# 2024年秋季学期《机器学习》课程

# 大作业

**（提交截至时间：2024年11月15日）**

学号：S324060088

姓名：王子旭

哈尔滨工程大学

2024年10月

## 第一部分 基础简答

问题1：过拟合与欠拟合（定义、产生的原因、解决的方法各是什么）。

问题2：L1正则与L2正则（有哪些常见的[正则化](https://so.csdn.net/so/search?q=%E6%AD%A3%E5%88%99%E5%8C%96&spm=1001.2101.3001.7020" \t "https://blog.csdn.net/weixin_41524411/article/details/_blank)方法？作用各是什么？区别是什么？为什么加正则化项能防止模型过拟合）。

间题3：模型方差和偏差（能解释一下机器学习中的方差和偏差吗？哪些模型是降低模型方差的？哪些模型是降低模型偏差的？举例说明一下）。

问题4：奥卡姆剃刀（说一说机器学习中的奥卡剪梯刀原理）。

问题5：模型评估指标（回归模型和分类模型各有哪些常见的评估指标？各白的含义是什么？解释一下AUC？你在平时的实践过程中用到过哪些评估捐标？为什要选择这些指标）。

问题6：优化算法（机器学习中常见的优化算法有哪些？梯度下降法原理推导）。

问题7：激活函放（神经网络模型中常用的激活函数有哪些？说一下各自的特点)。

问题8：核函数（核函数的定义和作用是什么？常用的核函数有哪些？你用过些核函数？说一下高斯核函数中的参数作用）。

问题9：梯度消失与梯度爆炸（解释一下梯度消失与梯度爆炸问题，各自有什么解决方案）。

问题10：有监督学习和无监督学习（说一下有监督学习和无监督学习的特点，举例说明一下)。

问题11：生成模型与判别模型（你知道生成模型和判别模型吗？各自的将点是什么？哪些模型是生成模型，哪些模型是判别模型）。

问题12：线性回归（线性回归模型的原理、损失函数、正则化项)。

问题13：KNN模型（KNN模型的原理、三要素、优化方案以及模望的优缺点）。

问题14：朴素贝叶斯（朴素贝叶斯模型的原理推导，拉普拉斯平滑，后验概率最大化的含义以及模型的优/缺点）。

问题15：决策树（决策树模型的原理、特征评价指标、剪枝过程和原理、几种常见的决策树模型、各自的优/缺点）。

问题16：随机森林模型（RF模型的基本原理，RF模型的两个“随机”。从偏差和方差角度说一下RF模型的优缺点，以及RF模型和梯度提升树模型的区别）。

问题17：AdaBoost（AdaBoost模型的原理推导、从偏差和方差角度说一下AdaBoost、Adaboost模型的优点）。

问题18：梯度提升树模型（GBDT模型的原理推导、使用GBDT模型进行特征组合的过程、GBDT模型的优缺点）。

问题19：XGBoost（XGBoost模型的基本原理、XGBoost模型和GBDT模型的异同点、XGBoost模型的优缺点）。

问题20：Logistic Regression回归模型（LR模型的原理、本质，LR模型的损失函数，能否使用均方损失、为什么）。

问题21：支持向量机模型（SVM模型的原理，什么是“支持向量”？为什么使用拉格朗日对偶性？说一下KKT条件、软间隔SVM和硬间隔SVM的异同点。SVM怎样实现非线性分类？SVM常用的核函数有哪些？SVM模型的优缺点各是什么）

问题22：K-Means类（K-Means聚类的过程和原理是什么？优化方案有哪各自优缺点是什么)。

问题23：层次聚类（层次聚类的过程、原理和优缺点）。

间题24：密度聚类（密度聚类的基本原理和优缺点）。

问题25：谱聚类（谱聚类的基本原理和优缺点）。

问题26：深度神经网络（深度神经网终模型的原理，反向传播的推导过程，常用的激活函数，梯度消失与梯度爆炸问题怎么解决？说一下深度神经网络中的Dropout、早停、正则化）。

