各位同学，大家好！

我是本学期《数学建模创新实践教育》任课老师王志俊，感谢大家选修该门课程。

本打算召集大家见个面，线下给大家聊一聊这门课的注意事项，现在的形势不允许，所以就建了 QQ 群，改成文字形式吧。

这是一门实践课，没有课堂教学，这些年都是采取“发布任务->解决问题->形成课程报告->评阅报告->给定成绩”的顺序进行的。本学期我们仍沿用此做法。

1) 关于数学建模，我们指导教师团队建立了在线开放课程（学习通邀请码：44157464），请大家自行加入进行学习。

2) 第十九届五一数学建模竞赛（51mcm.cumt.edu.cn）将于 2022 年 5 月 1 日上午 9:00 开始，届时我会指定其中一题作为本门课程题目，请大家组成团队（每队 1-3 名同学）共同解决此题，并形成课程报告。

3) 选择今年五一赛题目作为课程题目并不意味着大家必须参加五一赛。若有同学参加今年五一赛，题目不受前面指定的限制，可以用你们参赛题目作为课程题目。

4) 同一队只需提供一份课程报告。队友既可以是选修本门课程的同学，也可以是其他同学。若为其他同学，其信息（姓名、学号）不必出现在课程报告封面页（本文件第二页）。

5) 课程报告首页为封面页（本文件第二页），第二页为摘要页（本文件第三页），第三页开始为正文。课程报告应同时提交电子档和纸质档，且两者内容须一致。

6) 电子档必须为 pdf 格式，以便与五一赛论文格式匹配，便于查重；电子档以团队成员姓名命名，并于 2022 年 5 月 20 日 24:00 前发送至 wangzj@cumt.edu.cn

7) 纸质档于 2022 年 5 月 22 日 24:00 前交至数学学院 A333.

8) 本人是五一赛评审老师，能够接触所有参赛论文及查重结果，请大家遵守纪律，以团队为单位独立完成课程论文。若发现有同学盗用非本人参与的五一赛参赛

队论文，该同学成绩不及格，该五一赛参赛队（不论是否为本校参赛队）取消评奖资格。

2021-2022 学年第 2 学期

数学建模创新实践教育

课程论文

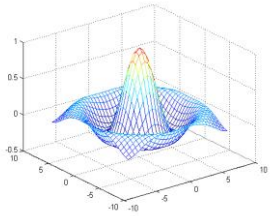
学号：00000000　00000000　06192081

姓名：李春阳　禚佳轩　胡钧耀

教师：王志俊

成绩：

五 一 数 学 建 模 竞 赛



题 目： **矿石加工质量控制问题**

摘　要：vvvv 依额哦 iv 从而 info 测慢慢揣摩二层平面开发，篇 CEO 排名 v 怕了发的，vvesdfv,sdfv,psdfmkpv，v 侧田还给别人说过二公分二通过特工 v 额电话不停有人沟通汇报和不太热输入法二维

关键词：支持向量机、CatBoost 算法、XGBoost 算法，投票法，对抗学习

一、问题背景与重述

1.1 问题背景

在绿色低碳发展的背景下，碳达峰与碳中和的“双碳”的理念越来越受到重视，当前我国已进入新的发展阶段，“双碳”目标紧迫且艰巨，这需要全体社会各界，特别是与环境相关的企业承担起这份责任。对于矿石加工企业来说，优化原矿的加工工艺，从而提高矿石加工合格率，就可以直接或间接地节约不可再生的矿物资源以及加工所需的能源，从而为节能减排助力。矿石加工是一个复杂的过程，在加工过程中，电压、水压、温度作为影响矿石加工的重要因素，直接影响着矿石产品的质量。矿工加工实际流程如下图所示。

