机器学习工具箱V2版本

**概要方案设计**



苏州同元软控信息技术有限公司

|  |  |  |  |
| --- | --- | --- | --- |
| 编制 | 袁友胜 | 生效日期 |  |
| 审核 |  | 批准 |  |

文件变更摘要

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **日期** | **版本** | **变更说明** | **修订** | **审核** | **批准** |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

目录

[1. 引言 1](#_Toc128407602)

[1.1 概述 1](#_Toc128407603)

[1.2 参考资料 1](#_Toc128407604)

[2. 需求分析 2](#_Toc128407605)

[2.1 功能要求 2](#_Toc128407606)

[2.1.1 聚类分析 2](#_Toc128407607)

[2.1.2 回归分析 2](#_Toc128407608)

[2.1.3 分类 2](#_Toc128407609)

[2.1.4 降维与特征提取 3](#_Toc128407610)

[2.1.5 工业统计 3](#_Toc128407611)

[2.2 运行环境要求 3](#_Toc128407612)

[3. 系统设计 4](#_Toc128407613)

[3.1 总体框架 4](#_Toc128407614)

[3.2 应用流程 5](#_Toc128407615)

[3.3 功能设计 5](#_Toc128407616)

[3.3.1 数据预处理 5](#_Toc128407617)

[3.3.2 模型构建 6](#_Toc128407618)

[3.3.3 模型评估 6](#_Toc128407619)

[3.3.4 神经网络 6](#_Toc128407620)

[3.4 预期效果 7](#_Toc128407621)

[4. 技术方案 7](#_Toc128407622)

[4.1 产品研发方案 7](#_Toc128407623)

[5. 应用验证 10](#_Toc128407624)

[5.1 基础案例简介 10](#_Toc128407625)

[5.1.1 使用t-SNE可视化高维数据 10](#_Toc128407626)

[5.1.2 逻辑回归模型的贝叶斯分析 10](#_Toc128407627)

[5.1.3 多元一般线性模型 11](#_Toc128407628)

[5.1.4 通过装袋决策树进行信用评级 11](#_Toc128407629)

[5.1.5 代理分裂 12](#_Toc128407630)

[5.2 综合案例简介 13](#_Toc128407631)

[5.2.1 机器学习模型与深度学习结合的应用场景设想： 13](#_Toc128407632)

[5.2.2 从RNA-Seq数据中识别差异表达基因： 13](#_Toc128407633)

[5.2.3 时间序列回归I:线性模型： 13](#_Toc128407634)

[5.2.4 时间序列回归III:有影响的观察： 14](#_Toc128407635)

[5.2.5 使用条件GAN生成合成信号 15](#_Toc128407636)

[6. 研发计划 16](#_Toc128407637)

[6.1 总体计划 16](#_Toc128407638)

[6.2 函数开发计划 17](#_Toc128407639)

[6.3 应用案例开发计划 17](#_Toc128407640)

# 引言

## 概述

Syslab机器学习工具箱V1基对标Matlab机器学习工具箱，已开发基本功能函数99个。主要覆盖包括分类模块中分类器中朴素贝叶斯分类器，决策树分类器，判别分析，支持向量机分类器，半监督分类器，最近邻分类器，以及模型诊断和评估中交叉熵，交叉预测，重要性评估，交叉训练模型等，在回归模块中我们覆盖包括多元线性回归分析，广义线性回归分析，回归树分析，支持向量机回归分析，线性混合效应分析，多项式回归分析，非线性回归分析等，在聚类模块中我们覆盖包括最近邻聚类分析，kmeans聚类分析，层次聚类分析，隐马尔可夫模型分析，kd树模型分析，簇的可视化等，在降维模块中，我们主要覆盖主成分分析，因子分析，非负矩阵分解，普氏分析等，在工业统计中我们主要覆盖试验设计，生存时间数据分析，统计控制过程等。

机器学习工具箱V2基于V1版本继续开发，对核心功能降维分析，广义回归，模型评估，半监督等模块进行功能扩充，增加集成学习，增量学习，神经网络等核心功能。进一步完善机器学习工具箱的功能。其中包括流形学习，广义加性模型，线性鲁棒模型。混合效应，数据的预处理，参数优化等核心问题。计划在有监督、半监督和降维领域引入综合案例的开发应用。并至少完成215个函数开发指标。

## 参考资料

1. scipy官方文档

[Resampling and Monte Carlo Methods — SciPy V1.10.1 Manual](https://docs.scipy.org/doc/scipy/tutorial/stats/resampling.html)

1. statsmodels官方文档

[Installing statsmodels — statsmodels](https://www.statsmodels.org/stable/install.html)

1. scikit-learn官方文档

[User guide: contents — scikit-learn 1.2.1 documentation](https://scikit-learn.org/stable/user_guide.html)

1. MatLab机器学习工具箱帮助

[Statistics and Machine Learning Toolbox — 示例 (mathworks.cn)](https://ww2.mathworks.cn/help/stats/examples.html?category=index&exampleproduct=all&s_tid=CRUX_lftnav)

