数据预处理

- import numpy as np
- from sklearn import preprocessing
-
- 均值移除　达到特征均值几乎为0，而且标准差为1
  - preprocessing.scale(data)
- 范围缩放　有时候每个特征数值范围可能变化很大，因为将特征的数值范围缩放到河里的大小是非常重要的
  - preprocessing.MinMaxScaler(feature_range=(0, 1))
- 归一化处理　保证每个特征向量的值都被缩放到相同的数值范围
  - preprocessing.normalize(data, norm='l1')
- 二值化　二值化用于将数值特征向量转换为布尔类型向量
  - preprocessing.Binarizer(threshold=1.4).transform(data)
- 独热编码　需要处理的数值都是稀疏地、散乱地分布在空间中，然而，我们并不需要存储这些大数值，这时就需要使用独热编码（One-Hot Encoding）。可以把独热编码看作是一种收紧（tighten）特征向量的工具。
  - preprocessing.OneHotEncoder()
  - 如果非重复计数的值是K，那么就把这个特征转换为只有一个值是1其他值都是0的K维向量，结果只能有一个是1，而且k是一列数据中，不相同数据的个数s