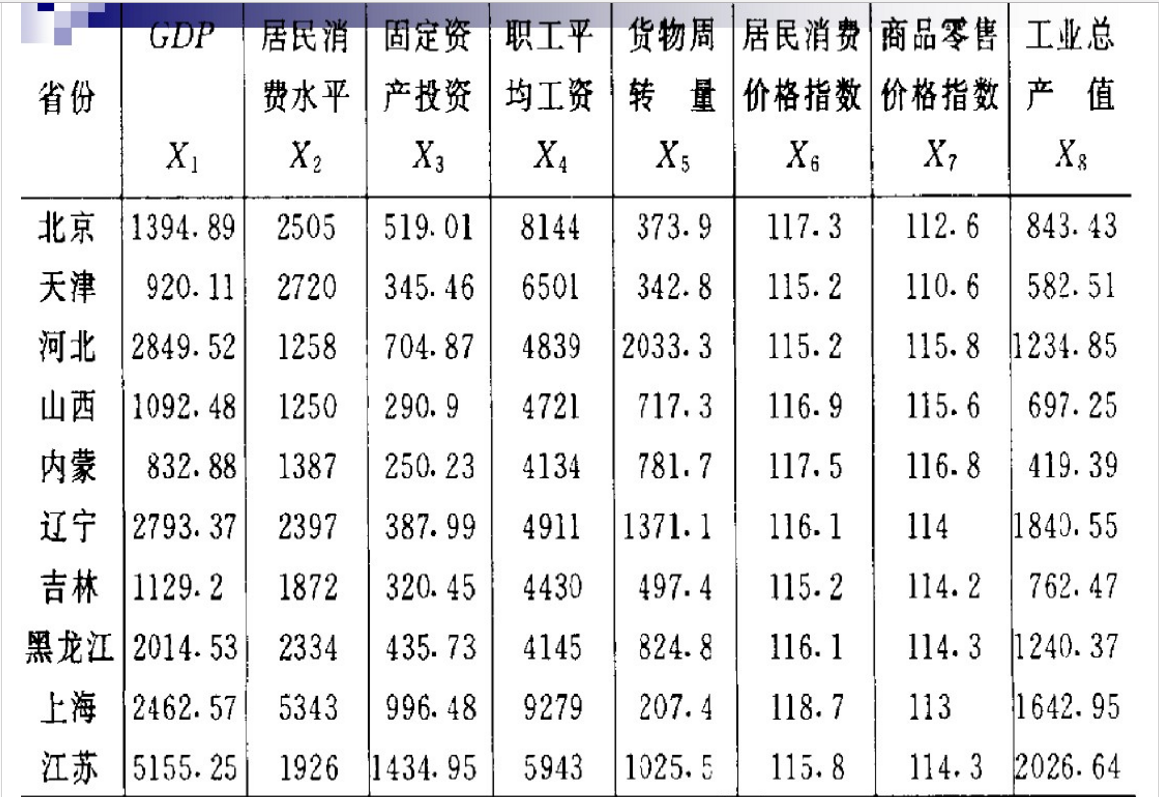
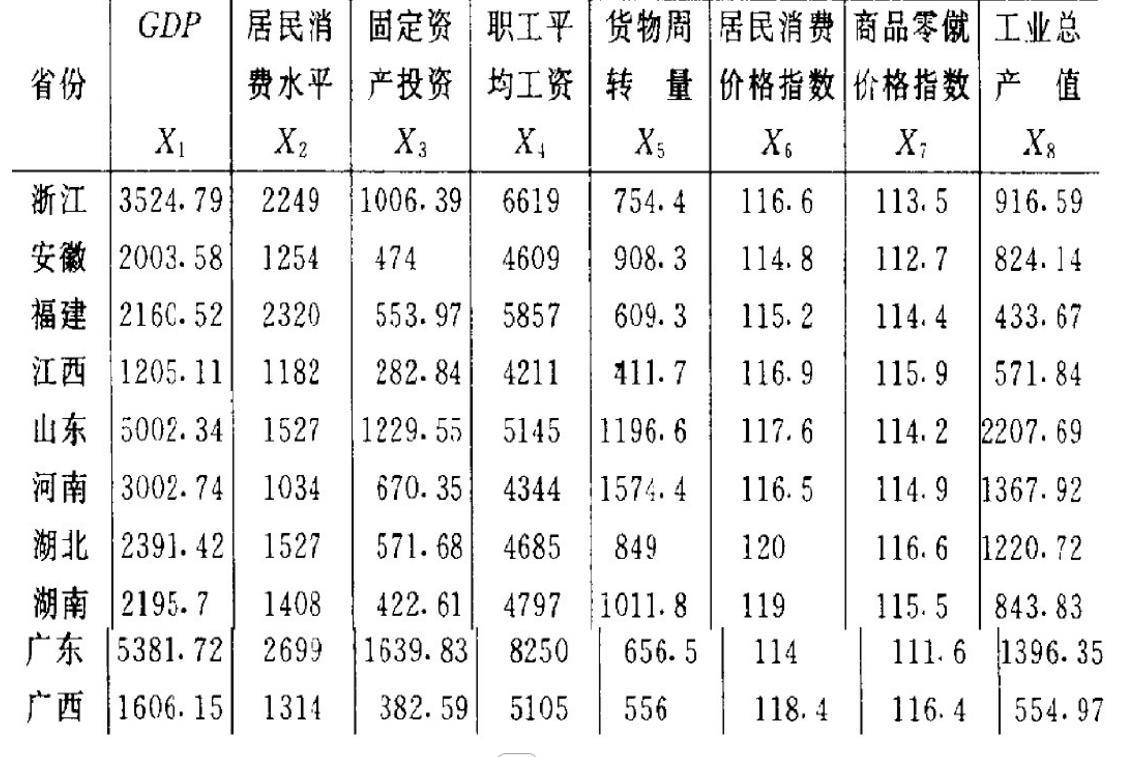
