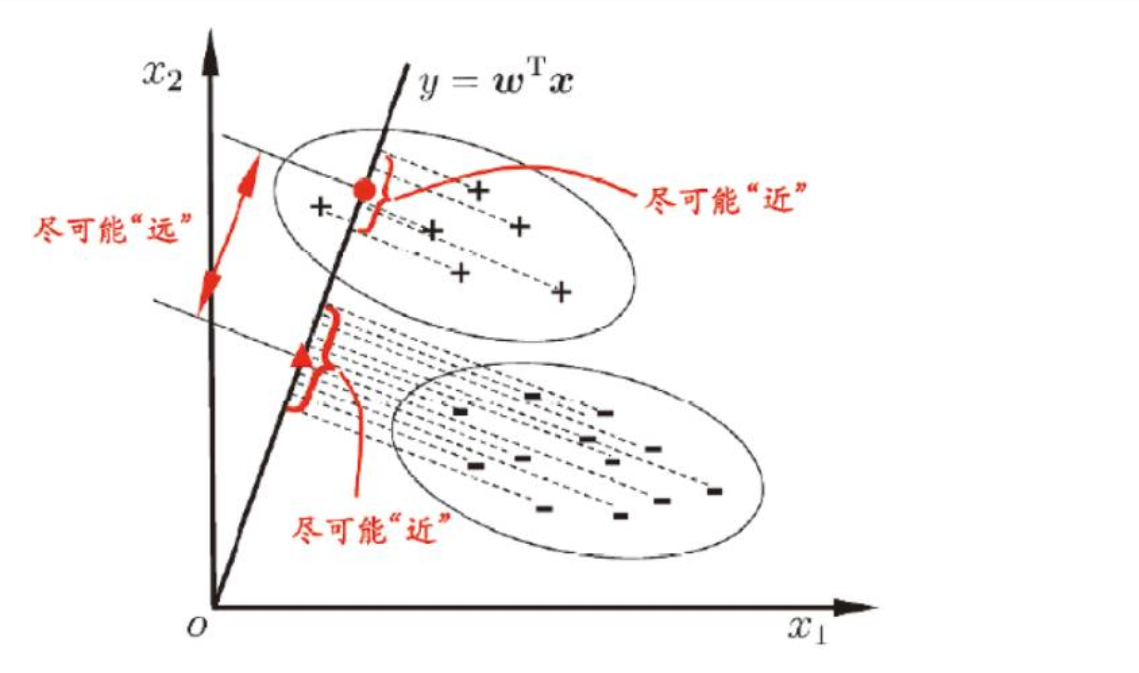
**线性判别分析**

是一种经典的降维方法，线性判别分析简称LDA。

LDA是一种特征降维算法，遇到特征维数过大而导致过拟合的问题，可以通过LDA算法降低特征维度，解决过拟合的问题。

LDA是一种监督学习的降维技术，它的数据集的每个样本是有类别输出的，这点和PCA不同。PCA是不考虑样本类别输出的无监督降维技术。

LDA的思想可以用一句话概括，就是“投影后类内方差最小，类间方差最大”，如下图所示。



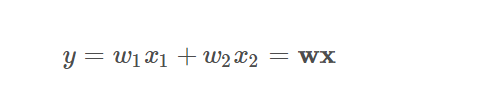
**公式的推导：**

假如我们有2种数据集

X1={x11,x21,...,xn1}

X2={x12,x22,...,xn2}

我们希望有一条直线



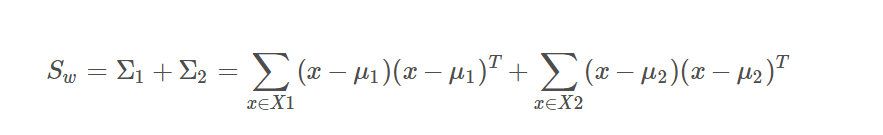
使得将两种数据集中的样本带入后可以获得比较明显的分类效果。

例如将X1数据集中的样本带入函数后y的值范围在[-1,0],而将X2数据集中的样本带入函数后y的值范围在[1,2]，这样就可以很轻松的进行分类了。

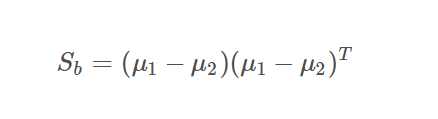
那么为了达到这样的效果，首先要让同类的样本投影点尽可能小,即期望投影后同类样本数据方差IMG_257小。