# 需求分析

## 功能要求

覆盖分类、回归、聚类、降维和特征提取等功能具备对标产品机器学习工具箱函数数量（269个）80%以上 (不少于215个)。

### 聚类分析

将物理或抽象对象的集合分成由类似的对象组成的多个类的过程被称为聚类。由聚类所生成的簇是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其他簇中的对象相异。“物以类聚，人以群分”，在自然科学和社会科学中，存在着大量的分类问题。聚类分析又称群分析，它是研究（样品或指标）分类问题的一种统计分析方法。聚类分析起源于分类学，但是聚类不等于分类。聚类与分类的不同在于，聚类所要求划分的类是未知的。

### 回归分析

回归分析是一种数学模型，当因变量和自变量为线性关系时，它是一种特殊的线性模型。最简单的情形是一元线性回归，由大体上有线性关系的一个自变量和一个因变量组成。若有k个自变量和1个因变量，则因变量的值分为两部分：一部分由自变量影响，即表示为它的函数，函数形式已知且含有未知参数；另一部分由其他的未考虑因素和随机性影响，即随机误差。当函数为参数未知的线性函数时，称为线性回归分析模型；当函数为参数未知的非线性函数时，称为非线性回归分析模型。当自变量个数大于1时称为多元回归。

### 分类

分类是一个有监督的学习过程，目标数据库中有哪些类别是已知的，分类过程需要做的就是把每一条记录归到对应的类别之中。由于必须事先知道各个类别的信息，并且所有待分类的数据条目都默认有对应的类别，因此分类算法也有其局限性，当上述条件无法满足时，我们就需要尝试聚类分析。

### 降维与特征提取

在数据的预处理阶段，特征提取和数据降维是提升模型表示能力的一种重要手段。特征提取主要是从数据中找到有用的特征，用于提升模型的表示能力，而数据降维主要是在不减少模型准确率的情况下减少数据的特征数量。

### 工业统计

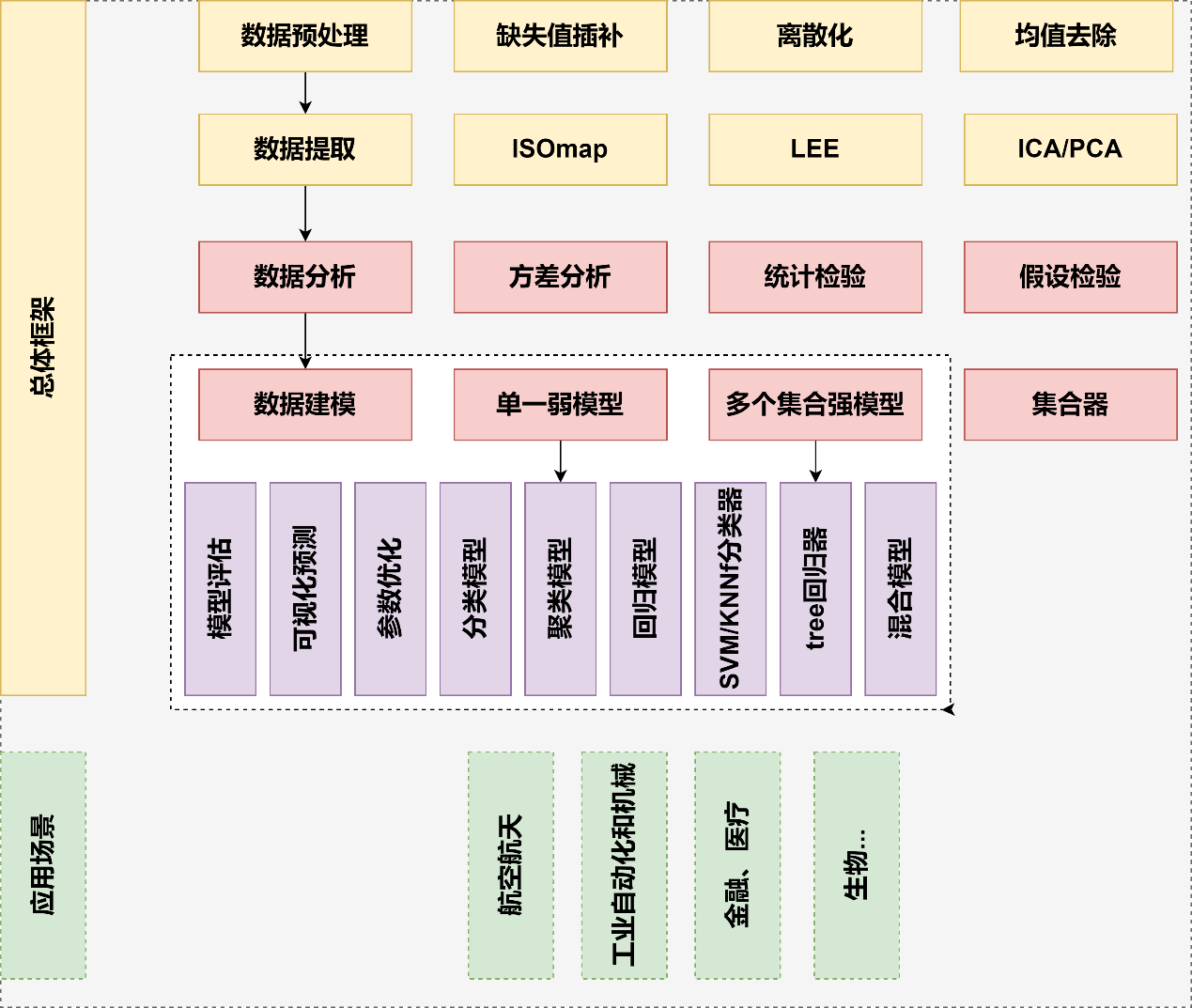
对工业经济现象数量方面的资料进行搜集、整理、分析、研究工业经济宏观发展与微观生产经营活动的数量关系，反映工业经济活动发展规律性的统计方法论和实务专业统计。