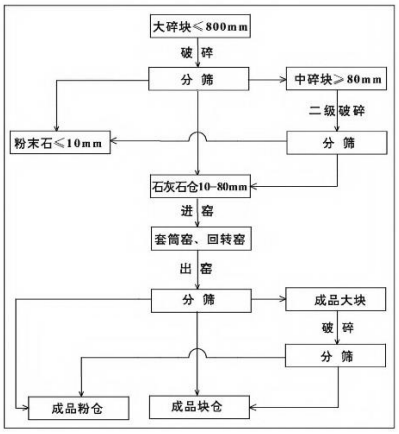


图 1.1-1 矿工加工实际流程

为了方便建模，假设矿石加工过程需要经过系统 I 和系统 II 两个环节，两个环节不分先后，其他条件（电压、水压等）保持不变。简化工艺流程如下图所示。

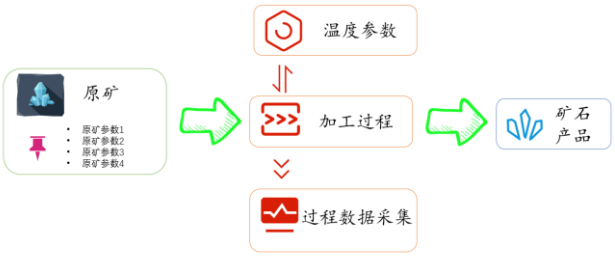


图 1.1-2 矿工加工实际流程

1.2 问题重述

针对该生产车间 2022 年 1 月 13 日至 2022 年 4 月 7 日的生产加工数据，进行数学建模，完成下列问题：

(1) 问题一要求对附件 1 中生产加工数据，研究原矿参数和系统设定温度对产品质量指标之间的关系，建立数学模型，给出利用系统温度预测产品质量的方法，并预测对应温度的产品质量指标。

(2) 问题二要求进一步探究原矿参数，系统设定温度，产品质量指标之间的关系，在问题一的基础上，分析问题一的逆映射，探究产品目标质量所对应的系统温度并预测对应产品质量指标下可能的系统温度。

(3) 问题三要求利用附件 2 的数据，通过增加过程数据，利用过程数据改进问题一模型，给出指定系统设定温度，预测矿石产品合格率的方法，并给出合格率预测结果。

(4) 问题四要求在问题三基础上，进一步分析灵敏度和准确性。判断能否达到 2022 年 4 月 10 日和 2022 年 4 月 11 日产品的合格率要求，如果可以达到，给出系统设定温度。

二、问题分析

2.1 题目整体分析

首先将附件所给的数据进行数据清洗，进行整合，整理出系统温度、产品指标以及原矿参数对应小时的数据样本。处理过程中，我们对一小时内的系统温度取均值作为对应时间的数值，而原矿参数设置为当天内 24 小时原矿参数相同。

对于该问题，我们首先需要观察数据的分布情况，探究数据之间的相关性，挖掘其数据本身的特征。为此我们选取四项指标的检测数据进行绘图。四项指标的趋势图如下图所示，从图中可以看出，参数 A，B，C 的波动范围小，D 波动范围大，指标对外界因素的变换更加敏感，可能更容易受到影响。

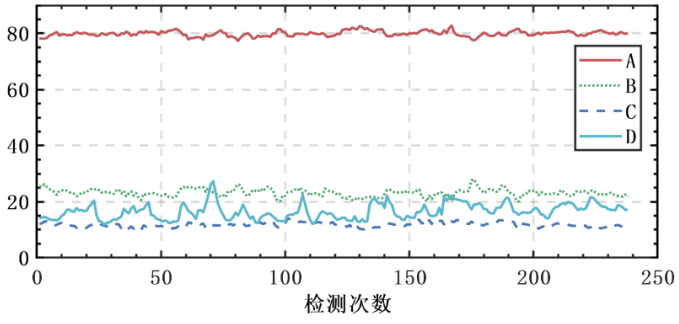


图 2.1-1 四种矿石产品检测指标时序变化图

选取两个系统温度为变量按时序进行绘图。两个系统温度的趋势图如下图所示，可以看出系统温度波动幅度都较大，特别是温度 1，波动范围较广，可能是主要影响矿石产品质量的因素之一。

