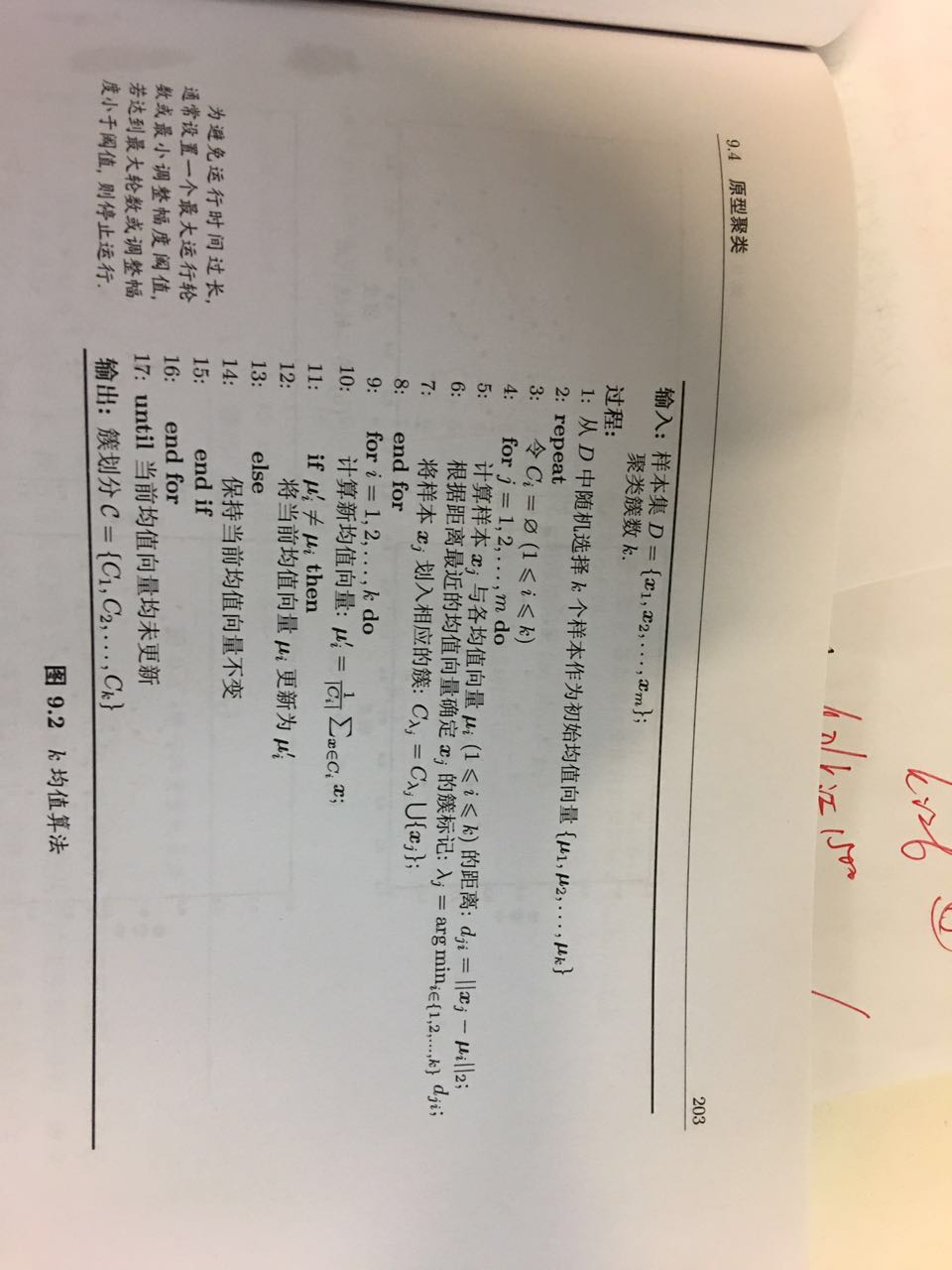
**聚类**

1. **Kmeans(贪心策略,通过迭代优化求近似解)**：

（k为参数，把n个对象分成k个簇，使簇内具有较高的相似度，而簇间的相似度较低。采用平方误差准则http://files.chinaaet.com/images/2010/10/18/4435423708146.gif），算法流程：

****

1. **层次聚类**：

先将每个对象作为一个簇，然后合并这些原子簇为越来越大的簇，直到所有对象都在一个簇中，或者某个终结条件被满足。

(1) 将每个对象看作一类，计算两两之间的最小距离；  
(2) 将距离最小的两个类合并成一个新类；  
(3) 重新计算新类与所有类之间的距离；  
(4) 重复(2)、(3)，直到所有类最后合并成一类。

**3、SOM聚类（**自组织映射神经网络， 即Self Organizing Maps (SOM)**）**：

假设在输入对象中存在一些拓扑结构或顺序，可以实现从输入空间(n维)到输出平面(2维)的降维映射，其映射具有拓扑特征保持性质,与实际的大脑处理有很强的理论联系。

(1) 网络初始化，对输出层每个节点权重赋初值；  
(2) 将输入样本中随机选取输入向量，找到与输入向量距离最小的权重向量；  
(3) 定义获胜单元，在获胜单元的邻近区域调整权重使其向输入向量靠拢；  
(4) 提供新样本、进行训练；  
(5) 收缩邻域半径、减小学习率、重复，直到小于允许值，输出聚类结果。

**4、FCM聚类（**模糊C均值(Fuzzy C-means)**）**：

用模糊数学的方法进行聚类分析，就是模糊聚类分析

(1) 标准化数据矩阵；  
(2) 建立模糊相似矩阵，初始化隶属矩阵；  
(3) 算法开始迭代，直到目标函数收敛到极小值；  
(4) 根据迭代结果，由最后的隶属矩阵确定数据所属的类，显示最后的聚类结果。

**5、聚类算法比较：**

k-means和FCM相对优于其他，k-means聚类算法的初始点选择不稳定，是随机选取的，这就引起聚类结果的不稳定。层次聚类虽然 不需要确定分类数，但是一旦一个分裂或者合并被执行，就不能修正，聚类质量受限制；FCM对初始聚类中心敏感，需要人为确定聚类数，容易陷入局部最优解；SOM与实际大脑处理有很强的理论联系。但是处理时间较长，需要进一步研究使其适应大型数据库。

**问题与答案：**

**1）k-means怎样判断好坏**

聚类性能度量：

外部指标：聚类结果与参考模型比较，Jaccard系数（JC），FM指数，Rand指数

内部指标：直接考查聚类结果，DB指数，Duun指数

《西瓜书》

**2）k均值算法怎么工作、决策树怎么工作**

**3）用伪代码实现一种你熟悉的聚类算法？**

**k-means**

**4）用过哪些聚类算法，解释密度聚类算法。聚类算法中的距离度量有哪些？**

k-means、高斯混合聚类（这俩属于原型聚类）/层次聚类/密度聚类

距离度量：Lp距离（也叫闵可夫斯基距离，p=2时为欧式距离，对有序属性进行度量）；VDM（无序属性）；距离度量学习（学习相似度度量）