## 运行环境要求

|  |  |
| --- | --- |
| 系统内存 | 大于等于8G |
| 剩余存储空间 | 大于等于80G |
| 操作系统 | Windows 10以上或centos7.9以上 |
| 软件系统 | Julia1.7以上。Python3.9及以上 |

# 系统设计

## 总体框架



Syslab 机器学习工具箱总体框架设计分为数据预处理、数据提取、数据分析、数据建模等过程。

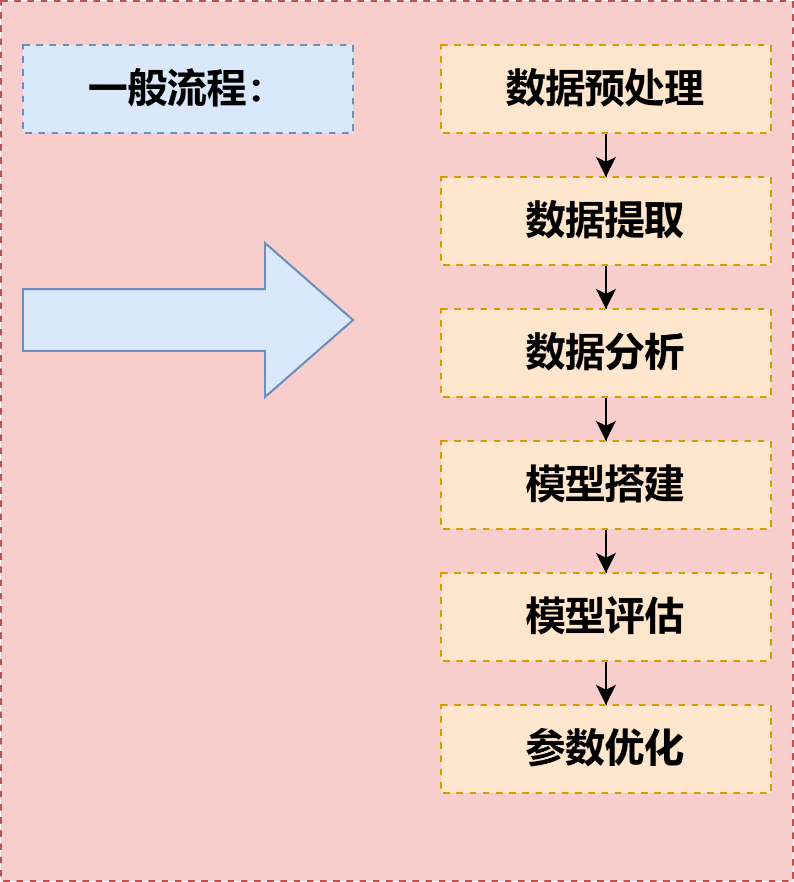
数据预处理结合标准化，均值去除，方差缩放，非线性变换离散化，缺失值的插补等方法。

数据提取结合ISOmap、LEE、PCA、ICA等主要方法。

数据分析结合方差分析，假设检验，统计检验等统计学方法对数据进行关联性分析。

模型建模中输入层，支持数学Array、Vector和DataFrame等Julia形式数据导入；接着参数优化、模型评估、输出可视化等步骤。应用场景主要在航空航天、工业自动化和机械、金融、医疗设备、生物等。

## 应用流程



## 功能设计

### 数据预处理

在现实生活问题中，我们得到的原始数据往往非常混乱、不全面，机器学习模型往往无法从中有效识别并提取信息。数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已，在采集完数据后，机器学习建模的首要步骤以及主要步骤便是数据预处理。

真实场景中可能存在的问题：

1. 各特征(变量)的尺度(量纲)和数量级差异大
2. 存在噪声：包含错误和异常值
3. 存在缺失值
4. 存在冗余特征

…

因此我们也提供了一些处理方法，标准化、均值去除、方差缩放、缺失值处理、缺失值的插补、离散化等。其中降维模块我们在V1版本主成分分析、因子分析、非负矩阵分解等基础方法上增加独立成分分析、流形学习等。