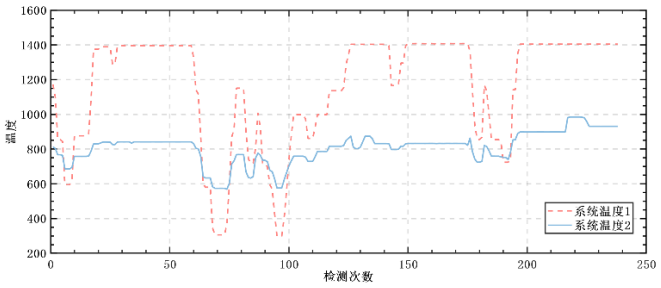


图 2.1-2 两个系统温度的时序变化图

选取四个原矿参数为变量按时序进行绘图。四项原矿参数的趋势图如下图所示，由图可以看出，原矿参数 1 变化幅度较大，原矿参数 2，3 呈现负相关，原矿参数 4 较为平缓。因此，后面重点探究原矿参数 1，2，3 对其结果的影响。

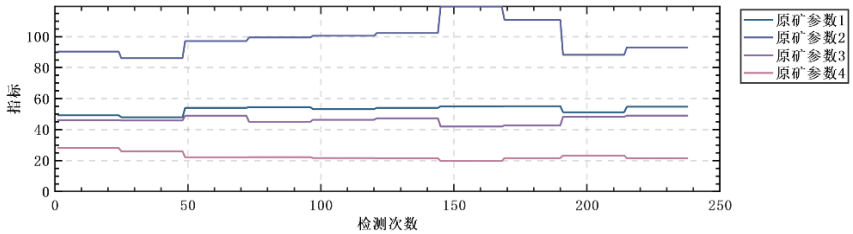


图 2.1-3 四项原矿参数的时序变化图

此外，我们绘制了其变量相关性统计图，如下所示。从图片中可以看出，变量之间相关性不是很明显，其中指标 A，B 负相关，系统温度 1 与系统温度 2 相关性明显，原矿参数之间互相影响。

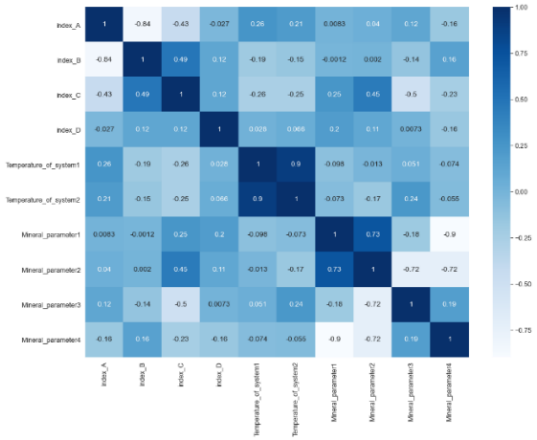


图 2.1-4 变量相关性热力图

为了进一步探究，绘制两两变量之间散点图和核密度估计图。如下图所示，右上角为两两变量之间的散点图，左下角为核密度估计，对角线为变量本身特征。可以发现其变量之间存在多维度影响，难以直接用线性关系进行拟合。

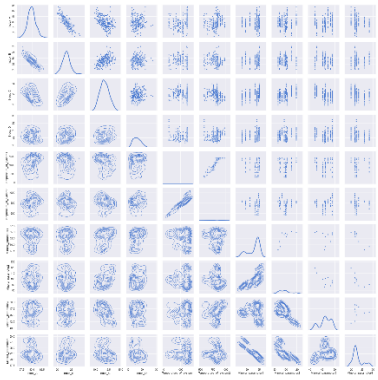


图 2.1-5 两两变量之间散点图和核密度估计图

2.2 问题一的分析

附件 1 给出了以往时间的原矿加工系统温度数据，问题 1 要求在给定原矿参数和系统温度下，给出产品质量预测结果。

根据上面对数据样本的分析，针对样本高维特征和样本数量不多的特点，在问题一中，决定采用支持向量机进行回归预测，可以有效挖掘高维度特征之间的关系，针对问题一的产品质量指标 A，B，C，D 构建多输出支持向量机模型（MSVR）。由于样本数据偏小的原因，为了增强鲁棒性，引入投票法进行平均（Voting），通过搭配不同核函数实现差异化学习，使用投票法后，最终的预测结果是多个回归模型预测结果的平均值。