然后我们还要让两个类的中心IMG_258投影后的距离尽量大，IMG_259要尽量大。

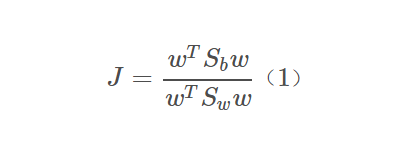
定义“类内散度矩阵”



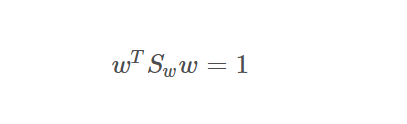
定义“类间散度矩阵”



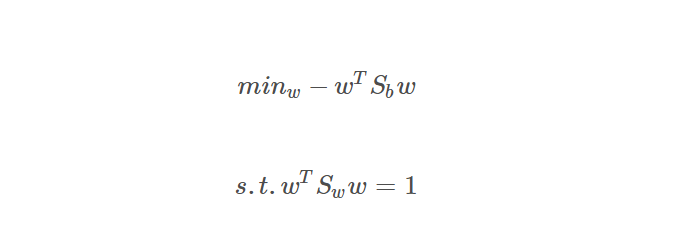
则我们的代价函数可写为：



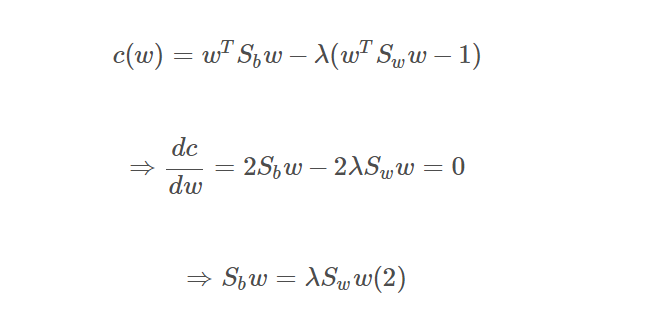
我们希望最大化上述代价函数，由于代价函数的分子分母都有2个w，因此J的最大解与w的长度无关，只与方向有关，令



