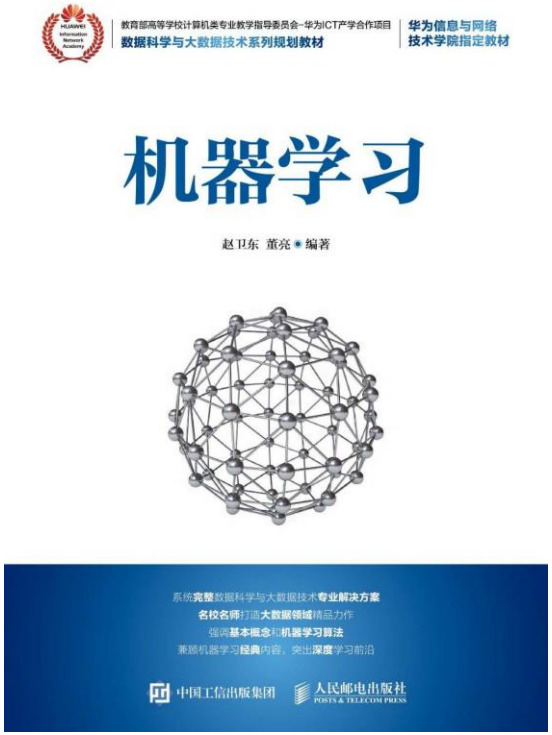


机器学习
PCA和LDA

复旦大学 **赵卫东** 博士

wdzhao@fudan.edu.cn

高维数据降维

- 机器学习领域中的降维就是指采用某种映射方法，将原高维空间中的数据点映射到低维度的空间中。在原始的高维空间中，包含有冗余信息以及噪声信息。图像识别中如果噪声太多会造成误差，降低识别准确率；通过降维，可以减少冗余信息所造成的误差，提高识别的精度。此外，通过降维可以寻找数据内部的本质结构特征
- 降维的本质是学习一个映射函数$f: x->y$，其中x是原始数据点的表达，目前最多使用向量表达形式。y是数据点映射后的低维向量表达，通常y的维度小于x的维度。y可能是显式的或隐式的、线性的或非线性的函数。目前大部分降维算法处理向量表达的数据

主成分分析

- 主成分分析是最常用的线性降维方法，它的目标是通过某种线性投影，将高维的数据映射到低维的空间中，并期望在所投影的维度上数据的方差最大，以此使用较少的维度，同时保留较多原数据的维度
- 尽可能如果把所有的点都映射到一起，那么几乎所有的区分信息都丢失了，而如果映射后方差尽可能的大，那么数据点则会分散开来，特征更加明显。PCA是丢失原始数据信息最少的一种线性降维方法，最接近原始数据
- PCA算法目标是求出样本数据的协方差矩阵的特征值和特征向量，而协方差矩阵的特征向量的方向就是PCA需要投影的方向。使样本数据向低维投影后，能尽可能表征原始的数据。协方差矩阵可以用散布矩阵代替，协方差矩阵乘以（n-1）就是散布矩阵，n为样本的数量。协方差矩阵和散布矩阵都是对称矩阵，主对角线是各个随机变量（各个维度）的方差

主成分分析

- 设有m条n维数据，PCA的一般步骤如下
  - 将原始数据按列组成n行m列矩阵X
  - 计算矩阵X中每个特征属性（n维）的平均向量M（平均值）
  - 将X的每行（代表一个属性字段）进行零均值化，即减去M
  - 按照公式$C = \frac{1}{\mathrm{m}} XX^T$求出协方差矩阵
  - 求出协方差矩阵的特征值及对应的特征向量
  - 将特征向量按对应特征值从大到小按行排列成矩阵，取前k（k < n）行组成基向量P
  - 通过$Y = PX$计算降维到k维后的样本特征