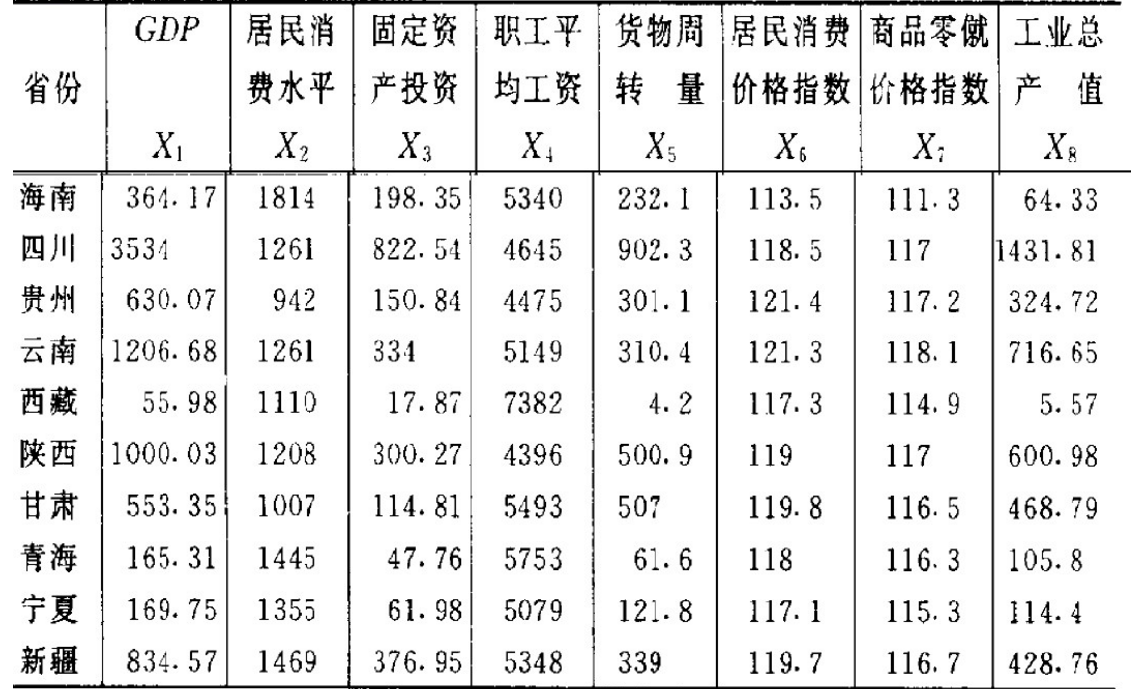
# 聚类分析