则（1）式等价于

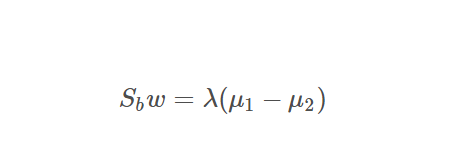


加入拉格朗日乘子并求导求得下式：

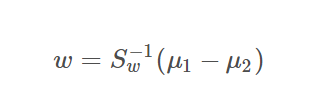


由于IMG_260的方向恒为IMG_261

可以令



带入（2）式可得

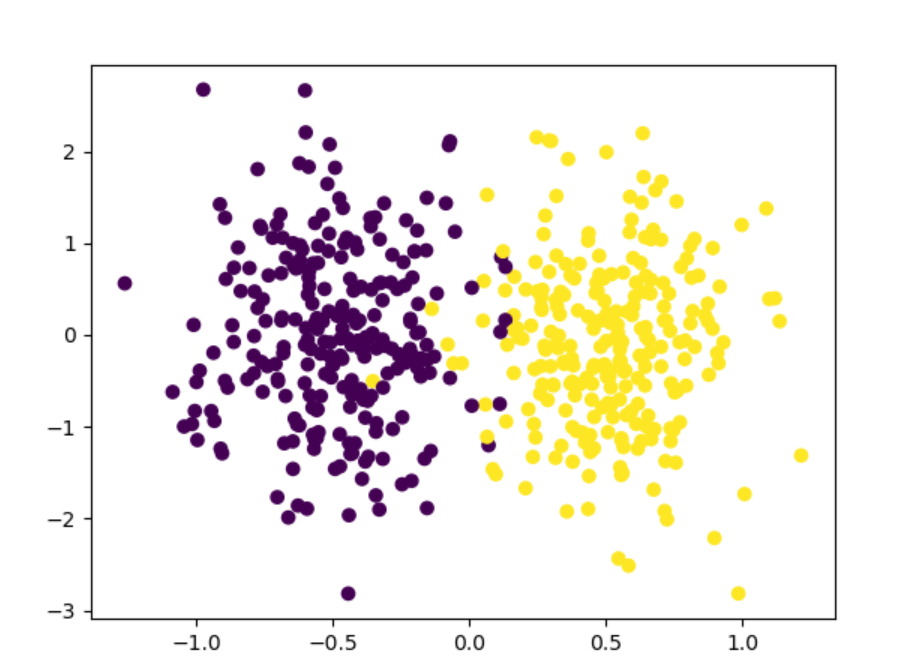


**LDA.py**

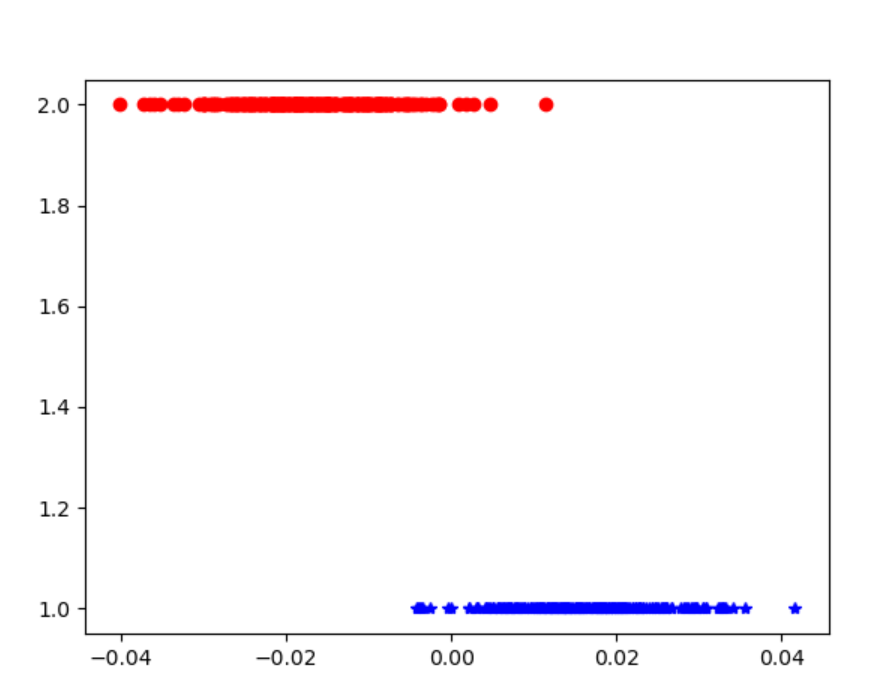
**二分类问题**

代码输出如下：

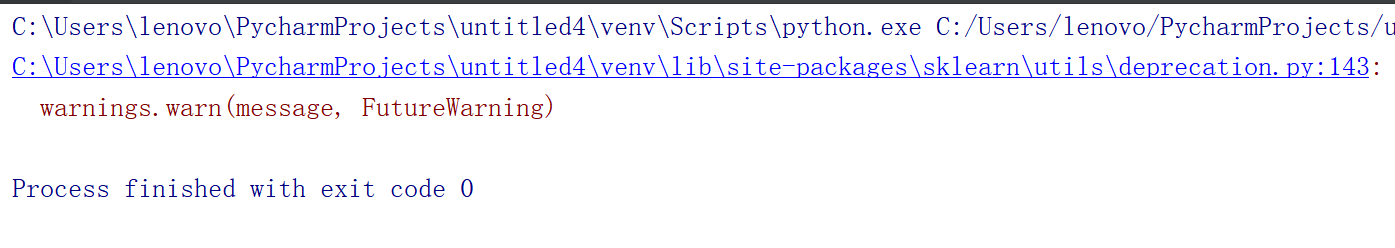
分类前的数据：



分类后的数据：



但是代码报错了：



FutureWarning: The sklearn.datasets.samples\_generator module is deprecated in version 0.22 and will be removed in version 0.24. The corresponding classes / functions should instead be imported from sklearn.datasets. Anything that cannot be imported from sklearn.datasets is now part of the private API.warnings.warn(message, FutureWarning)

