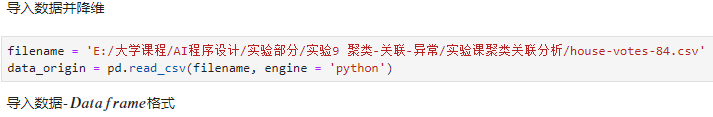
# 实验9\_任务1\_聚类

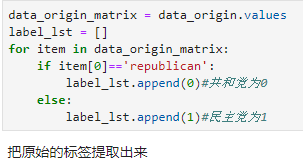
# 问题：

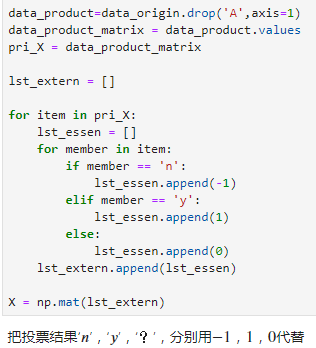
任务描述：数据中给定议员对于16项不同议题的投票情况（属性表示为⽀持Y、反对N、弃权？）,数据的第⼀列为议员所属的党派类别，请使⽤聚类⽅法对上述数据进⾏聚类分析：

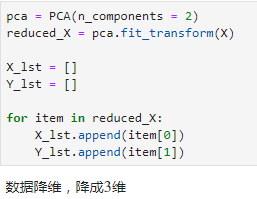
1. 尝试定量的评价的聚类结果；
2. 尝试⽐较不同聚类⽅法或者不同初始情况对聚类结果的影响；