2.3 问题二的分析

利用附件 1 数据，假设原矿参数和产品目标质量已知，需要估计产品目标质量所对应的系统温度。

通过问题一的分析，我们了解到矿石加工过程中，不同变量和产品质量指标之间的关系，因此在问题二中，可以探究其逆映射，但是针对支持向量机非线性问题的核函数的选择没有通用标准，难以选择一个合适的核函数，同时，问题一多项式核 poly，高斯核函数 rbf 对非线性内核存在误差，因此，引入随机森林发展而来的集成学习（Ensemble Learning），在问题一模型上进行改进。通过对比各类集成学习，最后确定引入 XGBoost 和 CatBoost 两个算法，用以平衡算两个法的性能。进一步完善模型，根据模型可以很轻易求出问题一、二的结果。

2.4 问题三的分析

附件 2 给出了生产车间一段时间内的生产加工数据及过程数据，满足下表销售条件的产品视为合格产品。

问题需要在给定系统设定温度和原矿参数、过程数据的情况下，预测矿石产品合格率。

表 2.4-1 矿石合格标准

指标	指标 A	指标 B	指标 C	指标 D
销售条件	77.78 - 80.33	<24.15	<17.15	<15.62

对附件 2 的数据做同样的数据分析，拥有相同的样本特征，利用给出合格标准，统计样本合格率，结果如下图所示。由观察可知，矿石产品的整体合格率偏低，如果直接利用模型进行预测的产品质量指标按标准进行归类合格，其预测存在一定误差，在误差范围内容易影响结果。设计对抗学习，利用问题二改进的模型进行回归预测，生成产品质量指标，与原产品质量指标进行混合，构建分类模型。利用对抗思想，评估生成的样本数据与真实数据的接近程度作为评估标准，通过对比，合理利用误差，得到在误差允许范围内的分类模型（这里选用随机森林），利用改进的预测模型和对抗分类模型，给出合格率预测结果。

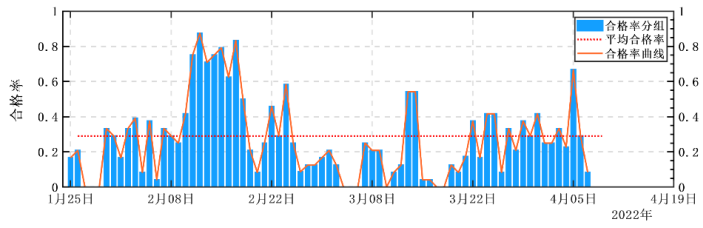


图 2.4-1 产品样本合格率时序图

2.5 问题四的分析

利用附件 2 的数据，给出在指定合格率的条件下，设定系统温度的方法，并对模型的准确性和敏感性进行分析。

由于整体合格率偏低，在 80%以上样本数较少，在前述问题中建立的模型基础上求其逆映射存在困难，这里采取优化算法来寻找结果。构建非线性映射函数，将合格率作为因变量进行研究。因为自变量温度的取值有无限种取值可能，所以对于此全局寻优问题，我们采用启发式算法中的粒子群优化算法（PSO）求解。如果无解，则代表达不到题目所给定合格率要求。

三、模型假设及符号说明

3.1 模型假设

我们已经结合实际情况对本问题进行了系统且具体的分析，根据上述分析我们做出如下合理的假设，以期利用恰当的数学模型对该问题进行详细的解答。

- 假设系统温度与调温指令设定的温度相同；
- 假设每次温度调节后的两个小时不会输入新的调温指令，系统温度变化幅度不大，可以用两小时的平均值来表示这两小时内任一时刻的温度；