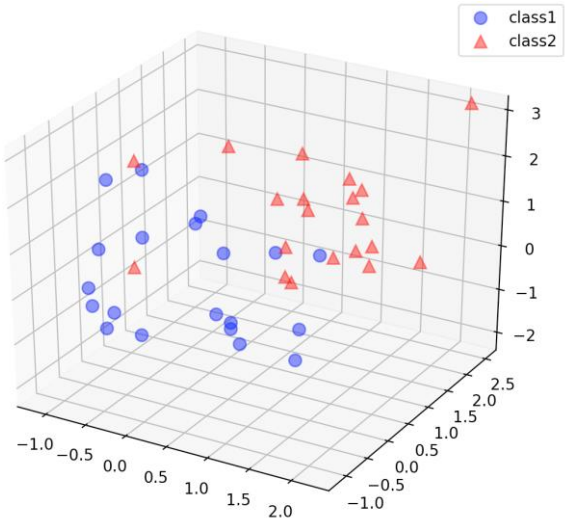
主成分分析

- 基于sklearn（Python语言下的机器学习库）和numpy随机生成2个类别共40个3维空间的样本点，生成的代码如下：

```
mu_vec1 = np.array([0,0,0])
cov_mat1 = np.array([[1,0,0],[0,1,0],[0,0,1]])
class1_sample = np.random.multivariate_normal(mu_vec1, cov_mat1, 20).T
mu_vec2 = np.array([1,1,1])
cov_mat2 = np.array([[1,0,0],[0,1,0],[0,0,1]])
class2_sample = np.random.multivariate_normal(mu_vec2, cov_mat2, 20).T
```

