QG工作室数据挖掘组中期考核

首先恭喜大家坚持到了QG工作室数据挖掘组中期考核，坚持到这里不容易。下面就是我们简简单单的中期考核了，具体考核要求如下：

（1）从下列10个算法中任选两个实现，有余力的同学可以实现更多算法（基础任务：实现两个；建议任务：实现四个）

（2）算法需实现两种形式：1. 只允许使用numpy,pandas,matplotlib三个主要库进行算法的编写实现；不允许调用Sklearn库。2.使用Sklearn库中的模型实现算法。将自己实现的算法和调用库的算法进行对比评估。

（3）部分数据集放在了data文件夹中，或者附带了下载链接。对数据集信息不了解的可以自行查看相关的数据集介绍文件，或者自行百度查找相关数据集介绍。

（4）这次中期考核要求实现详细文档。详细文档内容应包括数据集的处理，算法步骤和思想，算法实现结果评估，不足和优化之处等（可以自行编写）。详细文档封面放在附件中。

（5）中期考核还需要制作答辩ppt，答辩时间为3分钟，大家把握好时间制作ppt。

注意：中期考核时限为一星期，各位需要在下星期五大组培训之前将自己的代码、详细报告文档和中期答辩ppt放到github并且发给对应的导师。文档需要word格式和pdf格式。

一、K-means算法

1. 考核要求

1）基本任务：

内容一：理解k-means算法的思想

内容二：使用python实现k-means算法

内容三：使用Iris数据集进行测试

内容四：对参数k进行调整，记录结果

2）进阶任务：

实现二分k-means代码并测试，并进行比较

2. 数据集

数据集名称：Iris 数据集

二、k近邻算法

1.考核要求

1）基本任务：

内容一：理解KNN算法的思想

内容二：使用Python实现KNN算法

内容三：使用Iris数据集进行测试

内容四：记录测试结果

2）进阶任务：

自行优化KNN算法并进行代码测试

2. 数据集

数据集名称：Iris 数据集

三、 Apriori 算法

1. 考核要求：

内容一：理解 Apriori 算法的思想。

内容二：使用 Python 实现 Apriori 算法 。

内容三：使用 UCI 上面的 mushroom 数据集进行算法测试。

内容四：修改支持度和置信度，进行多次测试。

内容五：记录测试结果

2. 数据集：

数据集名称：mushroom 数据集

四、 线性回归算法

1. 考核要求：

内容一：理解 LinearRegression 算法的思想。

内容二：使用 Python 实现 LinearRegression 算法 。

内容三：使用 housing 数据集进行算法测试。

内容四：记录测试结果

内容五：自行优化线性回归算法（例如：梯度下降，岭回归等）

2. 数据集：

数据集名称：housing数据集

五、决策树算法

1.考核要求：

内容一：理解C4.5，ID3，CART决策树的算法思想（三者选一种）

内容二：使用python实现其中一种算法

内容三：使用对应的数据集进行算法测试

内容四：记录测试结果

内容五：尝试进行剪枝操作；实现其他两种决策树算法

2.数据集：

1）C4.5数据集： lenses数据集 （数据集介绍：https://wenku.baidu.com/view/12d9cc6548d7c1c708a145c5.html）

2）ID3数据集：同上，lenses数据集

3）CART数据集： （1）回归树：forestfires数据集

（2）分类树：Iris数据集

六、朴素贝叶斯算法

1. 考核要求：

内容一：理解简单朴素贝叶斯决策树的算法思想

内容二：使用python实现朴素贝叶斯算法

内容三：使用垃圾邮件数据集进行算法测试

内容四：记录测试结果

内容五：尝试实现高斯朴素贝叶斯和多分类朴素贝叶斯算法

2.数据集：

1）普通朴素贝叶斯：垃圾邮箱数据集

2）高斯朴素贝叶斯：iris数据集

3）多分类朴素贝叶斯：新闻分类数据集

七、DBSCAN算法

1. 考核要求：

内容一：理解DBSCAN的算法思想

内容二：使用python实现DBSAN算法

内容三：使用iris数据集进行算法测试

内容四：记录测试结果

内容五：自行优化算法

2. 数据集：

iris数据集

八、Logistic算法

1. 考核要求：

内容一：理解Logistic的算法思想

内容二：使用python实现该算法

内容三：使用adult数据集进行算法测试

内容四：记录测试结果

内容五：自行优化算法

2. 数据集：

adult数据集

九、bp神经网络

1. 考核要求

内容一：理解反向传播
内容二：使用Python实现bp神经网络算法
内容三：算法实战（使用mnist数据集进行分类，或使用波士顿房价数据集进行回归）
内容四：评估模型

2. 数据集

（1）mnist数据集：
获取渠道1：直接下载，url: http://yann.lecun.com/exdb/mnist/
获取渠道2：通过sklearn下载，参考代码：
from sklearn.datasets import fetch_mldata
mnist = fetch_mldata("MNIST original")

（2）波士顿房价数据集：
获取渠道1：通过sklearn加载，参考代码：
from sklearn.datasets import load_boston
house = load_boston()
获取渠道2：直接下载，url: https://archive.ics.uci.edu/ml/machine-learning-databases/housing/

十、支持向量机

1. 考核要求

内容一：理解支持向量机的算法思想

内容二：使用python实现该算法

内容三：使用iris数据集进行算法测试

内容四：记录测试结果

内容五：尝试优化（非线性分类(核函数)以及松弛变量，用拉格朗日乘子等方法）

2. 数据集

Iris数据集