3.2 符号说明

表 3.2-1 符号意义表

符号	含义
$E(X)$	数据X的期望
σ_X	数据X的方差
$Cov(X,Y)$	数据X, Y的协方差
ρ_{XY}	数据X, Y的相关性
$P(M)$	事件M发生的概率
$Index$	指标
T	系统温度
Mp	原矿参数
Pp	过程数据
C	惩罚因子
W	$Sigmoid$核函数
MAE	平均绝对误差
R^2_Score	拟合度指数

四、模型的建立与求解

4.1 问题一模型的建立与求解

4.1.1 模型的建立

附件 1 给出了以往时间的原矿加工系统温度数据，问题 1 要求在给定原矿参数和系统温度下，给出产品质量预测结果。

根据上面对数据样本的分析，针对样本高维特征和样本数量不多的特点，在问题一中，决定采用支持向量机进行回归预测，可以有效挖掘高维度特征之间的关系，针对问题一的产品质量指标 A，B，C，D 构建多输出支持向量机模型（MSVR）。由于样本数据偏小的原因，为了增强鲁棒性，引入投票法进行平均

（Voting），通过搭配不同核函数实现差异化学习，使用投票法后，最终的预测结果是多个回归模型预测结果的平均值。

4.1.1.1 支持向量机

支持向量机（Support Vector Machine，SVM）是一个功能强大并 且全面的机器学习模型，它能够执行线性或非线性分类、回归，甚至 是异常值检测任务。它是机器学习领域最受欢迎的模型之一。

4.1.1.2 多输出支持向量机

4.1.1.3 投票法

如果所有分类器都能够估算出类别的概率（即有 predict_proba（）方法），那么你可以将概率在所有单个分类器上平 均，然后让 Scikit-Learn 给出平均概率最高的类别作为预测。这被称为 软投票法。通常来说，它比硬投票法的表现更优，因为它给予那些高度 自信的投票更高的权重。

4.1.2 模型的求解

MSE:

1 0.4723346929506919

2 0.8289606288947904

3 0.28831513595446234

4 2.0489838374249176

MAE:

1 0.5055064155858265

2 0.6568059898649369

3 0.4135665874203815

4 0.9580151293228684

可解释的方差分数:

1 0.4778962771998032

2 0.5393486523104958

3 0.8121783178100314

4 0.8039681153773822

r2_score:

1 0.4700354613730384

2 0.5362651784982123

3 0.8108728729727801

4 0.803788072638778

mean_absolute_percentage_error:

1 0.006383776911476698

2 0.027776726419015237

3 0.037487449788125876

4 0.06517875501938798

表 4.1-1 问题 1 结果

时间	系统 I 设定温度	系统 II 设定温度	指标 A	指标 B	指标 C	指标 D
2022-01-23	1404.89	859.77				
2022-01-23	1151.75	859.77				

4.2 问题二模型的建立与求解

4.2.1 模型的建立

利用附件 1 数据，假设原矿参数和产品目标质量已知，需要估计产品目标质量所对应的系统温度。

通过问题一的分析，我们了解到矿石加工过程中，不同变量和产品质量指标之间的关系，因此在问题二中，可以探究其逆映射，但是针对支持向量机非线性问题的核函数的选择没有通用标准，难以选择一个合适的核函数，同时，问题一多项式核 poly，高斯核函数 rbf 对非线性内核存在误差，因此，引入随机森林发展而来的集成学习（Ensemble Learning），在问题一模型上进行改进。通过对比各类集成学习，最后确定引入 XGBoost 和 CatBoost 两个算法，用以平衡两个算法的性能。进一步完善模型，根据模型可以很轻易求出问题一、二的结果。

4.2.1.1 集成学习

你聚合一组预测器（比如分类器 或回归器）的预测，得到的预测结果也比最好的单个预测器要好。这 样的一组预测器称为集成，所以这种技术也被称为集成学习，而一个 集成学习算法则被称为集成方法。

