# 问题

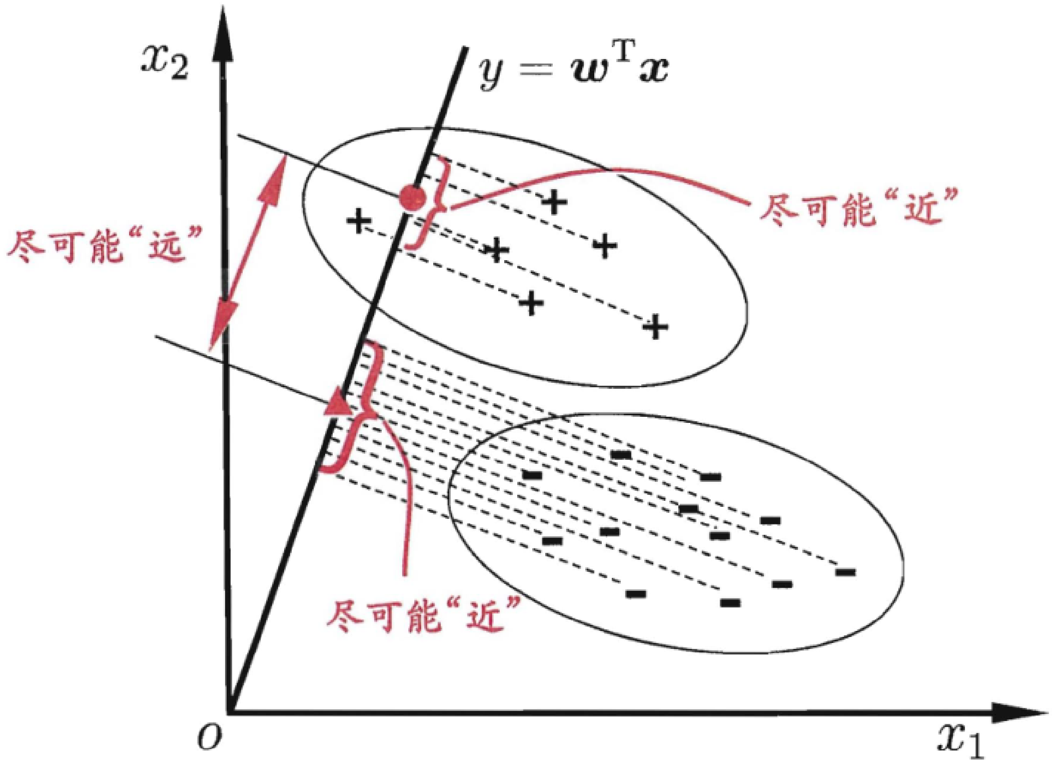
线性判别分析（Linear Discriminant Analysis，LDA）是机器学习中常用的降维方法之一，本文旨在介绍LDA算法的思想，其数学推导过程可能会稍作简化。

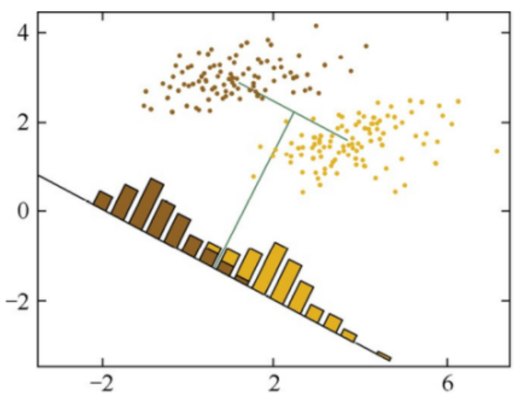
# LDA的思想

● LDA是一种线性的、有监督的降维方法，即每个样本都有对应的类别标签（这点和PCA）。

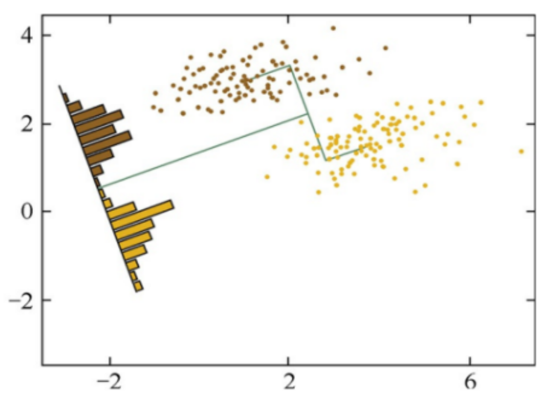
● 主要思想：给定训练样本集，设法将样本投影到一条直线上，使得同类的样本的投影尽可能的接近、异类样本的投影尽可能地远离（即**最小化类内距离和最大化类间距离**）。

> 下面分别通过《机器学习》和《百面机器学习》两本书中的图片先来直观地理解一下LDA的思想。



（a）最大化两类投影中心距离准则下得到的分类结果



（b）使得投影后样本区分性更高的投影方式

● 为什么要将最大化类间距离和最小化类内距离同时作为优化目标呢？

先看上面第二张图的左图（a），对于两个类别，只采用了最大化类间距离，其结果中两类样本会有少许重叠；而对于右图（b），同时最大化类间距离和最小化类内距离，可见分类效果更好，同类样本的投影分布更加集中了。当然，对于二维的数据，可以采用将样本投影到直线上的方式，对于高维的数据，则是投影到一个低维的超平面上，这应该很好理解。