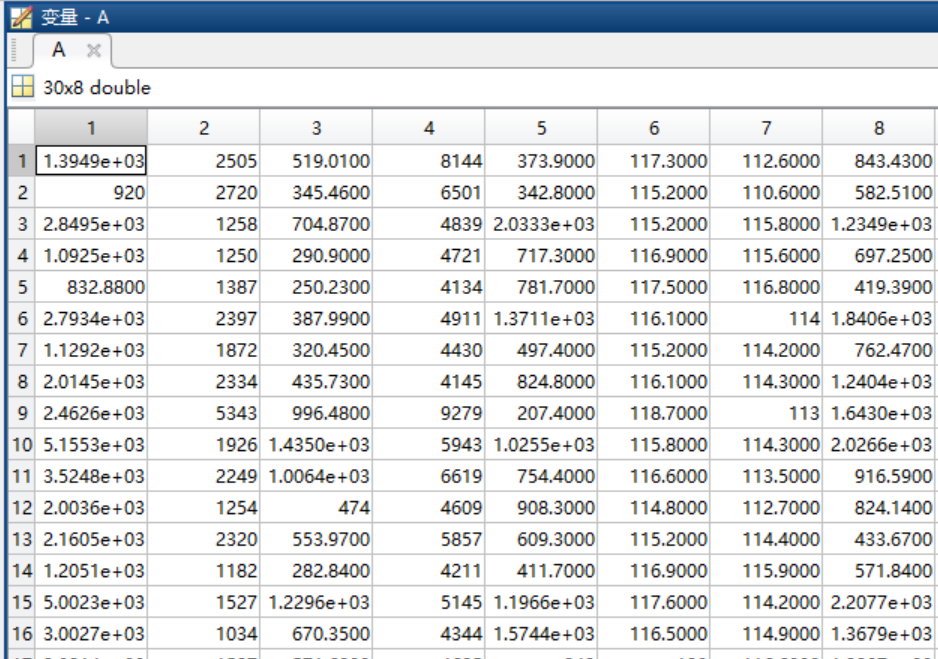
本次我选择的数据是30个省份的反映经济发展状况的八个属性，并将这30个城市聚类，观察是否合理。



