

计算机与信息技术学院

**实验名称 贝叶斯分类&聚类&决策树综合实验报告**

**姓 名 龙敏**

**学 号 18281046**

**日 期 2020.12.22**

1. 贝叶斯分类

# 使用sklearn的GaussianNB、BernoulliNB、MultinomialNB完成spambase垃圾邮件分类任务

# 使用各个模型得到结果如下：

# 

结合实验结果,可以得到结论：

GaussianNB的查全率最高，查准率最低，综合效果较好；

BernoulliNB的精度、查准率、F1值最高，综合效果最好；

MultinomialNB的精度、查全率、F1值均为最低，综合效果最差。

1. **实现一个高斯朴素贝叶斯分类器**

使用spambase数据集，通过自己设计的高斯朴素贝叶斯分类器，计算模型的查准率，查全率，F1值得到结果如下表所示：



# 实现带有拉普拉斯修正的朴素贝叶斯

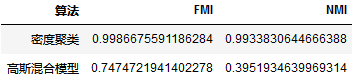
拉普拉斯修正的作用：

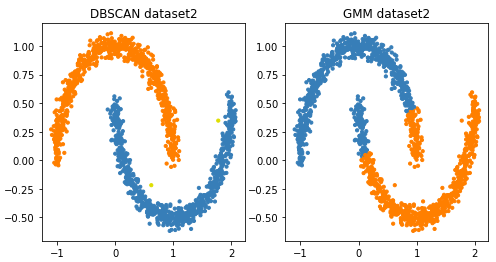
Laplace smoothing 是属于贝叶斯估计范畴的，具体而言就是加入一个先验分布。但在我们没有很好的先验知识的时候，我们一般不要自作主张地加入偏见，因此此时引入无信息先验是比较合理的（即均匀分布，对待所有情况一视同仁）拉普拉斯平滑对应于共轭先验参数取1的情况，即均匀分布。

使用实现的分类器，并使用现有的spambase.data数据集（实验代码见附件）。得到模型评价指标如下：

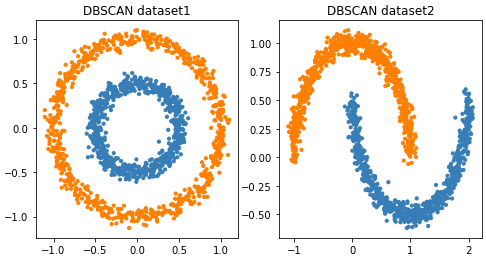


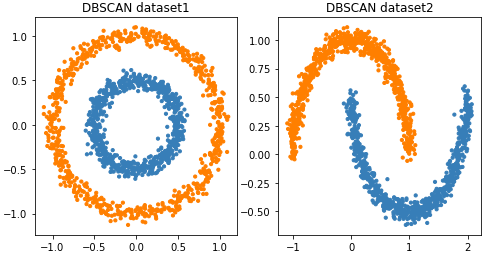
1. 聚类
2. **使用sklearn的DBSCAN和GaussianMixture完成聚类**
3. 使用密度聚类和高斯混合模型在数据集X2上进行聚类，绘制聚类后的效果。计算FMI和NMI值，及结果如下所示：

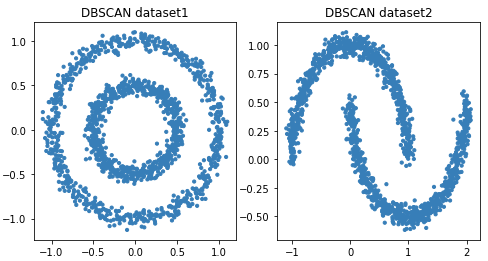




1. 选做：尝试调整密度聚类的eps的值，绘制聚类结果如下：

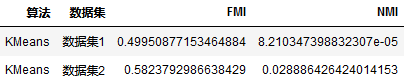
Esp=0.2:

Esp=0.23：

Esp=0.24：

# 2． 使用sklearn的KMeans完成聚类

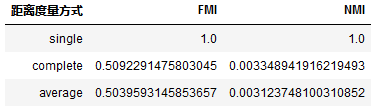
实现Kmeans聚类算法在两个数据集上，指标计算结果如下：

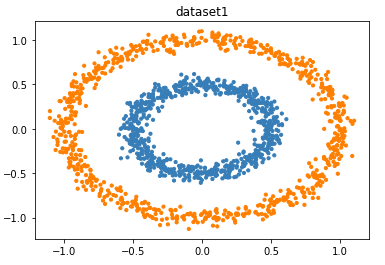


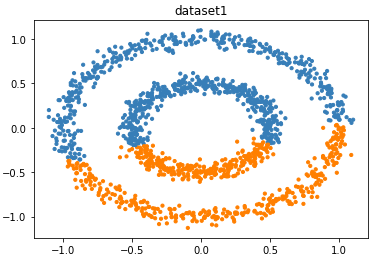
**3. 使用sklearn的AgglomerativeClustering完成层次聚类**

