聚类项目：

1. 实验数据: UCI 的 iris 数据集
2. 用传统的 kmeans 算法对 iris 数据集进行聚类，将 K=3，随机执行 3 次，输出 3 次的结果，结果包括：
   输出 3 个簇中心，
   统计每个簇内的样本数
   比较三次的 SSE，比较三次的 purity
3. 用传统的 kmeans 算法对 iris 数据集进行聚类，将 K=2,3,4,5,6, 输出结果包括：
   输出每次的 SSE
   将多次的 SSE 画折线图
   选出最佳的 K 值

评价标准：purity 的定义

$$\text{purity}（\text{W}，\ \text{C}）=\frac{1}{\text{N}}\sum_{\text{k}}\max_{\text{j}}|w_k \cap c_j|$$

SSE：

$$J_e=\sum_{i=1}^{k}\sum_{x\in C_i}|x-m_i|^2$$

PCA 项目：

1.对 UCI 的 iris 的数据集进行 PCA 处理
2.输出 pca 的特征值(由大到小排序)，填写下表:累计方差用 np.cumsum(a)

序号	特征值	方差比	累计方差比
1			

3.画出方差比折线图,累计方差比折线图,
4.K=2 时，画出原始数据和重构数据散点图，画出数据在新特征空间散点图任意选择 2 个特征）

基于 Python 实现的 PCA：

- loadDataset(filename,delim='\t')
- cov(?)建立协方差矩阵
  - cov(X,0) = cov(X) 除数是 n-1(n 为样本个数),除数 n-1 是为了得到协方差的无偏估计
  - cov(X,1) 除数是 n

- linag.eig(?)求特征向量和特征值
- argsort :对特征值矩阵进行由大到小排序，返回对应排序后的索引排序