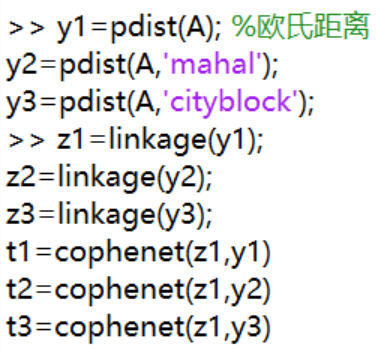


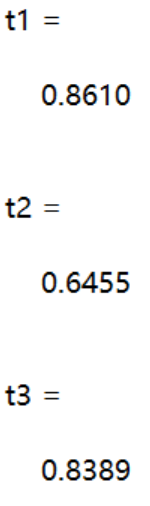


数据如上所示。

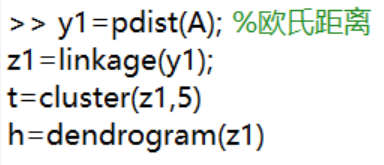


将数据输入数组A中





t1最大，判断出用欧式距离最好。



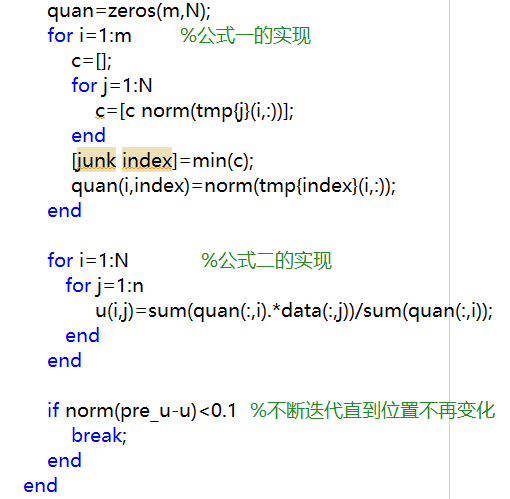
用欧式距离将城市分为5类，设想是将城市分类成：发达、较发达、中等、较不发达、不发达

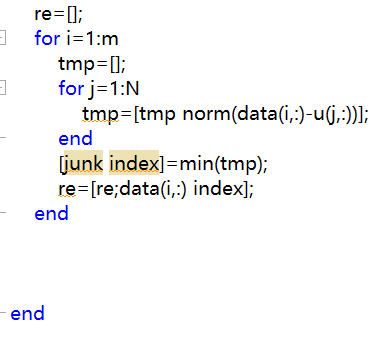
结果是：北京为第一类，江苏和山东为第三类，广东为第四类，上海为第五类，其余为第二类。

分析：虽然最发达的城市被挑选了出来，但是它们并没有被划分成一类，而是分成了不同类；不发达的地区，如西藏，也没有被挑选出来，而是被划入了大多数的省份里，所以这个分类有欠缺。

所以再构造kmeans方法：







N是数据一共分多少类

data是输入的不带分类标号的数据

u是每一类的中心

re是返回的带分类标号的数据



再调用它，在数组u中查看分类结果：

北京、上海为第一类；

山西、内蒙、吉林、江西、广西、贵州为第二类；

河北、辽宁、黑龙江、安徽、福建、河南、湖北、湖南、四川为第三类；

天津、西藏、海南、甘肃、青海、宁夏、新疆、云南、陕西为第四类

广东、江苏、浙江、山东为第五类。

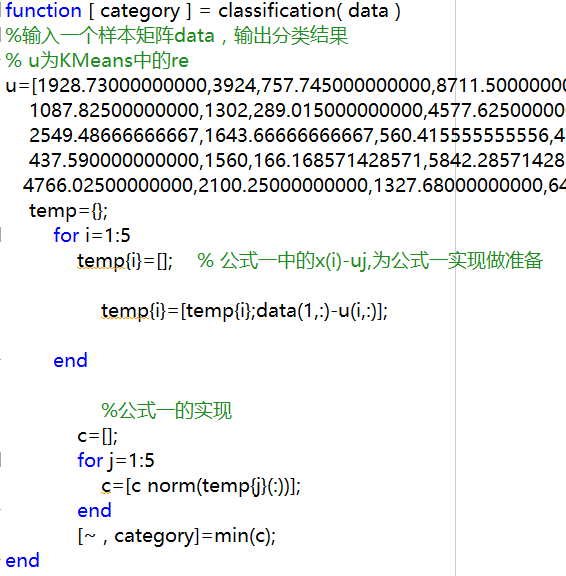
这个分类从直观上看要比之前的合理很多，相同级别的几乎都聚拢为一类。

经济发达的层次为：第一类 第五类 第三类 第二类 第四类

然而，其中仍有瑕疵：天津分在了第五类

可以说，k-means聚类是最适合这类样本的。

接着构造分类器：





将西藏的数据传入C并调用分类器，得出结果：，与聚类结果相符。