 无关，对于二值分类，最多投影到 1 维。

2、LDA 不适合对非高斯分布样本进行降维。



上图中红色区域表示一类样本，蓝色区域表示另一类，由于是 2 类，所以最多投影到 1 维上。不管在直线上怎么投影，都难使红色点和蓝色点内部凝聚，类间分离。

3、LDA 在样本分类信息依赖方差而不是均值时，效果不好。



上图中，样本点依靠方差信息进行分类，而不是均值信息。LDA 不能够进行有效分类，因为 LDA 过度依靠均值信息。

4、LDA 可能过度拟合数据。

6. LDA 的一些变种

1、非参数 LDA

非参数 LDA 使用本地信息和 K 临近样本点来计算 S_B , 使得 S_B 是全秩的，这样我们可以抽取多余 $C-1$ 个特征向量。而且投影后分离效果更好。

2、正交 LDA

先找到最佳的特征向量，然后找与这个特征向量正交且最大化 fisher 条件的向量。这种方法也能摆脱 $C-1$ 的限制。

3、一般化 LDA

引入了贝叶斯风险等理论

4、核函数 LDA

将特征 $x \rightarrow \Phi(x)$ ，使用核函数来计算。

7. 一些问题

上面在多值分类中使用的

$$S_B = \sum_{i=1}^C N_i (\mu_i - \mu)(\mu_i - \mu)^T$$

是带权重的各类样本中心到全样本中心的散列矩阵。如果 $C=2$ （也就是二值分类时）套用这个公式，不能够得出在二值分类中使用的 S_B 。

$$S_B = \sum_{i=1}^C (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

因此二值分类和多值分类时求得的 S_B 会不同，而 S_W 意义是一致的。

对于二值分类问题，令人惊奇的是最小二乘法和 Fisher 线性判别分析是一致的。

下面我们证明这个结论，并且给出第 4 节提出的 y_0 值得选取问题。

回顾之前的线性回归，给定 N 个 d 维特征的训练样例 $x^{(i)}\{x_1^{(i)}, x_2^{(i)}, \dots, x_d^{(i)}\}$ (i 从 1 到 N)，每个 $x^{(i)}$ 对应一个类标签 $y^{(i)}$ 。我们之前令 $y=0$ 表示一类， $y=1$ 表示另一类，现在我们为了证明最小二乘法和 LDA 的关系，我们需要做一些改变

$$\begin{cases} y = \frac{N}{N_1}, \text{ 样例属于有 } N_1 \text{ 个元素的类 } C_1 \\ y = -\frac{N}{N_2}, \text{ 样例属于有 } N_2 \text{ 个元素的类 } C_2 \end{cases}$$

就是将 0/1 做了值替换。

我们列出最小二乘法公式

$$E = \frac{1}{2} \sum_{i=1}^N (w^T x^{(i)} + w_0 - y^{(i)})^2$$

w 和 w_0 是拟合权重参数。

分别对 w_0 和 w 求导得

$$\begin{aligned} \sum_{i=1}^N (w^T x^{(i)} + w_0 - y^{(i)}) &= 0 \\ \sum_{i=1}^N (w^T x^{(i)} + w_0 - y^{(i)}) x^{(i)} &= 0 \end{aligned}$$

从第一个式子展开可以得到

$$w^T N\mu + Nw_0 - \sum_{i=1}^N y^{(i)} = w^T N\mu + Nw_0 - \left(N_1 \frac{N}{N_1} - N_2 \frac{N}{N_2}\right) = 0$$

消元后，得

$$w_0 = -w^T \mu$$

$$\mu = \frac{1}{N} \sum_{i=1}^N x^{(i)} = \frac{1}{N} (N_1 \mu_1 + N_2 \mu_2)$$

可以证明第二个式子展开后和下面的公式等价

$$\left(S_w + \frac{N_1 N_2}{N} S_B\right) w = N(\mu_1 - \mu_2)$$

其中 S_w 和 S_B 与二值分类中的公式一样。

由于 $S_B w = (\mu_1 - \mu_2) * \lambda_w$

因此，最后结果仍然是

$$w = S_w^{-1}(\mu_1 - \mu_2)$$

这个过程从几何意义上去理解也就是变形后的线性回归（将类标签重新定义），线性回归后的直线方向就是二值分类中 LDA 求得的直线方向 w 。

好了，我们从改变后的 y 的定义可以看出 $y > 0$ 属于类 C_1 ， $y < 0$ 属于类 C_2 。因此我们可以选取 $y_0 = 0$ ，即如果 $y(x) = w^T x + w_0 > 0$ ，就是类 C_1 ，否则是类 C_2 。

写了好多，挺杂的，还有个 topic 模型也叫做 LDA，不过名字叫做 Latent Dirichlet Allocation，第二作者就是 Andrew Ng 大牛，最后一个他导师 Jordan 泰斗了，什么时候拜读后再写篇总结发上来吧。