人工智能与机器学习综合实践期末考核内容与要求

一、基本知识点总结（20 分）

1、整理所学过的所有机器模型，根据类别：分类模型，回归模型和无监督模型进行总结（10 分）

要求：

（1）总结 sklearn 中每类模型中都调用了哪些方法（函数名）,常用的参数及含义，

（2）每种方法的原理（简单概括）、

（3）每类模型的评价指标及含义

2、总结数据预处理、特征优化、模型优化等方法（10 分）

（1）数据预处理有哪些方法，具体功能

（2）特征优化有哪些方法，简单描述

（3）模型优化参数调优具体方法有哪些?

二、实践操作（75 分）

（一）信用卡虚假交易识别（30 分）

数据集 creditcard.csv

数据说明：数据文件来自一个欧洲的信用卡发卡机构，记录了 2013 年 9 月中某两天的信用卡刷卡活动。经过事后分析判断，在数据文件记录的 284807 次交易行为中，有 492 次虚假交易行为。因为涉及到商业秘密和客户隐私，数据文件是原始记录经过主成分分析法 PCA（从较多维度数据中提取出有价值的较少维度的数据）以后提取出来的。

一共有 284807 条记录，每条记录对应一次交易的持卡人信息或者交易信息。每条记录由 28 个字段，用 V1，V2……V28 表示，除了交易时间和交易金额是原始数据之外，其他字段都经过了转换。“Class”属性取 1，则表示是一个虚假交易；0 则表示是正常交易。

1、数据清晰与因素分析（15 分）

（1）读入交易数据，显示数据概况（基本信息）

（2）查看欺诈交易与正常交易的数据量对比

（3）查看正常交易/欺诈交易不同类别下交易时间和交易金额的关系

（4）对 V1-V28 个属性进行分析，根据其特征重要性选择其中一半的属性作为特征

2、模型训练（15 分）

（1）按 7：3 的比例将数据集分成训练集和测试集

（2）分别使用逻辑回归、随机森林、支持向量机三种模型分别对训练集进行训练，输出混淆矩阵、正确率、精确率、召回率

（3）对比三种模型的结果，选择一种模型对测试集进行测试，输出测试集五折交叉验证的正确率

（二）综合案例（45 分）

案例背景：准确预测汽车备件的需求量，有助于厂家合理规划其生产流水线，从而避免出现备件紧缺或备件多余耗费仓库存放空间的情况。

数据集：sales_details.csv

用于分析的数据属性特征有：创建日期、物料编号、应发库、销售订单数量和交货数量

具体要求：

1、数据预处理（15 分）

（1）缺失值统计与处理（少量删除）

（2）异常值处理：原数据集中汽车备件的应发库范围为 9 个城市，检查其对应的内容是否属于这 9 个城市之一，不属于的删除处理。

（3）数据格式转换 1：原始数据集中销售订单数量和交货数量两项中的数值使用了千位分隔符，会被误认为字符串格式，因此需要清除数据中的所有千位分隔符，并转换数据类型。

（4）数据格式转换 2：原始数据中创建日期精确到日，由于汽车备件的销量与年份和月份都存在一定的关系，所以需要对数据进行转换，将年份与月份单独提取，作为独立的两列。

（5）数据编码：将物料编号和应发库两个特征数值化

（6）统计数据：处理完数据后，按照创建年、创建月、物料编号和应发库四个维度分组后对销售订单数量进行求和。

2、构建数据集（5 分）

（1）得到按年、月、物料编号与应发库四个维度进行统计的数据后，将原数据集再次随机打乱，防止相似数据集中影响训练质量

```
shuffled_indices=np.random.RandomState(seed = 15).permutation(len(df_new))
df_new=df_new.iloc[shuffled_indices]
```

（2）划分训练集和测试集：将数据划分为训练数据集、验证数据集和测试数据集，其中训练集用来获得预测模型，验证集用来确定控制模型复杂程度的参数，而测试集则用来检验最终选择的最优的模型的性能如何。3 个数据集的比例是 7：1：2

3、模型训练（10 分）

（1）使用决策树回归模型进行训练，其中 max_depth 取值范围从 2 到 15，基于验证集查找最优的决策树深度及模型

（2）使用训练好的模型对测试集进行测试，输出测试集的性能指标

（3）模型调试：选择其他回归模型对数据进行预测，输出测试集的性能指标，并对比结果

4、时间序列分析（5 分）

（1）基于 1（5）的结果将销售订单数量根据年月进行求和

（2）将年月日期作为 x 轴，统计结果作为 y 轴，画折线图，观察 2013——2015 年的总体趋势变化

5、聚类（10 分）

（1）基于 1（5）的结果，去除第一列，转换为数组。

（2）选用一种聚类方法，对数据进行聚类，输出类别标签及每个类别的中心点，对中心数据进行说明

（3）输出当聚类个数 K 等于不同值时的轮廓系数得分

三、总结与思考（5 分）

对该门课程学习过程中遇到的问题及期末考核实践题目中遇到的问题的总结与思考。

总体要求：

（1）所有内容用 word 整理，

（2）封面要求：山东财经大学

名称：人工智能与机器学习综合实践大论文

学院：

专业班级：

学号：

姓名：

（3）word 文件名：学号+姓名

第一部分知识点总结，重点是归纳总结所学过的内容，切记简单罗列，内容不超过 5 页。

两个人重复率 80%以上期末成绩直接为零，重复 50%以上，会扣分。