### 模型构建

我们在模型构建方面在V1版本的多元线性模型、多项式回归模型、线性混合效应模型、广义线性模型等增加了线性鲁棒模型、广义加性模型、广义混合效应，由于单个模型性能的局限性，也进一步引入集成学习，组合这里的多个弱监督模型以期得到一个更好更全面的强监督模型，集成学习潜在的思想是即便某一个弱分类器得到了错误的预测，其他的弱分类器也可以将错误纠正回来。单个学习器我们称为弱学习器，相对的集成学习则是强学习器。我们借鉴了bagging、AdaBoost、Xgboost、GBDT等核心算法逻辑。

### 模型评估

评价指标针对不同的机器学习任务有不同的指标，同一任务也有不同侧重点的评价指标。主要有分类、回归、聚类等。

1. 分类模型评价指标：

其中包括准确率、混淆矩阵、精准率、召回率、F1-score、ROC曲线。

1. 回归模型评价指标：

平方根误差RMSE,均方误差、r2-score、部分依赖图、主体成分重要性。

1. 方差分析与假设检验：

单因素方差分析、多因素方差分析、以及新增多元方差检验、拒绝原假设检验等。

### 神经网络

此实现不适用于大规模应用程序。对于更快的基于 GPU 的实现，以及为构建深度学习提供更大灵活性的框架体系结构。请参阅syslab深度学习工具箱。

借鉴scikit-learn多层感知机的核心算法实现一种监督学习算法，回归或分类，可以存在一个或多个非线性层。

## 预期效果

对核心功能进行增强，补充机器学习工具箱算法模型，主要集中在四个方面，集成学习：参考Bagging、AdaBost、Histogram-Based GradientBoosting等集成方法。流形学习我们引入ISOmap、LLE、ICS、TSNE、LE等核心算法，在稳定性方面我们新构建robustfit等函数，以及广义加性模型等，增加数据预处理功能形成数据预处理-模型构建-模型预测-模型评估-参数优化-增量学习-集成学习-模型评估生态闭环。

本产品以MWORKS中的工具箱的形式存在，功能亮点包括但不限于：

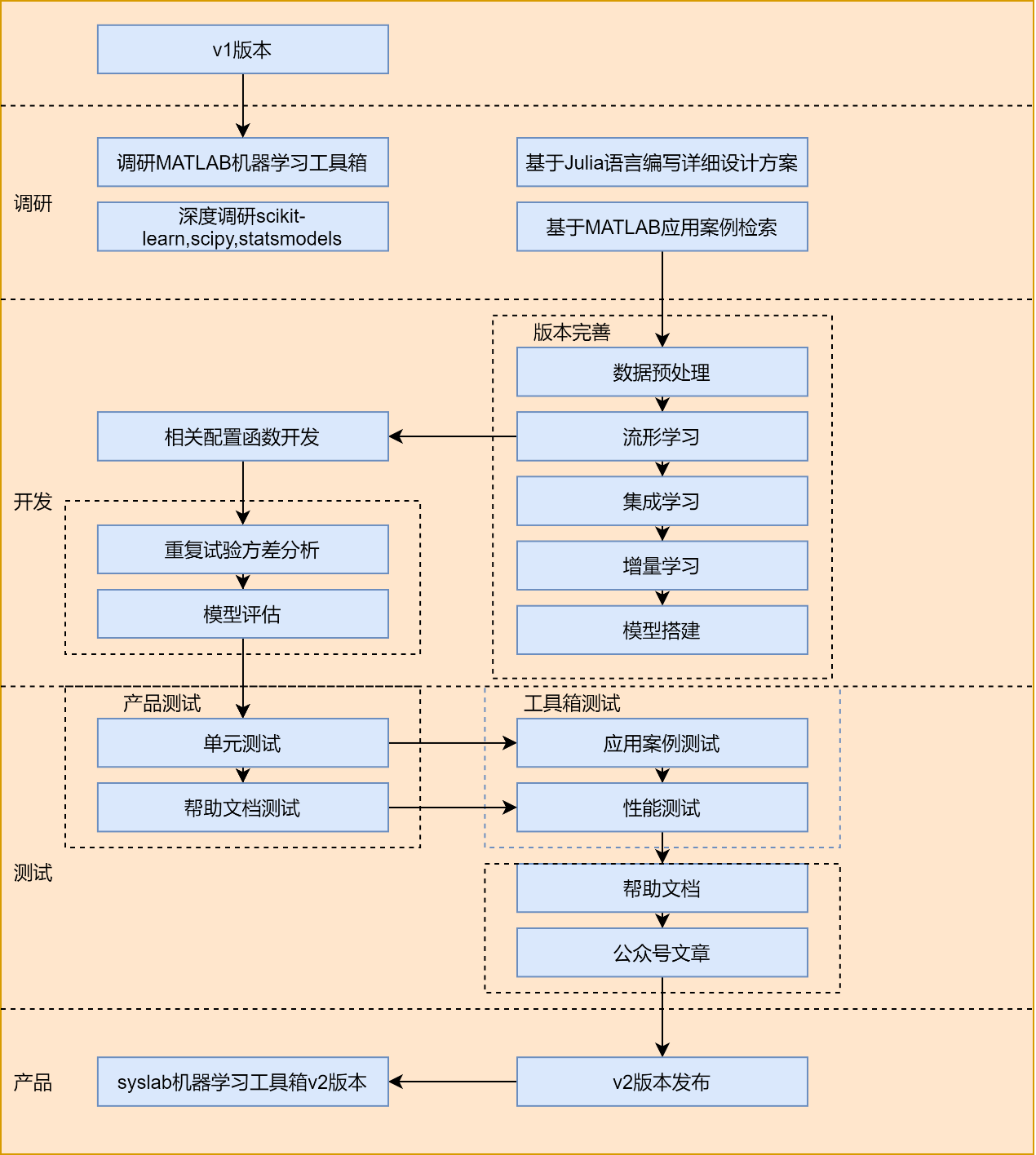
1. 相较于V1版本，基础功能更为完善，支持更丰富的函数库，更多机器学习训练模型。
2. 提供较为完整的数据预处理模块，达到数据预训练标准。
3. 多样化的数据提取方法，使得模型构建更为严谨。
4. 由弱模型到强模型的完善，提供更强大稳定的模型构建方法。
5. 提供较为完整的数据分析方法，使得数据更有解释性。
6. 提供更完整的模型评估方法