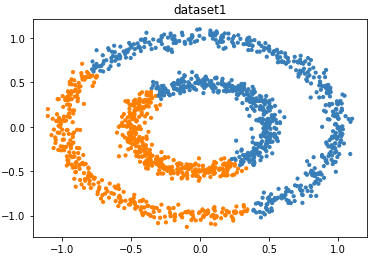
请你完成"single"，"complete"和"average"三种距离在两个数据集上的聚类任务，并绘制出聚类结果，计算6种情况的FMI和NMI结果如下：

1. 数据集1

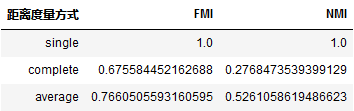


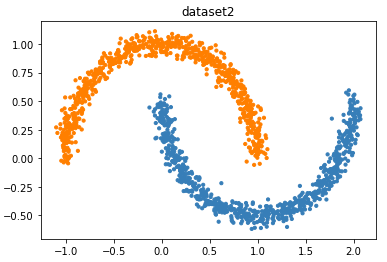
Single：

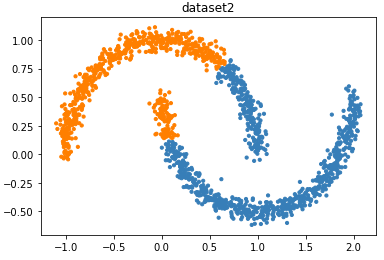
Complete:

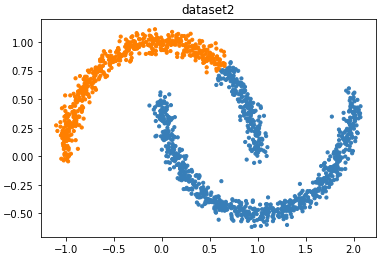
Average:

1. Dataset2



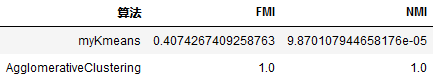
Single:

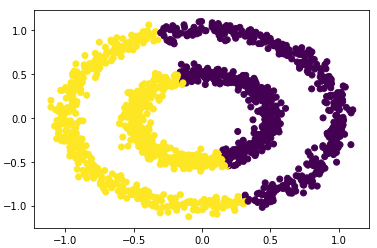
Complete:

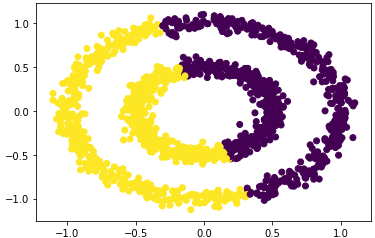
Average:

# 4. 实现K-means

# 使用提供的数据使用myKmeans和层次聚类算法(AgglomerativeClustering)对该数据进行聚类。计算出两个模型的FMI和NMI值，并对聚类结果可视化结果如下：



myKmeans：

AgglomerativeClustering：

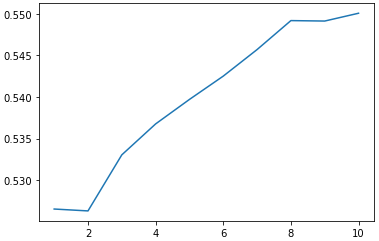
1. 决策树

# 1.决策树处理分类任务

针对dota2比赛结果预测问题，计算最大深度为10时，十折交叉验证的精度(accuracy)，查准率(precision)，查全率(recall)，F1值 ，结果如下图所示：



改变最大深度，绘制决策树的精度变换图绘制最大深度从1到10的决策树十折交叉验证精度的变化图如下：

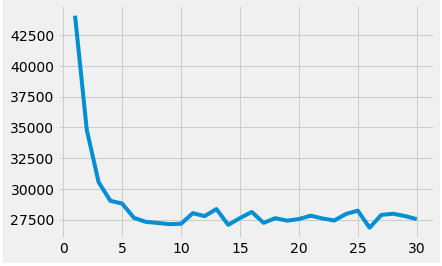


**2. 决策树处理回归任务**

使用sklearn.tree.DecisionTreeRegressor完成kaggle房价预测问题，计算最大深度为10的决策树，训练集上十折交叉验证的MAE和RMSE如下表所示：



改变最大深度，绘制决策树的精度变换图，绘制最大深度从1到30，决策树在训练集和测试集上MAE的变化曲线如下：



从图中可以看出，树最大深度达到10的时候效果很好。随着树深度的继续增大，会产生过拟合。而减小，则会产生欠拟合

# 3. 实现决策树

使用LendingClub Safe Loans数据集，实现信息增益、信息增益率、基尼指数三种划分标准,使用给定的训练集完成三种决策树的训练过程,计算三种决策树在最大深度为10时在训练集和测试集上的精度，查准率，查全率，F1值如下所示：



# 4. 实现预剪枝

实现使用信息增益率划分的预剪枝，计算出带有预剪枝和不带预剪枝的决策树的精度，查准率，查全率和F1值(最大深度为6，使用信息增益率)如下：

