第2小组的数据挖掘作业2

**小组成员：高鹏昺 蒋世豪 李进雄 周亮 苏金涛 刘昊轩**

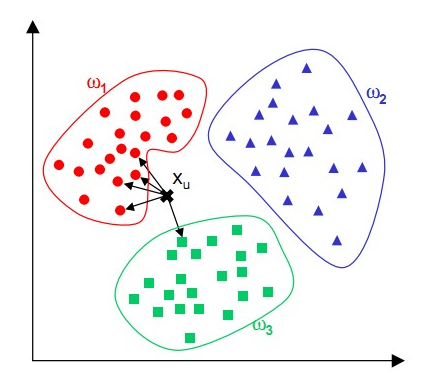
**方法 ：采用KNN算法进行差异甲基化区域数据挖掘**

**完成人　15051722 刘昊轩 q:525270633**

1.对于采用KNN算法的初步构想

KNN是最简单最基础的数据挖掘算法，重点在于ｋ的选取，数据标准化和数据处理上的时间复杂度，有多种包引入(numpy,sk-learn,pandas等)。对于交叉验证，采用Ｋ折交叉验证和留一法，比较分类精度。

2.基于KNN的预测方法建立



（1）相关原理介绍

对于一个被预测的新样本x，K近邻分类器将（在当前的d维空间）找到与该样本x最近的k个样本（来自于训练数据集）（这k个样本称为该样本的k个最近邻）；k个最近邻样本中所占类别数最多的那个类就是被预测样本的类别。

kNN算法的核心思想是如果一个样本在特征空间中的k个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。该方法在确定分类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。 kNN方法在类别决策时，只与极少量的相邻样本有关。由于kNN方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别的，因此对于类域的交叉或重叠较多的待分样本集来说，kNN方法较其他方法更为适合。

（2）具体的操作实现

①导入数据

②标准化数据（**0-1标准化**和**Z-Score**）

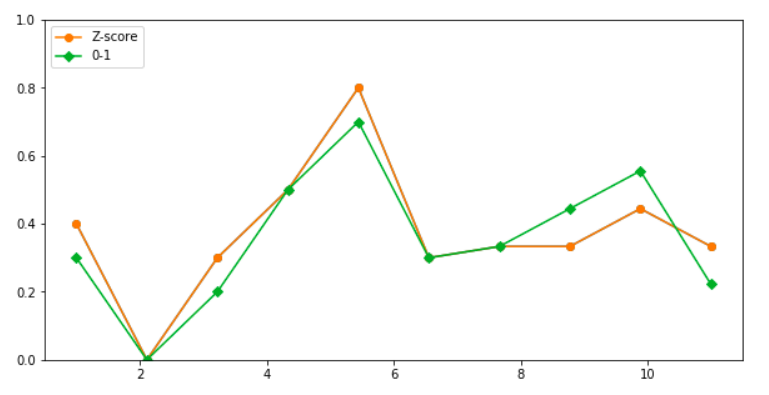
③根据上面的理论论述，构造KNN算法代码。

④调用KNN算法，输入数据，返回k折交叉验证和留一法在不同标准化数据集上的分类精度，比较不同K值的分类精度，画图

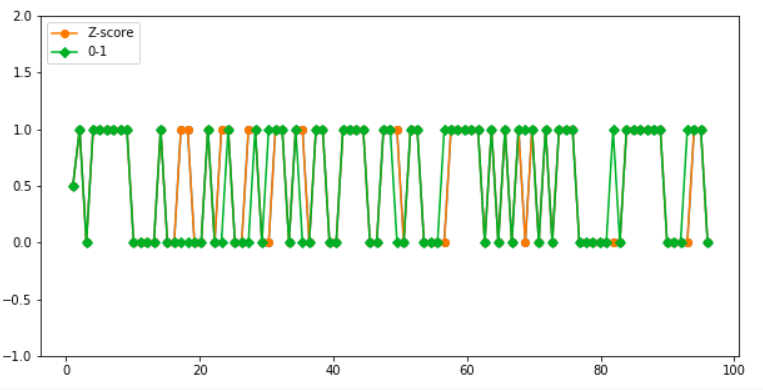
3.输出结果

在Ｋ值探索阶段我们可以通过比较不同ｋ值的分类精度，决定选取的ｋ值．通过python可视化作图得到下图。将不同k值的k折和留一法分类精度打印，放Output\_KNN.txt.

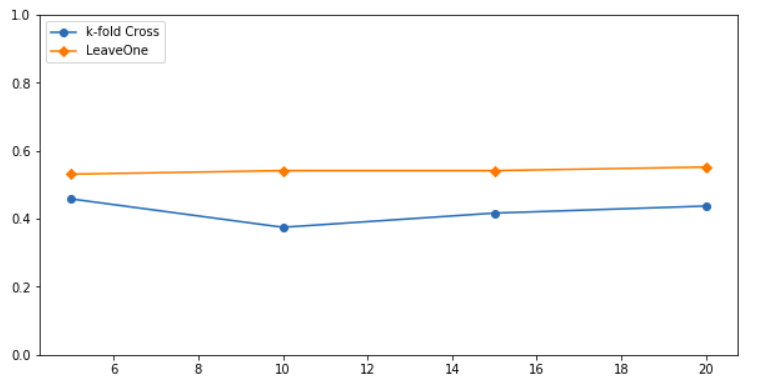
**10折交叉验证分布图**(knn.k=10)



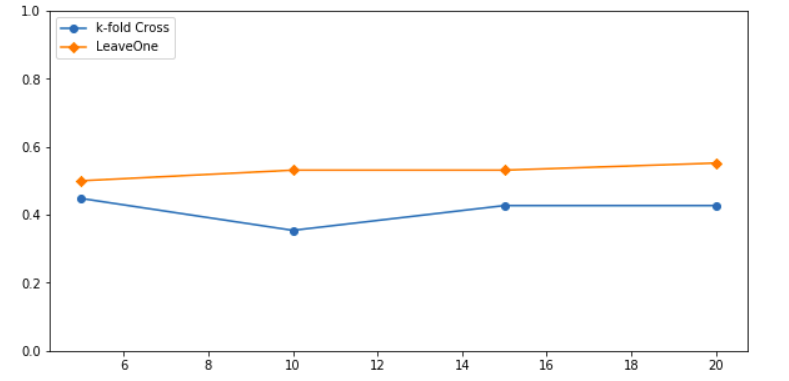
**留一法分布图**(knn.k=10)



**不同**k**值下的分类精度（Z-Score）**



**不同**k**值下的分类精度（0-1标准化）**



4.模型程序

程序位置：KNN文件夹

环境：Python2.7 Jupyter notebook

可执行文件：KNN-k折-留一法.ipynb KNN-k折-留一法.html

结果输出文件：Output\_KNN.txt