# 技术方案

## 产品研发方案

Syslab 机器学习工具箱产品路线图将机器学习工具箱研发分为三个阶段，第一阶段基于Julia开源库以及scikit-learn、Statsmodels、scipy等机器学习开源库，仿照Matlab机器学习工具箱相关的功能函数进行开发，完成回归、分类、聚类、降维相关函数开发。第二阶段主要完成机器学习工具箱V1版本应用验证，并完成综合应用案例的开发。第三阶段，对已完成的V1版本进行代码优化，依据客户使用反馈进行改进，提高算法的运行速度与功能的易用性，发布机器学习工具箱V2版本。

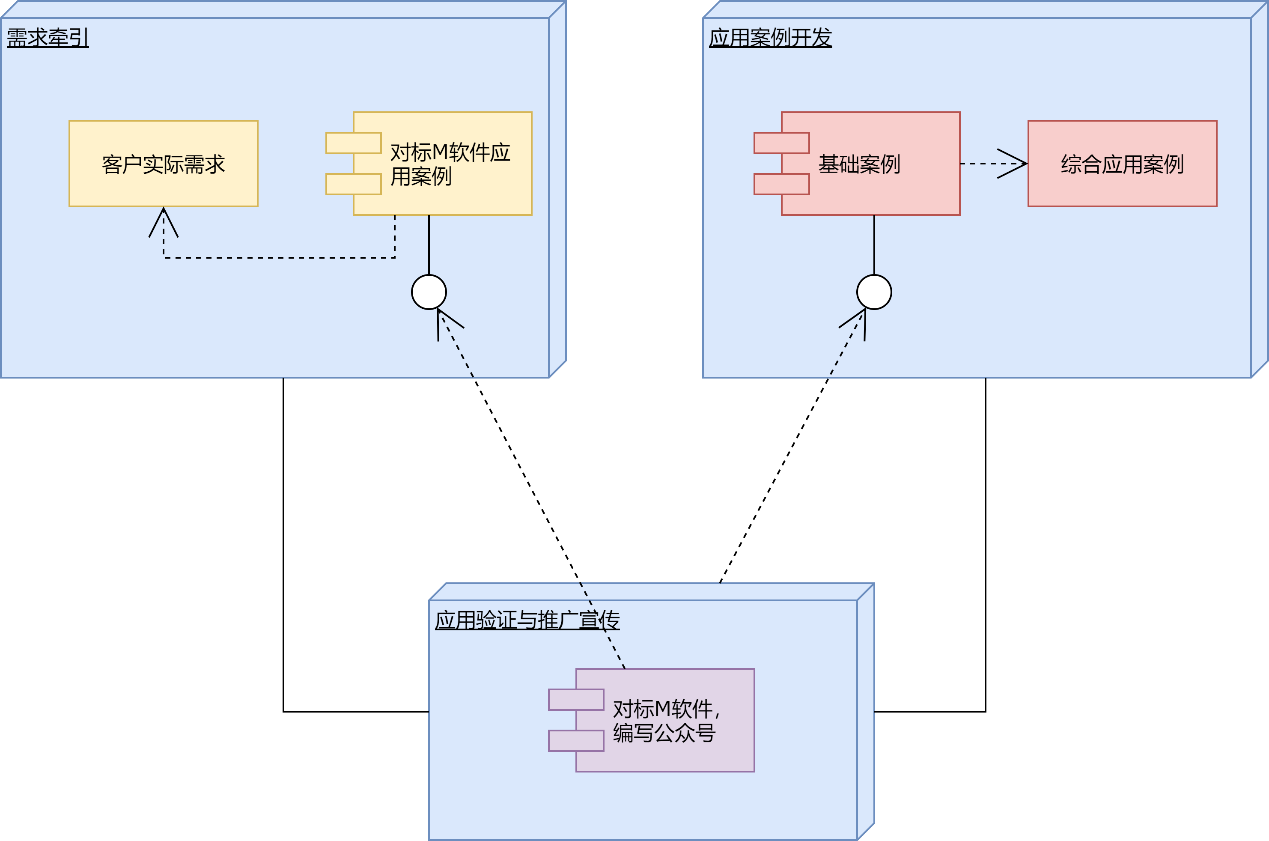
示意图如下：



机器学习工具箱V2版本，根据“调研🡪开发🡪测试🡪产品”开发流程，核心工作可分为4大块：

1. 前期调研，调研Matlab机器学习工具箱，Julia机器学习库、scikit-learn、statsmodels、scipy等机器学习库。完成Syalab机器学习工具箱V2版本开发方案，为接下来的开发工作提供依据和指导。
2. 版本完善，指对V1版本的进一步完善，包括增量学习，流形学习，集成学习，模型诊断评估，数据预处理和bug修复等工作。
3. 函数开发，包括robustfit、fitrgam和 LogitGam等相关配置函数开发。
4. 工具箱验证，包括性能验证和应用验证。性能验证指Syslab 机器学习工具箱内部求解器性能对比，和与 Matlab 对比；应用验证指针对经典案例或实际问题的验证。

# 应用验证



## 基础案例简介

### 使用t-SNE可视化高维数据

这个例子展示了如何使用tsne函数可视化MNIST数据[1]，它由手写数字的图像组成。这些图像的灰度为28 \* 28像素。每个图像都有一个从0到9的关联标签，这是图像所代表的数字。tsne将数据的维数从原始的784维减少到使用PCA的50维，然后使用t-SNE Barnes-Hut算法减少到2或3维。

### 逻辑回归模型的贝叶斯分析

此示例说明如何使用 slicesample 对逻辑回归模型进行贝叶斯推断。

统计推断通常基于最大似然估计 (MLE)。MLE 选择能够使数据似然最大化的参数，是一种较为自然的方法。在 MLE 中，假定参数是未知但固定的数值，并在一定的置信度下进行计算。在贝叶斯统计中，使用概率来量化未知参数的不确定性，因而未知参数被视为随机变量。

### 多元一般线性模型

这个例子展示了如何使用MLR建立一个多元的一般线性模型进行估计。

该数据包含了1985年以来205辆进口汽车的测量样本。

