机器学习课程论文要求

**一、论文研究内容（共两个方向）**

**每位同学可选其中“一个”进行研究及论文撰写。**

**1、机器学习不同算法运算效果比较与分析**

论文内容要求：

**①**文件夹data（csv）中含有17个数据集，从中任选三个数据集，选一种机器学习通过训练三种不同数据集进行比较分析。

**②**针对data（csv）中wine数据集，

选用三种不同的机器学习算法，完成葡萄酒数据集的分类预测（至少用三种机器学习算法实现，如支持向量机、神经网络、决策树等），并进行不同算法优劣的深入比较分析。可自定义分类评价函数，并比较定义不同的分类评价函数对预测结果的影响。

**③**扩展内容，可针对某种机器学习算法，进行算法优化改进等操作，在完成本题目①②要求之后，创新性进行算法实验。

**备注:**

1. 实验结果的比较分析，多采用图表进行展示；
2. 对不同算法的分类结果进行比较分析；
3. 分析比较不同算法的运行时间及过拟合程度；
4. 可说明采用了哪些方法防止过拟合；
5. 可评估wine数据集中不同特征对预测结果的影响，进行因子排序。

数据集信息说明示例如下：，大家可以自行在网上查找更详细说明





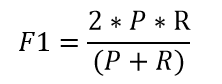
**2、糖尿病患病遗传风险预测与分析**

数据集：https://tianchi.aliyun.com/competition/entrance/231638/information

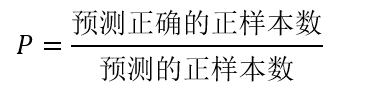


Train中是训练数据，test中是测试数据，用于评估模型的最终分数，label为1表示患病，0表示不患病。

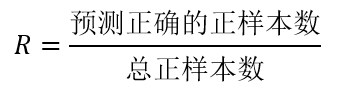
评分指标如下，F1越大越好。



其中P为准确率，计算公式如下：



R为召回率，计算公式如下（其中正样本数定义为数值为1的样本数）：



**备注:**

1. 至少用三种机器学习算法实现；
2. 评估比较不同算法的预测结果；
3. 说明防止过拟合采用了哪些方法；
4. 可评估比较不同的缺失值处理方法对预测结果的影响。

**二、论文格式要求**

1、按照提供的论文模板撰写论文，字体大小严格按照模板的要求，尤其注意要采用**“双栏”**。

2、论文的正文内容，陈述算法原理，建模步骤，实验结果及分析等，要求条理清楚，描述准确，避免口语化文字，**禁止大段拷贝代码**。

3、论文对应代码，另存为.py文件，与论文一起**打包**提交

4、论文页数要求：**至少四页**，论文命名方式为：**学号-姓名-论文题目.doc，**其中“论文题目”自拟，可根据所选研究内容及研究方法来命名，例如“基于\*\*\*的糖尿病风险预测”。

5、论文提交时间：2019年12月20日24点之前

论文提交邮箱：[565088286@qq.com](mailto:565088286@qq.com)

提交文件命名方式：学号-姓名-论文题目.doc