4.2.1.2 随机森林

随机森林是决策树的集成，通常用 bagging（有时是 pasting）方法训练，

4.2.1.3 XGBoost

提升法（boosting，最初被称为假设提升）是指可以将几个弱学习器结合成一个强学习器的任意集成方法。大多数提升法的总体思路是循环训练预测器，每一次都对其前序做出一些改正。流行的 Python 库 XGBoost（该库代表 ExtremeGradient Boosting）中提供了梯度提升的优化实现，该软件包最初是由 Tianqi Chen 作为分布式（深度）机器学习社区（DMLC）的一部分开发的，其开发目标是极快、可扩展和可移植。

4.2.1.4 CatBoost

4.2.1.5 平衡算法的性能

4.2.2 模型的求解

表 4.2-1 问题 2 结果

时间	指标 A	指标 B	指标 C	指标 D	系统 I 设定温度	系统 II 设定温度
2022-01-24	79.17	22.72	10.51	17.05		
2022-01-24	80.10	23.34	11.03	13.29		

4.3 问题三模型的建立与求解

4.3.1 模型的建立

该问题的需求是，需要在给定系统温度、原矿参数、过程数据的情况下，预测矿石产品合格率。这里需要计算合格率，还加上了过程参数为新的自变量。

由于矿石产品的整体合格率偏低，直接使用之前的预测模型预测产品质量指标按标准进行归类，其预测存在一定误差。本题模型的思路是设计对抗学习，利用问题二改进的模型进行回归预测，生成虚拟的产品质量指标，与原产品质量指标进行混合，构建分类模型。最后评估生成样本和真实样本的相似度，作为合格标准，得到在误差允许范围内的随机森林分类模型。

4.3.1.1 生成器与对抗学习思想

生成器

以随机分布作为输入（通常是高斯分布），并输出一些数据（通常是图像）。你可以将随机输入视为要生成的图像的潜在表征（即编码）。因此你可以看到，生成器提供的功能与变分自动编码器中的解码器相同，并且可以使用相同的方式来生成新图像（只需馈入一些高斯噪声，就会输出一个新图片）。但是我们很快就会看到，它的训练方式大不相同。

4.3.1.2 修正随机森林分类模型

4.3.2 模型的求解

表 4.3-1 问题 3 结果

时间	系统 I 设定温度	系统 II 设定温度	合格率
2022-04-08	341.40	665.04	80%
2022-04-09	1010.32	874.47	99%

4.4 问题四模型的建立与求解

4.4.1 模型的建立

利用附件 2 的数据，给出在指定合格率的条件下，设定系统温度的方法，并对模型的准确性和敏感性进行分析。

由于整体合格率偏低，在 80%以上样本数较少，在前述问题中建立的模型基础上求其逆映射存在困难，这里采取优化算法来寻找结果。构建非线性映射函数，将合格率作为因变量进行研究。因为自变量温度的取值有无限种取值可能，所以对于此全局寻优问题，我们采用启发式算法中的粒子群优化算法（PSO）求解。如果无解，则代表达不到题目所给定合格率要求。

4.4.1.1 粒子群算法（PSO）

粒子群优化算法是一种基于群体物种或者粒子研究的随机优化方法，其思想来源于人工生命和演化计算理论。

基本的全局 PSO 算法如下所述。在以D维度的目标搜索空间中，有M个粒子组成一个粒子集群，其中第i个粒子表示为一个D维的向量$\boldsymbol{x}_i=(x_{i1},\ x_{i2},\ ...,\ x_{iD}),\ i=1,\ 2,\ ...,\ m$，即第$i$个粒子表示为$D$维的搜索空间中的位置记为$\boldsymbol{x}_i$，每个粒子所处的坐标都可以认为是可能的解，将$\boldsymbol{x}_i$带入设计好的目标函数，即可计算出这个粒子在当前坐标的适应值，根据适应值的大小衡量$\boldsymbol{x}_i$的优劣。第i个粒子的移动速度或者步长表示为一个D维的向量$\boldsymbol{v}_i=(v_{i1},\ v_{i2},\ ...,\ v_{iD})$。对于最优位置有两种，一种是第$i$个粒子自身搜索到的最优位置，第二种是整个群体搜索到的最优位置，分别为$\boldsymbol{p}_i=(p_{i1},\ p_{i2},\ ...,\ p_{iD})$和$\boldsymbol{p}_g=(p_{g1},\ p_{g2},\ ...,\ p_{gD})$。对粒子执行如下操作：

$$\begin{aligned} v_{id} &= v_{id}+c_1r_1(p_{id}-x_{id})+c_2r_2(p_{gd}-x_{gd}) \\ x_{id} &= x_{id}+v_{id} \end{aligned} \tag{1}$$

