**《大数据分析基础》期末考核作业**

1. 给定一个“IT行业收入表”（income.xlsx）,其中包含了不同工龄的薪水情况，请编写一个python程序，构建一个用于预测薪水的一元线性回归模型，绘制散点图进行模型可视化，并对模型进行评估。（15分）
2. 给定一个“肿瘤数据表”(tumour.xlsx**),** 其中包含肿瘤的6个特征变量，分别是最大周长，最大凹陷度，平均凹陷度，最大面积，最大半径，平均灰度值。目标变量为肿瘤性质，0代表为恶性，1代表为良性。请编写一个python程序，按照8：2的比例来划分训练集和测试集，运用朴素贝叶斯模型来预测肿瘤为良性肿瘤还是恶性肿瘤，并绘制ROC（Receiver Operating Characteristic）图，对模型进行评估。（15分）
3. 给定某国家国民生产总值平减指数（GNP deflator）时间序列如excel文件（data1.xlsx），请用python编写程序，完成如下功能：
4. 判断该序列的平稳性；
5. 判断该序列是否为纯随机序列；
6. 绘制该序列时序图、自相关图和偏自相关图；
7. 采用合适的时间序列模型对上述序列进行拟合，并对未来10年的序列值进行预测。

（30分）

1. 分析和理解SVR（支持向量机回归）方法原理，通过查阅相关资料，简述求解SVR问题的SMO（Sequential Minimal Optimization，序列最小化算法）原理，并结合教材第6章案例，分析SVR方法有什么优点和缺点？如何改进SVR？（20分）
2. 阅读和理解教材11章中的“电子商务网站用户行为分析及推荐服务”综合案例，并成功运行该案例，分析运行结果，掌握协同过滤算法的原理。通过查阅相关文献，了解对协同过滤算法的改进方法。**本题需要提交：**
3. 运行结果截图，包括准确率、召回率和F1三个指标截图;
4. 至少一种改进的协同过滤算法的描述(主要方法或流程等)。

（20分）