挖掘初始模型总结

1.案例中如何选择挖掘工具？

1）概括模型输入输出组的特征：本案例的输入组为 7 个数值型数据，输出组为 1 个类别型数据

2）本案例的挖掘目标：宋朝官员的特征描述

通过输入、输出组变量的特征和挖掘目标决定使用聚类算法

2. K-means 算法的基本流程是什么？

1）随机选取 K 个数据点作为 K 个起始群中心

2）计算剩余的每个数据点到 K 个群中心的距离，并将数据点分配到距离最近的群中心

3）分配完毕后，重新计算各群的中心

4）迭代上述 2、3 步，直到达到停止条件

停止条件包括中心位置不再发生变化，达到限定的最大迭代次数，或者中心的选择范围在一定区间等。

3. 在 Python 中如何进行 K-means 聚类？

1）引入 K-means 算法

从 python 的 scikit-learn 机器学习库中引入 k-means 算法：

from sklearn.cluster import KMeans

2）对官员进行聚类

KMeans 算法的参数有 11 个，通过设置具体参数观察聚类结果。本次案例使用 KMeans 中最常用的参数 n_clusters 对官员进行聚类 ,设置参数 n_clusters 等于 2 将官员聚为两类：

KMeans(n_clusters=2).fit_predict(politic_relation_rlt)

括号中标红的部分为聚类使用的数据集，即数据合并后的结果

3）筛选出聚类结果

①添加新列为聚类结果

②使用 query() 分别筛选出聚类后的两种类别官员