机器学习第七次作业

1. 针对西瓜数据3.0α计算样本1-8上两两之间的闵可夫斯基距离（书9.18式）. 对形如$\boldsymbol{x}_i = (x_{i,1}; \ldots; x_{in})$的样本, 两个样本$\boldsymbol{x}_i$与$\boldsymbol{x}_j$间的闵可夫斯基距离为：

$$\mathrm{dist}_{mk}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{u=1}^{n} |x_{iu} - x_{ju}|^p \right)^{\frac{1}{p}} .$$

   本题中，p取值为1,2,3.

2. 针对西瓜数据3.0计算属性“色泽”上两个离散值“青绿”和“乌黑”之间的VDM距离；属性“纹理”上两个离散值“清晰”和“模糊”之间的VDM距离（书9.21式）. 对属性u上两个离散值的a与b间的VDM距离为：

$$\mathrm{VDM}_p(a, b) = \sum_{i=1}^{k} \left| \frac{m_{u,a,i}}{m_{u,a}} - \frac{m_{u,b,i}}{m_{u,b}} \right|^p .$$

   本题中，p取值为1,2,3，类别按照数据集的真实类别判定.

3. 编程实现k均值算法在给定的数据集（glass, fourclass和西瓜数据4.0）上进行测试。设置三组不同的k值（数据类别数目的1，2，3倍），使用不同的初始化方式（初始化尽可能不同，如果是随机初始化，请给出随机数种子），分析结果。讨论k的取值和初始化方式对聚类结果的影响。本题中任选1-2种聚类指标进行结果分析即可.