在这里，建模城市和公路MPG的二元响应(列14和列15)。

对于预测因素，请使用轴距(第3列)、整备重量(第7列)和燃料类型(第18列)。前两个预测器是连续的，在本例中是居中和缩放的。燃料类型是一个类别变量，有11和20两个类别，因此需要一个虚拟指标变量进行回归。

### 通过装袋决策树进行信用评级

这个例子展示了如何构建一个自动化的信用评级工具。

信用风险管理的基本任务之一是确定借款人的信用等级。评级是根据客户的信用等级来对其进行排名的:评级越高意味着客户风险越低;相似的等级意味着相似的风险水平。等级分为两类:信用评级和信用评分。信用评级是一小部分独立的类别，通常用字母标记，如“AAA”，“BB-”等。信用评分是数字等级，如“640”或“720”。信用等级是监管框架的关键要素之一，如巴塞尔协议II(见巴塞尔银行监管委员会[3])。

确定信用等级需要分析借款人的信息。如果借款人是个人，相关信息可能是个人收入、未偿还债务(抵押贷款、信用卡)、家庭规模、居住状况等。对于公司借款人，可以考虑某些财务比率(例如销售额除以总资产)、行业等等。这些关于借款人的信息被称为特征或预测因子。不同的机构使用不同的预测因子，他们也可能有不同的评级类别或评分范围来对客户进行排名。对于向大量潜在借款人提供的相对小额贷款(例如，信用卡)，通常使用信用评分，并且对借款人进行评级的过程通常是自动化的。对于中小企业和大型企业的大额贷款，通常使用信用评级，评级过程可能涉及自动算法和专家分析的结合。

有评级机构跟踪公司的信用状况。然而，大多数银行都制定了一种内部方法，为客户分配信用等级。如果客户没有被评级机构评级，内部评级可能是必要的，但即使存在第三方评级，内部评级也可以提供对客户风险状况的补充评估。

这个例子展示了syslab如何帮助信用评级过程的自动化阶段。特别地，这个示例利用了统计和机器学习工具箱™中现成的统计学习工具之一，即被称为袋装决策树的分类算法。

本例假设历史信息以数据集的形式提供，其中每个记录包含借款人的特征和分配给它的信用评级。这些评级可能是内部评级，由一个遵循现有政策和程序的委员会分配。或者，评级可能来自评级机构，其评级被用于“启动”一个新的内部信用评级系统。

现有的历史数据是起点，它被用来训练袋装决策树，将自动化的信用评级。在统计学习的词汇中，这个训练过程属于监督学习的范畴。然后使用分类器为新客户分配评级。在实践中，这些自动或预测的评级很可能被视为试验性的，直到由专家组成的信用委员会对其进行审查。本例中使用的分类器类型还可以促进这些评级的修订，因为它为预测的评级(分类分数)提供了确定性度量。