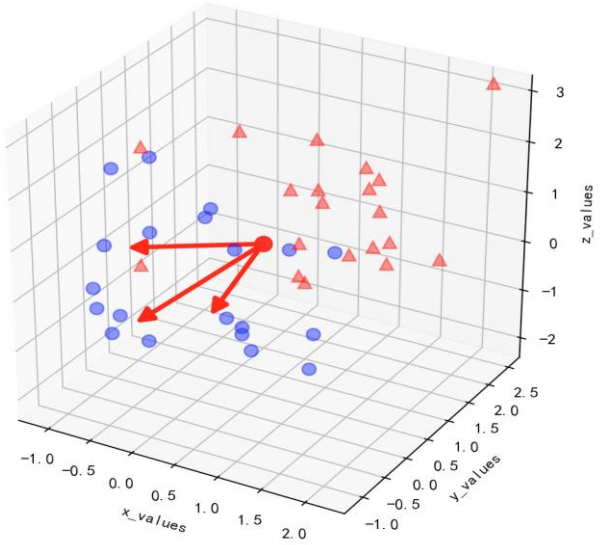
主成分分析

- 生成的两个类别class1_sample和class2_sample的样本数据维度为3维，即样本数据的特征数量为3个，将其置于3维空间中展示



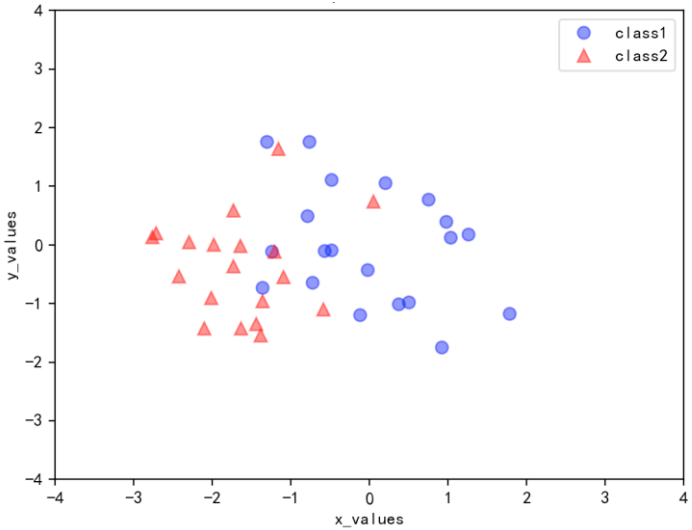
主成分分析

- 计算40个点在3个维度上的平均向量



主成分分析

- 二维空间分布

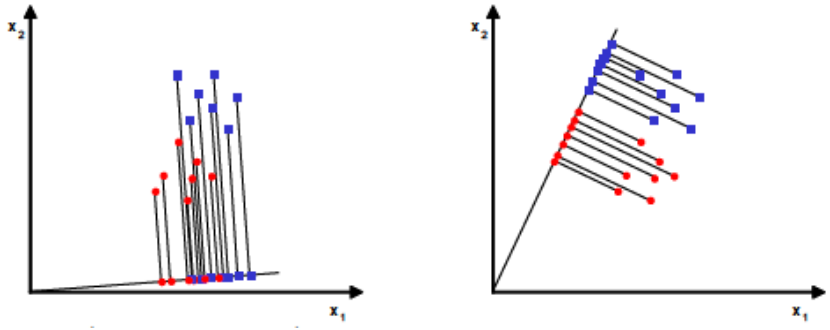


线性判别分析

- 线性判别分析LDA是一种有监督的线性降维算法。与PCA不同，LDA是为了使降维后的数据点尽可能地容易被区分
- 线性判别分析的原理是对于给定的训练集，设法将样本投影到一条直线上，使得同类的投影点尽可能接近，异类样本的投影点尽可能远离；在对新样本进行分类时，将其投影到这条直线上，再根据投影点的位置来确定新样本的类别。PCA主要是从特征的协方差角度，去找到比较好的投影方式。LDA更多地考虑了标注，即希望投影后不同类别之间数据点的距离更大，同一类别的数据点更紧凑

线性判别分析

- LDA的降维过程如下
  - 计算数据集中每个类别下所有样本的均值向量
  - 通过均值向量，计算类间散布矩阵S_B和类内散布矩阵S_W
  - 依据公式$S_W^{-1}S_BU = \lambda U$进行特征值求解，计算$S_W^{-1}S_B$的特征向量和特征值
  - 按照特征值排序， 选择前k个特征向量构成投影矩阵U
  - 通过$Y = X \times U$的特征值矩阵将所有样本转换到新的子空间中



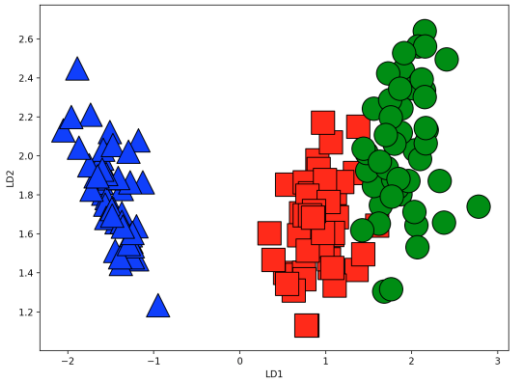
线性判别分析

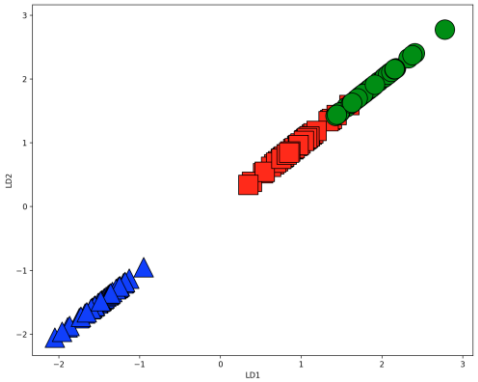
- 应用LDA技术对鸢尾花(Iris)的样本数据进行分析，鸢尾花数据集是20世纪30年代的经典数据集，它由Fisher收集整理，数据集包含150个数据集，分为3类，每类50个数据，每个数据包含4个属性。可通过花萼长度、花萼宽度、花瓣长度和花瓣宽度4个属性预测鸢尾花卉属于山鸢尾（Iris Setosa）、杂色鸢尾（Iris Versicolour）、维吉尼亚鸢尾（Iris Virginica）中的哪种类别，将类别文字转化为数字类别

序号	萼片长(cm)	萼片宽(cm)	花瓣长(cm)	花瓣宽(cm)	类别
145	6.7	3.0	5.2	2.3	2
146	6.3	2.5	5.0	1.9	2
147	6.5	3.0	5.2	2.0	2
148	6.2	3.4	5.4	2.3	2

线性判别分析

- 数据集中有4个特征，萼片长、萼片宽、花瓣长和花瓣宽，总共150行，每一行是一个样本，这就构成了一个4x150的输入矩阵，输出是1列，即花的类别，构成了1x150的矩阵。分析的目标就是通过LDA算法将输入矩阵映射到低维空间中进行分类






THANK
YOU