其中$i=1,\ 2,\ ...,\ m,\ d=1,\ 2,\ ...,\ D$；学习因子$c_1$和$c_2$是非负常数；$r_1$和$r_2$是介于$[0,1]$之间的随机数；$v_{id}\in[-v_{\max},v_{\max}]$，$v_{\max}$是自行设定的常数。如果迭代达到最大次数或者粒子群搜索到的最优值达到了设定的阈值，就结束算法。

4.4.2 模型的求解

表 4.4-1 问题 4 结果

时间	合格率	能否达到	系统 I 设定温度	系统 II 设定温度
2022-04-10	80%	是		
2022-04-11	99%	否	/	/

五、模型的评价与推广

5.1 模型的优点

5.1.1 问题一、二模型的优点

1. 非线性映射是 SVM 方法的理论基础,SVM 利用内积核函数代替向高维空间的非线性映射，样本量不是海量数据的时候，准确率高，泛化能力强。

2. 投票法是一种遵循少数服从多数原则的集成学习模型，通过多个模型的集成降低方差，从而提高模型的鲁棒性和泛化能力。

3. XGBoost 的决策树是 Level-wise 增长。Level-wise 可以同时分裂同一层的叶子，容易进行多线程优化，过拟合风险较小

4. Catboost 采用的策略在降低过拟合的同时保证所有数据集都可用于学习。

5.1.2 问题三模型的优点

1. 引入对抗学习 GAN 网络的思想，以半监督方式训练分类器，使得在误差范围能更准确的预测结果。

2. 对抗学习使得数据增强，具有更广的适用性。

3. 随机森林利用构建决策树，可以处理非线性特征且考虑了变量之间的相互作用。

5.1.3 问题四模型的优点

1. 粒子群算法方便简单且易于实现，对比其他搜索算法，其速度更快，方便易操作。

5.2 模型的缺点

5.2.1 问题一二模型的缺点

1. 模型较为复杂，通过投票法使得模型更加稳定的同时也产生了大量冗余，造成模型训练速度慢。

2. 其过拟合效果联合，可数据样本小时，如问题一样本数据远小于问题三，

使得部分模型欠拟合问题，只能减少测试集合，扩大训练集。

3. 由于 XGBoost 和 CatBoost 与支持向量机特性不同，在鲁棒性增强的同时，准确性有所下降。

5.2.2 问题三模型的缺点

1.对于特征的处理需要大量的内存和时间；

2.对抗训练学习了假的数据，存在误差影响，如果前面模型误差太大会导致误差进一步放大，影响结果。

5.2.3 问题四模型的缺点

1 有时粒子群在俯冲过程中会错失全局最优解；

2 应用 PSO 算法处理高度复杂问题时，算法可能过早收敛；

3 PSO 算法是一种概率算法，搜索过程带有随机性。

5.3 模型的推广或改进

1. 模型二对模型一进行了改进，弥补了支持向量机的曲线，也融合了 xgboost 和 catboost 的优点，在牺牲一部分准确性的情况下，使得模型更加稳定，对此可以更容易的利用在其他问题上。

2. 针对问题三中对抗学习，可以引入对抗网络和变分自编码器进行模拟，可以增强其模型稳定性和准确率，但也需要更多的样本数据，问题三中数据太少，无法构建神经网络进行预测。

3.在问题四中我们采用了 PSO 算法，在启发式算法中，还可以升级为人工鱼群法，加入觅食行为，聚群行为，追尾行为，随机行为等控制，提高搜索算法的精确度。

六、参考文献

[1]

七、附录

附录 1
介绍：该代码是某某语言编写的，作用是什么