## LDA算法优化目标

由上面的介绍我们知道，LDA算法的思想就是最大化类间距离和最小化类内距离，其优化目标就很直观了，那怎么用数学方式来表示呢？要解决这个问题，就得先看看怎么描述类间距离和类内距离。

- **类间距离**（以二分类为示例）

假设有$C_1$、$C_2$两类样本，其均值分别为 $\mu_1 = \frac{1}{N}\sum_{x \in C_1} x$ 和 $\mu_2 = \frac{1}{N}\sum_{x \in C_2} x$ 。很显然，要使得两类样本类间距离最大，则 $\mu_1$ 、$\mu_2$ 的距离应尽可能地大，则类间距离可描述为

$$||\omega^T \mu_0 - \omega^T \mu_1||_2^2, \text{其中，}\omega\text{为投影方向} \tag{1}$$

- **类内距离**

要使得样本在同类中距离最小，也就是最小化同类样本的方差，假设分别用 $D_1$、 $D_2$ 表示两类样本的投影方差，则有：

$$\begin{aligned} D_1 &= \sum_{x \in C_1} (\omega^T x - \omega^T \mu_1)^2 = \sum_{x \in C_1} \omega^T (x - \mu_1)(x - \mu_1)^T \omega \\ D_2 &= \sum_{x \in C_2} (\omega^T x - \omega^T \mu_2)^2 = \sum_{x \in C_2} \omega^T (x - \mu_2)(x - \mu_2)^T \omega \end{aligned} \tag{2}$$

因此，要使得类内距离最小，就是要最小化 $D_1 + D_2$。

- **优化目标**

由上面分析，最大化类间距离和最小化类内距离，因此可以得到最大化目标：

$$\begin{aligned} J(\omega) &= \frac{||\omega^T \mu_0 - \omega^T \mu_1||_2^2}{D_1 + D_2} \\ &= \frac{||\omega^T \mu_0 - \omega^T \mu_1||_2^2}{\sum_{x \in C_i} \omega^T (x - \mu_i)(x - \mu_i)^T \omega} \end{aligned} \tag{3}$$

为了化简上面公式，给出几个定义：

- **类间散度矩阵：**

$$S_b = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \tag{4}$$

- **类内散度矩阵：**

$$S_\omega = \Sigma_1 + \Sigma_2 = \sum_{x \in C_1} (x - \mu_1)(x - \mu_1)^T + \sum_{x \in C_2} (x - \mu_2)(x - \mu_2)^T \tag{5}$$

因此最大化目标可以简写为：

$$J(\omega) = \frac{\omega^T S_b \omega}{\omega^T S_\omega \omega} \tag{6}$$

> 这是一个广义瑞利商，可以对矩阵进行标准化操作（具体证明就不展开啦），因此，通过标准化后总可以得到 $\omega^T S_\omega \omega = 1$，又由于上面优化目标函数分子分母都是二次项，其解与 $\omega$ 的长度无关，只与方向有关，因此上面优化目标等价于以下最小化目标：

转化为最小化目标：

$$
\begin{aligned}
\underset{\omega}{min} \quad & -\omega^T S_b \omega \\
s.t. \quad & \omega^T S_\omega \omega = 1
\end{aligned}
\tag{7}
$$

由拉格朗日法，上式可得：

$$
\begin{aligned}
& S_b \omega = \lambda S_\omega \omega \\
& \text{即有，} S_\omega^{-1} S_b \omega = \lambda \omega
\end{aligned}
\tag{8}
$$

至此，我们的**优化目标就转化成了求矩阵 $S_\omega^{-1} S_b$ 的特征值，而投影方向就是这个特征值对应的特征向量**。

> 由于 $(\mu_1 - \mu_2)^T \omega$ 是个标量（因为 $\mu_1 - \mu_2$ 和 $\omega$ 同向时才能保证类间距离最大），
>
> 所以，对于 $S_b \omega = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \omega$ 而言，可以看出 $S_b \omega$ 始终与 $(\mu_1 - \mu_2)$ 的方向一致

因此，如果只考虑 $\omega$ 的长度而不考虑方向，则由：

$$
S_\omega^{-1} S_b \omega = \lambda \omega \qquad => \qquad \omega = S_\omega^{-1}(\mu_1 - \mu_2) \tag{9}
$$

**也就是说，我们只需求出样本的均值和类内的散度矩阵（即类内方差），即可求出投影方向。**

# LDA算法流程(推广至高维)

1.计算每类样本的均值向量 $\mu_i$。

2.计算类间散度矩阵 $S_\omega$ 和类内散度矩阵 $S_b$ 。

3.求矩阵 $S_\omega^{-1} S_b$ 的特征值即对应的特征向量，从大到小排序。

4.将特征值由大到小排列，取出前 k 个特征值对应的特征向量。

5.将 n 维样本映射到 k 维，实现降维处理。

$$
x_i^{'} = \begin{bmatrix} \omega_1^T x_i \\ \omega_2^T x_i \\ \vdots \\ \omega_k^T x_i \end{bmatrix} \tag{10}
$$

# 总结

● LDA是线性的、有监督的降维方法，其优点是善于对有类别信息的数据进行降维处理（与PCA的不同）。

● LDA因为是线性模型，对噪声的鲁棒性较好，但由于模型简单，对数据特征的表达能力不足。

● LDA对数据的分布做了一些很强的假设，比如每个类别都是高斯分布、各个类别的协方差相等，实际中这些假设很难完全满足。

> 关于LDA与PCA的区别，请看下回分解。

# 参考资料

周志华《机器学习》