百度以后，改动了

from sklearn.datasets.samples\_generator import make\_classification

改成了

from sklearn.datasets import make\_classification

就不报错了。代码正常运行了

**生成随机数据**

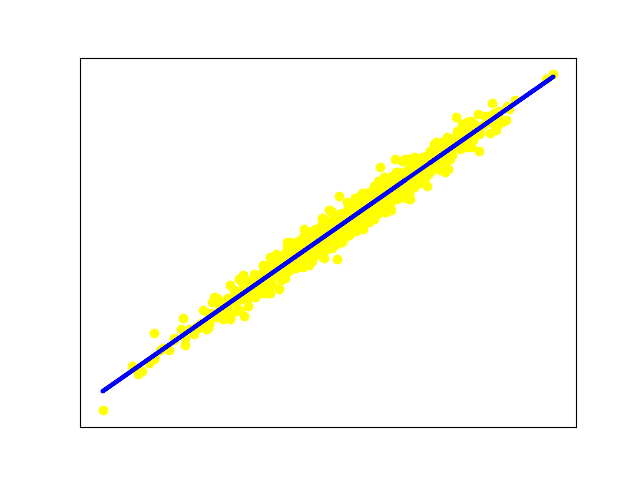
**LDA-3.py**

scikit-learn生成数据样本。

1）回归模型随机数据  
使用make\_regression生成回归模型数据。

几个关键参数有n\_samples（生成样本数）， n\_features（样本特征数），noise（样本随机噪音）和coef（是否返回回归系数）。

代码输出如下：



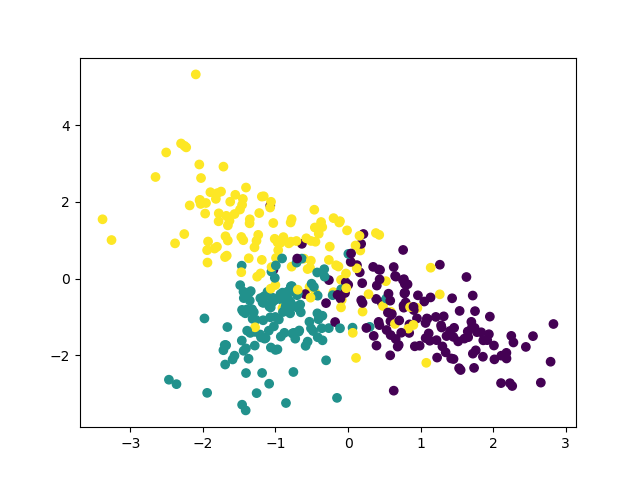
**LDA-4.py**

2)分类模型随机数据

用make\_classification生成三元分类模型数据。

几个关键参数有n\_samples（生成样本数）， n\_features（样本特征数）， n\_redundant（冗余特征数）和n\_classes（输出的类别数）

代码输出如下：



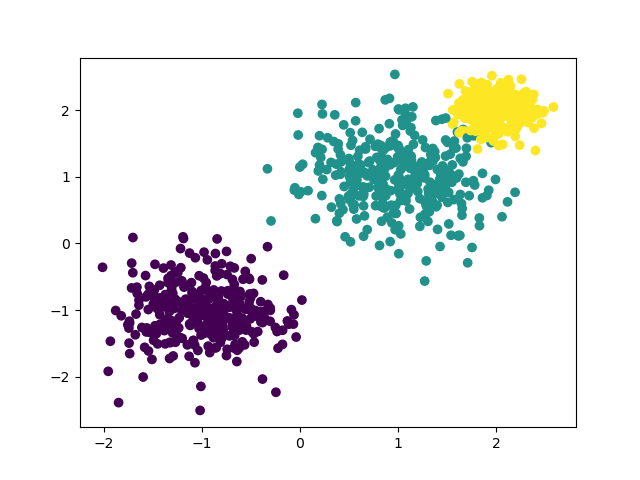
**LDA-5.py**

1. 聚类模型随机数据

用make\_blobs生成聚类模型数据。

几个关键参数有n\_samples（生成样本数）， n\_features（样本特征数），centers(簇中心的个数或者自定义的簇中心)和cluster\_std（簇数据方差，代表簇的聚合程度）。

