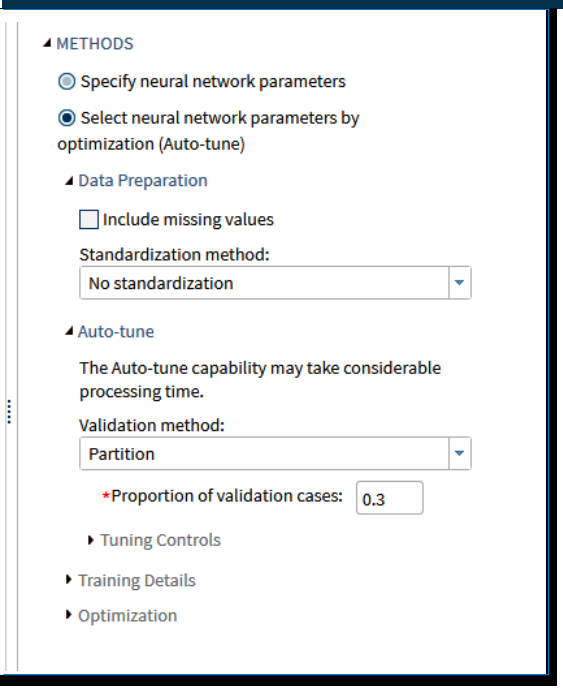
SAS®可视化数据挖掘和机器学习

单一集成内存环境满足解决最复杂问题的所有需求



SAS®可视化数据挖掘和机器学习可以做什么?

SAS 可视化数据挖掘和机器学习 (SAS Visual Data Mining and Machine Learning)将数据整理、数据探查、可视化、特征工程与现代统计、数据挖掘和机器学习技术，全部结合在单一可扩展的内存处理环境中。这样，可以更加快速准确地解决复杂业务问题，提高系统部署灵活性，建立易于管理、运行流畅的 IT 环境。

SAS®可视化数据挖掘和机器学习为什么重要?

这一解决方案帮助数据科学家和其他人员解决过去难以解决的业务问题，消除数据尺寸、数据多样性、分析深度有限和计算瓶颈带来的障碍。性能提升和创新算法得以显著提高生产效率，更加快速创造性地解决最复杂的问题。

SAS®可视化数据挖掘和机器学习为哪些用户而设计?

这款产品专为希望利用强大的自定义内存编程语言分析大量复杂数据、快速发现新的洞察结果的人员而设计，其中包括数据科学家、资深统计师、数据挖掘人员、工程师、研究员和科学家。

数据量不断增长。高水平数据科学家和分析师供不应求。企业难以及时解决日益复杂的问题。

无论分析每笔交易，识别新出现的欺诈模式，分析数量不断增加的社交媒体聊天改进客户体验，还是建立准确快速的建议系统预测近乎最好的方案，先进的机器学习软件可帮助企业解决最重要的问题。

SAS 可视化数据挖掘和机器学习支持将原始数据转换成新的分析结果的所有步骤。在单一集成式内存环境下，数据科学家可比以往更加迅速地访问、准备数据，进行探索性分析，构建并选优机器学习模型，生成预测模型的评分代码。

优点

- **提升数据科学家工作效率。**快速实现价值是企业分析项目成功的关键。SAS 可视化数据挖掘和机器学习有助于数据科学家更加迅速地获得高度准确的结果。
- **更加快速地解决复杂的分析问题。**这一解决方案采用先进的新型内存架构 SAS®Viya™，具有预测建模和机器学习功能，性能达到前所未有的水平。数据保存在内存中避免迭代分析过程中多次数据加载。分析模型处理时间按分秒，而不是小时计算，因此可比以往更加快速地找到难题的解决办法。
- **迅速推导多种方法找到最佳解决方案。**性能卓越的分布式分析引擎和功能丰富的机器学习模块便于快速轻松地推导多种方法。自动调节功能测试集成环境下不同场景，确定最佳预测模型，提供高度精确可信的答案。
- **克服大数据分析挑战。**将现代机器学习技术运用于大量结构化和非结构化文本数据，求出先前未知的分析结果。
- **利用自动生成的 SAS 评分代码快速部署预测模型。**自动生成机器学习模型多种编程语言易于实施的代码，加快实现价值的速度。
- **通过图像界面完成复杂的机器学习任务。**基于 web 的编程环境含有直观图形界面，便于配置通用机器学习任务。相关 SAS 代码自动生成，用于之后批处理和自动化。用户可在这种环境中共享数据源和代码片段，提高协作能力。

概述

目前，统计师和数据科学家往往利用多种编程语言或产品管理不同的数据挖掘任务和机器学习流程。采用 SAS 可视化数据挖掘和机器学习解决方案，他们可利用专门针对分析工作负载优化的快速内存处理引擎，在单一直观的编程环境下工作。数据科学家可迅速测试大量建模场景，高度自信地选择最佳预测模型，快速实现高级分析项目的价值。

基于 web 的灵活编程环境

从数据准备到建模、评估和评分，SAS Studio 提供基于 web 的界面，用于最常见的机器学习步骤。您可以选择采用 SAS 代码进行项目编程，也可以采用直观的图形界面支持机器学习管道最常见的任务。每项任务还可以生成各种场景下的 SAS 代码，供之后批处理、编辑和自动化。