在实践中，首先需要训练分类器，然后使用它为新客户分配信用评级，最后还需要分析或评估分类器的质量或准确性，这个过程也称为验证或反向测试。本例还讨论了一些可用的回测工具。

### 代理分裂

当缺少某个观察结果的最佳分割预测器值时，如果指定使用代理分割，则软件将使用最佳代理预测器将观察结果发送到左或右子节点。当你有缺失的数据时，树和具有代理分割的树的集合可以提供更好的预测。这个例子展示了如何通过使用带有代理分割的决策树来提高对缺失值数据的预测准确性。

## 综合案例简介

我们会根据机器学习与深度学习相结合在各个实际场景中应用设计。

### 机器学习模型与深度学习结合的应用场景设想：

这个例子展示了如何用机器学习和深度学习方法对雷达回报进行分类。机器学习方法采用小波散射特征提取与支持向量机相结合的方法。此外，还阐述了两种深度学习方法:使用SqueezeNet的迁移学习和长短期记忆(LSTM)循环神经网络。请注意，本例中使用的数据集不需要高级技术，但这里描述了工作流，因为这些技术可以扩展到更复杂的问题。

目标分类是现代雷达系统的一项重要功能。本例使用机器和深度学习对来自圆柱体和锥体的雷达回波进行分类。虽然本例使用的是合成的I/Q样本，但该工作流程适用于真实的雷达返回。

### 从RNA-Seq数据中识别差异表达基因：

RNA-Seq数据的典型差异表达分析包括对原始计数进行归一化，并进行统计检验以拒绝或接受原假设，即两组样本在基因表达上没有显着差异。这个例子展示了如何检查原始计数数据的基本统计数据，如何确定计数归一化的大小因子，以及如何使用负二项式模型推断表达差异最大的基因。

本例的数据集包括Brooks等人描述的实验中获得的RNA-Seq数据。[1]。作者研究了siRNA敲除pasilla的影响，pasilla是一个已知在果蝇剪接调控中起重要作用的基因。该数据集包括2个控制(未处理)样本的生物重复和2个敲除(处理)样本的生物重复。

### 时间序列回归I:线性模型：

这个例子介绍了多元线性回归模型背后的基本假设。这是关于时间序列回归的一系列例子中的第一个，为所有后续例子提供了基础。

时间序列过程通常由多元线性回归(MLR)模型描述，其形式为:

*yt*=*Xtβ*+*et*,

在*yt*是观察到的响应和*Xt*包括可观察预测因子的同期值列。β中的偏回归系数表示个体预测因子对*yt*变化的边际贡献*et*当所有其他的预测都是固定的。

*et*项*yt*的预测值和观测值之间的差异是否全面*et* 这些差异是由于过程波动(β的变化)，测量误差(*Xt*和*yt的*变化)，以及模型规范错误(例如，遗漏了预测因子或*Xt*和*yt* 它们还来自底层数据生成过程(DGP)的固有随机性，而模型试图表示这种随机性。通常假设*et*是由一个不可观测的具有平稳协方差的创新过程产生的与其他社会科学一样，经济数据通常是通过被动观察收集的，不需要对照实验的帮助。理论上相关的预测因子可能需要被实际可用的代理所取代。反过来，经济观察可能具有有限的频率、低可变性和强烈的相互依赖性。

这些数据缺陷导致OLS估计的可靠性和应用于模型规范的标准统计技术存在许多问题。系数估计可能对数据测量误差敏感，使得显著性检验不可靠。多个预测因子的同时变化可能会产生难以分离为单个影响的相互作用。反应中观察到的变化可能与预测因子中观察到的变化相关，但不是由其引起的。

在可用数据的上下文中评估模型假设是规范分析的目标。当一个模型的可靠性变得可疑时，实际的解决方案可能会受到限制，但彻底的分析可以帮助确定任何问题的来源和程度。

这是讨论指定和诊断MLR模型的基本技术的一系列示例中的第一部分。本系列还提供了一些一般策略，以解决在处理经济时间序列数据时出现的具体问题。

### 时间序列回归III:有影响的观察：

这个例子展示了如何在时间序列数据中检测有影响的观测值，并适应它们对多个线性回归模型的影响。这是关于时间序列回归的一系列示例中的第三个，在前面的示例中进行了演示。在考虑影响OLS估计的经验限制时，Belsley等人建议首先解决共线性问题。下一步是寻找有影响的观测值，这些观测值的单独或分组存在对回归结果有可测量的影响。我们将“有影响的观察”这一度量概念与“异常值”这一更主观的概念区分开来，后者可能包括任何不符合预期模式的数据。

我们首先从前面的示例时间序列回归II:共线性和估计方差加载相关数据，并继续分析那里提出的信用违约模型:

有影响力的观察以两种截然不同的方式出现。首先，它们可能是测量或记录错误的结果。在这种情况下，它们只是坏数据，不利于模型估计。另一方面，它们可能反映了创新过程的真实分布，表现出模型无法解释的异方差、偏态或瘦峰度。这样的观测可能包含异常的样本信息，但这对于准确的模型估计是必不可少的。单看数据很难确定有影响的观测的类型。最好的线索通常在产生残差序列的数据模型交互中找到。我们将在时间序列回归VI:剩余诊断示例中进一步研究这些问题。

对影响观测值的预处理有三个组成部分:识别、影响评估和调节。在计量经济学中，识别和影响评估通常基于回归统计。调整(如果有的话)通常是在删除数据(这需要对DGP进行假设)和实施适当的稳健估计程序(有可能掩盖异常但可能重要的信息)之间做出选择。

时间序列数据与横断面数据的不同之处在于，删除观测值会在样本的时间基础上留下“漏洞”。输入替换值的标准方法，如平滑，违反了CLM严格外部性的假设。如果时间序列数据表现出序列相关性，就像它们在经济环境中经常做的那样，删除观测值将改变估计的自相关性。通过残差分析诊断偏离模型规范的能力受到了损害。因此，建模过程必须在诊断和重新规范之间循环，直到可接受的系数估计产生可接受的残差系列。

### 使用条件GAN生成合成信号

这个例子展示了如何使用条件生成对抗网络生成合成泵信号。

生成对抗网络(GANs)可用于生成类似于网络实际数据输入的合成数据。当模拟计算成本很高或实验成本很高时，GANs非常有用。条件GANs (Conditional GANs)可以在训练过程中使用数据标签来生成属于特定类别的数据。

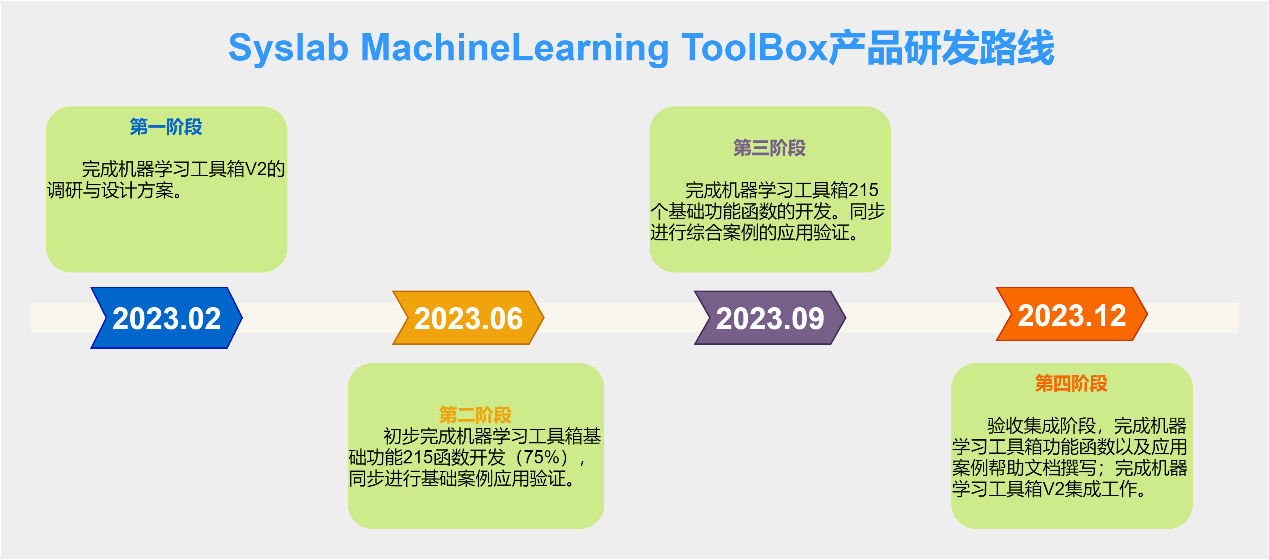
本例将泵sysplore模型获得的模拟信号作为CGAN训练数据集的“真实”数据。CGAN使用1-D卷积网络，并使用自定义训练循环和深度学习阵列进行训练。此外，本例使用主成分分析(PCA)来直观地比较生成信号和真实信号的特征。

# 研发计划

## 总体计划

Syslab 机器学习工具箱产品路线图将机器学习工具箱研发分为四个阶段，

第一阶段完成机器学习工具箱V2版本的设计方案。第二阶段在V1版本的基础上继续进行基础功能函数的开发，扩充至215个（完成75%），并同步进行机器学习工具箱机基础案例的应用。第三阶段完成机器学习工具所有基础功能函数并开始综合案例的开发，初步完成机器学习总体应用案例库的搭建。第四阶段是集成测试阶段，完成115个功能函数与应用案例帮助文档撰写，完成测试并完成工具箱集成工作。



## 函数开发计划

计划在2023.06.30完成基础功能函数215（75%）个，在2023.9.30完成215（100%）个基础功能函数。

## 应用案例开发计划

计划在2023.06.30完成至少两个基础案例和一个综合应用案例。

计划在2023.9.30完成至少四个基础案例和三个综合应用案例。

计划在2023.12.30完成至少五个基础案例和五个综合应用案例。