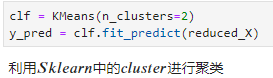
# 实验过程记录：

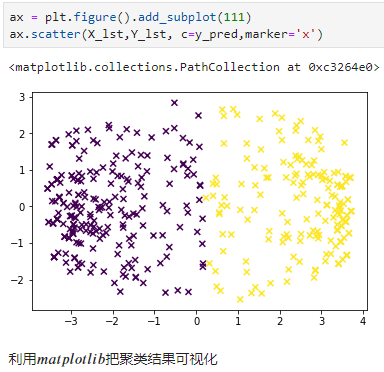


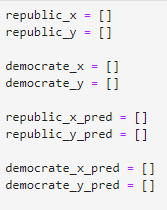


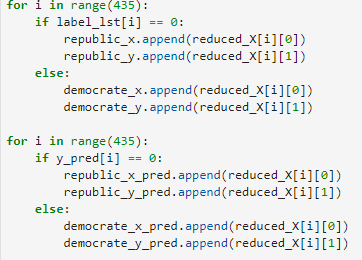


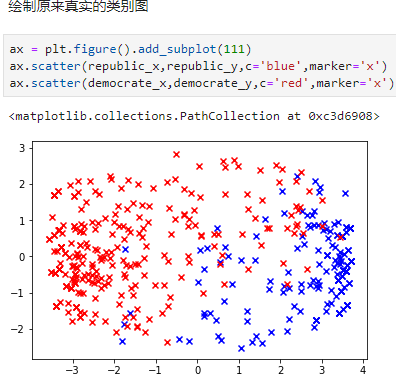


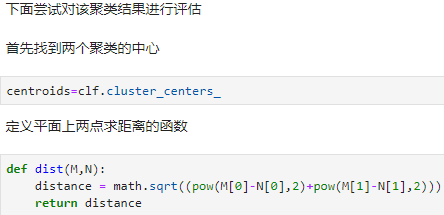


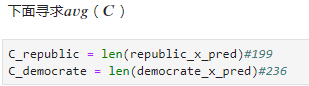




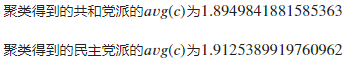


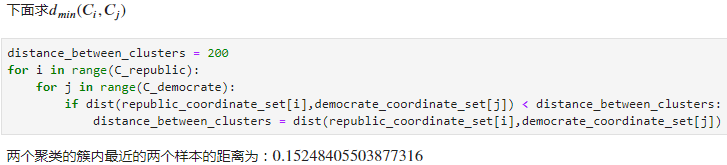


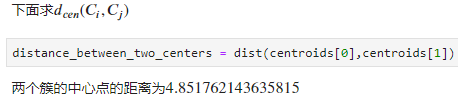


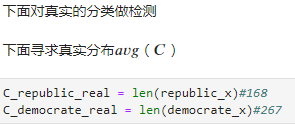




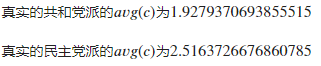


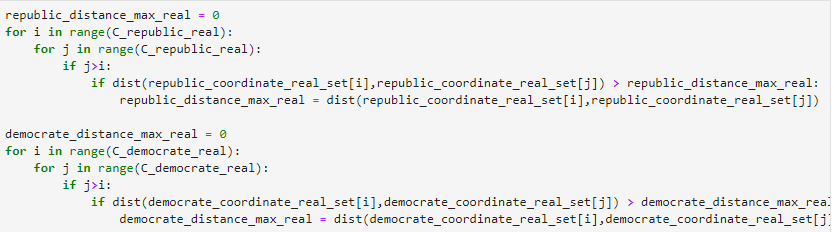


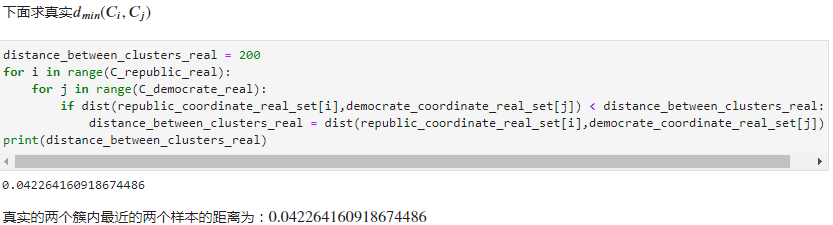










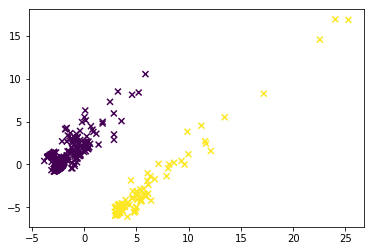


**问题1：**

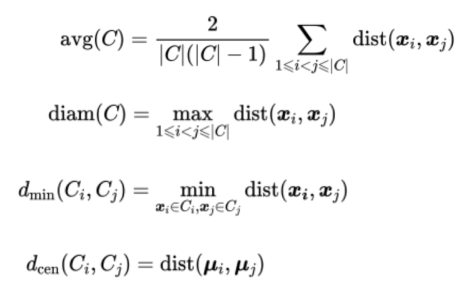
I．数据处理主要包括以下步骤：

1. 用1，-1，0分别代替，‘y’，‘n’，‘？’。
2. 把多维数据降维为2维（使用上述数据替代方法的原因也是为了在降维中尽量减少对数据的影响---保持‘y’，‘n’的同等效果，减少‘？’的影响）

注：尝试过用9来代替‘？’，结果产生了如下的聚类结果：

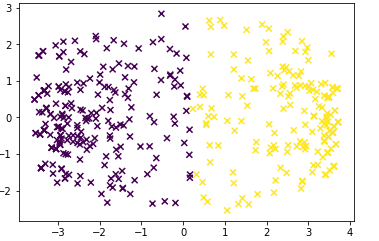


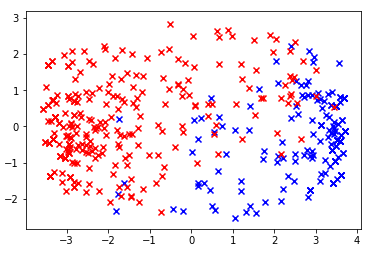
1. 在对聚类做评价时，采用了内部评价指标，即



II.聚类成果

1. 可视化：

----🡪此图为聚类结果的可视化图

----🡪此图是按照真实的党派标签的结果

1. 聚类评价

聚类得到的各个指标：

1）簇内平均距离：

聚类得到的共和党派的avg(c)为1.8949841881585363

聚类得到的民主党派的avg(c)为1.9125389919760962

1. 簇内最大距离：

聚类得到的共和党派的diam(c)为5.198664619227401

聚类得到的民主党派的diam(c)为5.315512164657216

1. 簇间最小距离：

两个聚类的簇最近的两个样本的距离为：0.15248405503877316

1. 簇中心距离：

两个簇的中心点的距离为4.851762143635815

真实分类的各个指标：

1）簇内平均距离：

真实的共和党派的avg(c)为1.9279370693855515

真实的民主党派的avg(c)为2.5163726676860785

2) 簇内最大距离：

真实的共和党派的diam(c)为6.734401503840723

真实的民主党派的diam(c)为7.111757611276915

3) 簇间最小距离：

真实的两个的簇最近的两个样本的距离为：0.042264160918674486

# 问题2：

由于在KMeans聚类算法中，每次的中心点是随机选择的，也就是说，每次运行聚类库函数的代码，都会随机选择不同的中心。所以说，即便每次运行了聚类函数，产生的聚类结果并无明显差异，而且所得评价指标的值均未发生改变，所以，我认为，选取不同的初始位点，对于聚类的最终结果没有影响。