问题27：K近邻模型与 K-Means模型的异同点。

问题28：ID3决策树、C4.5决策树、CART决策树的异同点。

问题29：Bagging模型与 Boosting模型的异同点。

问题30：GBDT模型与 XGBoost模型的异同点。

## 算法实践

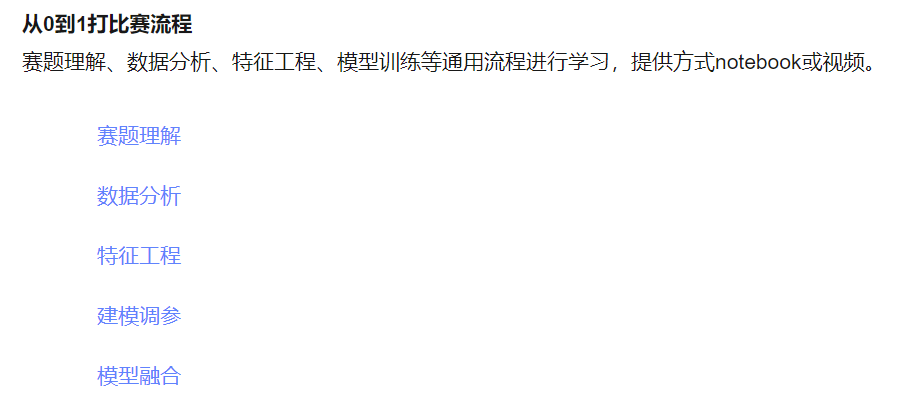
题目：[首页](https://tianchi.aliyun.com/home)>[天池大赛](https://tianchi.aliyun.com/competition/gameList/activeList)>[日常学习赛](https://tianchi.aliyun.com/competition/getstartList)>零基础入门数据挖掘 - 二手车交易价格预测

<https://tianchi.aliyun.com/competition/entrance/231784>

### 任务1：

学习**赛制与数据**页面中的[赛题理解](https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.15.1cd8593aw4bbL5&postId=95456)、[数据分析](https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.12.1cd8593aw4bbL5&postId=95457)、[特征工程](https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.6.1cd8593aw4bbL5&postId=95501)、[建模调参](https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.18.1cd8593aw4bbL5&postId=95460)、[模型融合](https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.3.1cd8593aw4bbL5&postId=95535)内容。

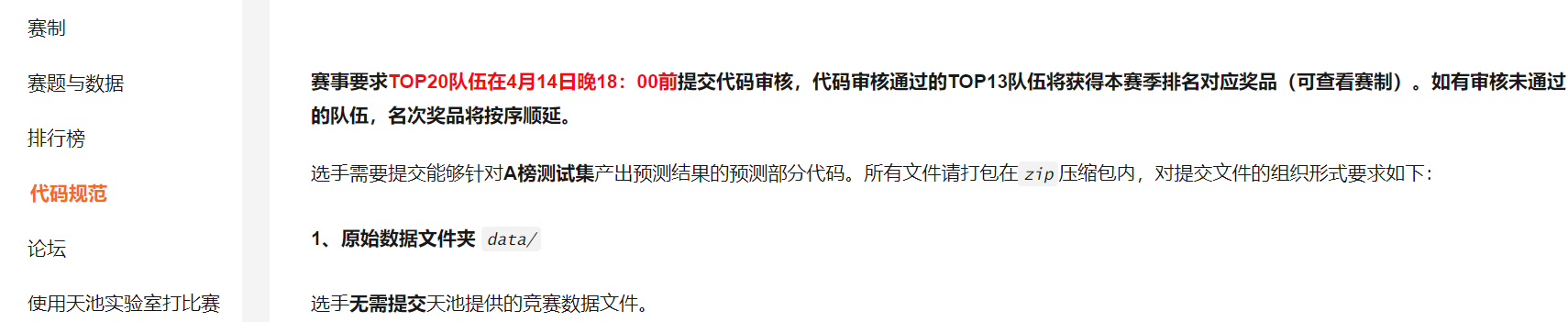




### 任务2：

完成二手车交易价格预测编程，程序提交方式按赛事页面**代码规范**下的要求组织、提交源代码。

https://tianchi.aliyun.com/competition/entrance/231784/customize191



所有文件请打包在zip压缩包内，对提交文件的组织形式要求如下：

1、原始数据文件夹 data/

2、用户数据文件夹 user\_data/

3、特征工程文件夹 feature/

4、模型训练文件夹 model/

5、预测结果输出文件夹 prediction\_result/

6、代码文件夹 code/

7、解决方案及算法介绍文件 README.md

在README.md（或其他文件格式如doc、pdf等皆可）文件中介绍自己的：

* 解决方案及算法，包含从原始图像到最终结果输出的整个逻辑流程以及算法详情。
* 代码运行说明，包括代码运行入口，若需额外输入参数请自行将参数写入运行代码。
* 若提交的代码在运行时有需要特殊注意的内容，也请在该文件中一并说明。