代码输出如下：



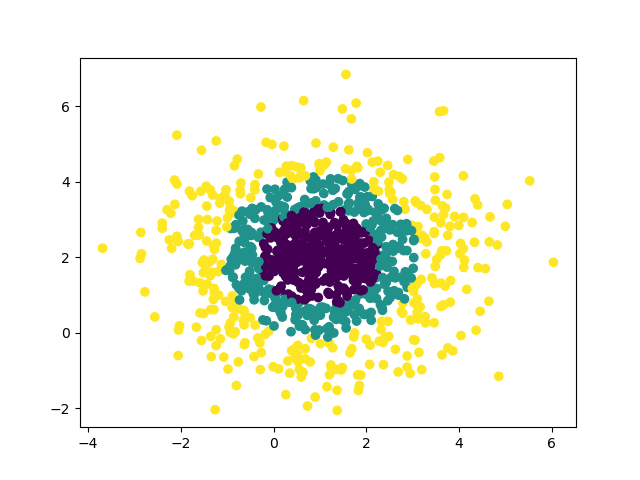
**LDA-6.py**

1. 分组正态分布混合数据

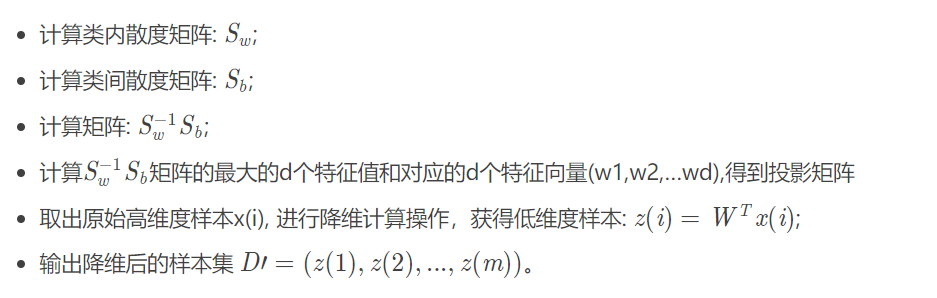
用make\_gaussian\_quantiles生成分组多维正态分布的数据。

几个关键参数有n\_samples（生成样本数）， n\_features（正态分布的维数），mean（特征均值）， cov（样本协方差的系数）， n\_classes（数据在正态分布中按分位数分配的组数）。

代码输出如下：

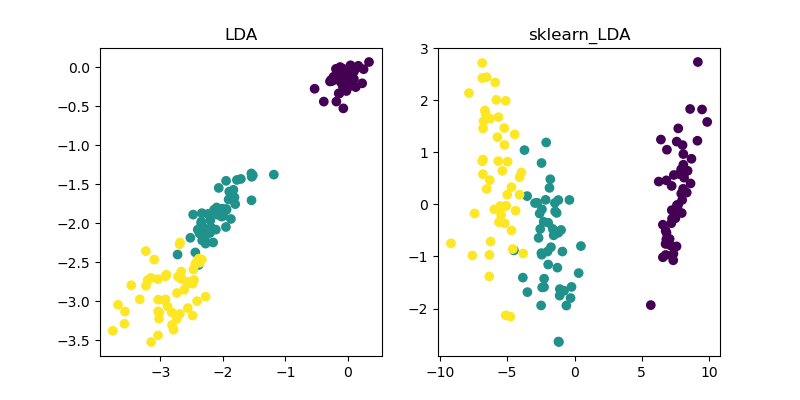
****

**LDA算法计算步骤：**



**LDA-2.py**

代码输出如下：



使用了自己编程实现的LDA和调用sklearn自带的LDA对iris数据进行LDA，

PCA追求的是在降维之后能够最大化保持数据的内在信息，并通过衡量在投影方向上的数据方差的大小来衡量该方向的重要性。

**PCA优缺点：**

优点：

1. 最小误差
2. 提取了主要信息

缺点：

PCA将所有的样本（特征向量集合）作为一个整体对待，去寻找一个均方误差最小意义下的最优线性映射投影，而忽略了类别属性，而它所忽略的投影方向有可能刚好包含了重要的可分性信息。

LDA所追求的目标和PCA不同，不是希望保持数据最多的信息，而是希望数据在降维后能够很容易地被区分开来。

**LDA与PCA**

相同点

1）两者均可以对数据进行降维。

2）两者在降维时均使用了矩阵特征分解的思想。

3）两者都假设数据符合高斯分布。

不同点

1）LDA是有监督的降维方法，而PCA是无监督的降维方法

2）LDA降维最多降到类别数k-1的维数，而PCA没有这个限制。

3）LDA除了可以用于降维，还可以用于分类。

4）LDA选择分类性能最好的投影方向，而PCA选择样本点投影具有最大方差的方向。这点可以从下图形象的看出，在某些数据分布下LDA比PCA降维较优。

**LDA小结**

LDA算法既可以用来降维，又可以用来分类，但是目前来说，主要还是用于降维。

优点

1）在降维过程中可以使用类别的先验知识经验，而像PCA这样的无监督学习则无法使用类别先验知识。

2）LDA在样本分类信息依赖均值而不是方差的时候，比PCA之类的算法较优。

缺点

1）LDA不适合对非高斯分布样本进行降维，PCA也有这个问题。

2）LDA降维最多降到类别数k-1的维数，如果我们降维的维度大于k-1，则不能使用LDA。当然目前有一些LDA的进化版算法可以绕过这个问题。

3）LDA在样本分类信息依赖方差而不是均值的时候，降维效果不好。

4）LDA可能过度拟合数据。