高扩展性内存分析处理

SAS 可视化数据挖掘和机器学习利用适用于多通道分析计算的下一代 SAS 内存分析技术。分析处理引擎为并行访问内存数据

主要特点

交互型分布式基于 web 的内存编程环境

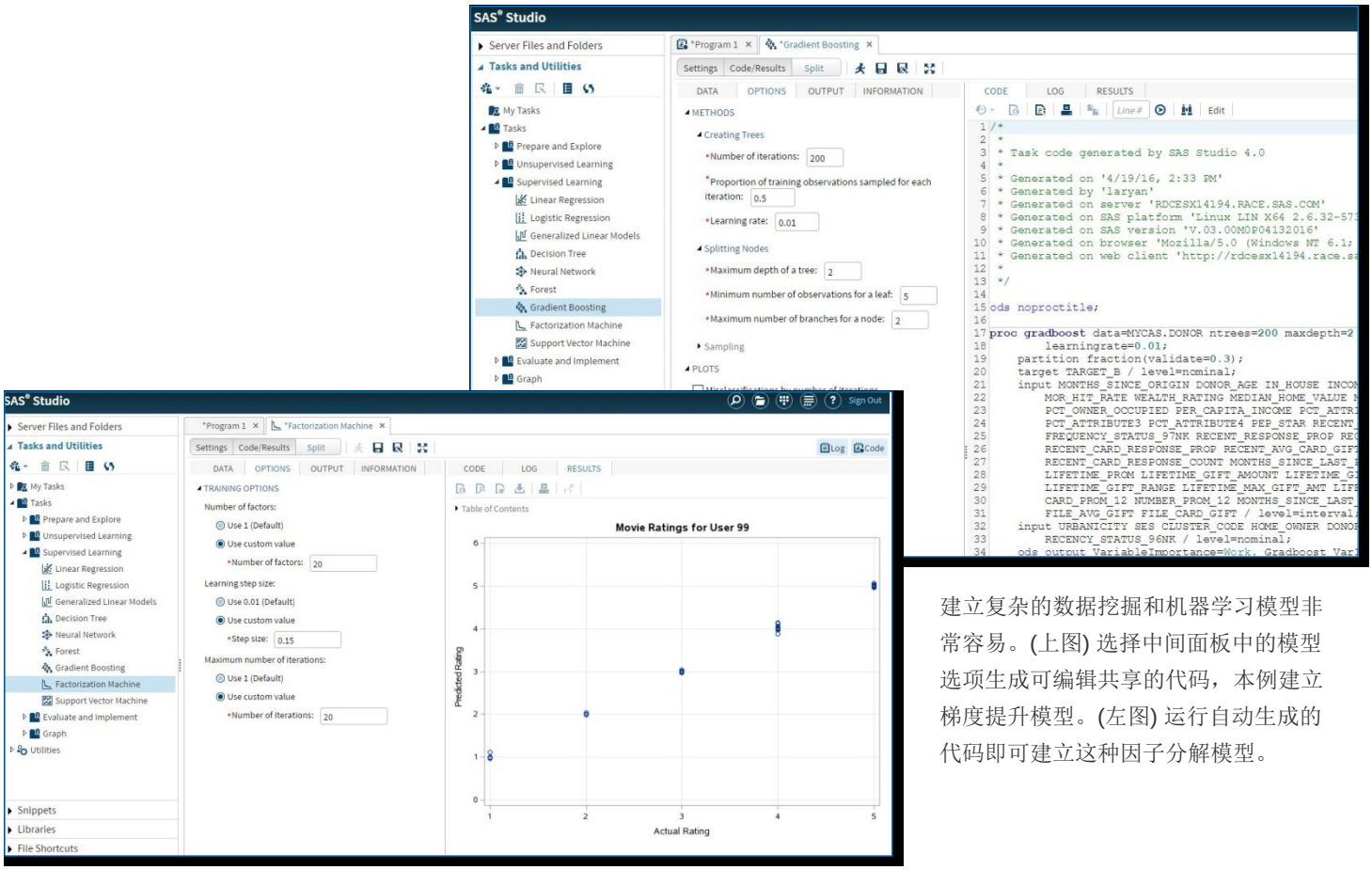
- 为程序员提供易维护的基于 web 的界面(SAS Studio)。
- SAS Studio 交互式图形任务支持点击式机器学习。
- SAS Studio 任务生成 SAS 代码快速启动并自动完成机器学习任务。
- 协作环境便于共享数据、代码和最佳实践。

高扩展性内存分析处理

- 大型数据集复杂分析计算分布式内存处理缩短提供答案的时间。
- 分析任务关联在一起作为一项内存作业，不需要重新加载数据，或将中间结果写入磁盘。
- 大量用户可并行访问内存中的相同数据源，提高工作效率。
- 凡需要的数据和中间结果可保存在内存中，缩短处理延迟时间。
- 内置工作负载管理确保有效利用计算资源。
- 内置故障切换管理功能保证完成提交的作业。

分析数据准备

- 分布式 SAS 数据步语言：
  - SAS 数据步代码在分布式计算环境下并行运行。
  - 控制每个执行节点并行级别及接合的节点数量。
- 数据汇总和基数分析：



建立复杂的数据挖掘和机器学习模型非常容易。(上图) 选择中间面板中的模型选项生成可编辑共享的代码，本例建立梯度提升模型。(左图) 运行自动生成的代码即可建立这种因子分解模型。

提供安全的多用户环境。大量用户可以共同利用相同的原始数据同时建模。

数据和分析工作量在单个服务器内核之间，或大规模计算集群节点之间自动分配，利用并行架构大大加快处理速度。凡需要的数据、表格和对象可保存在内存中，从而提高内存处理效率。

灵活扩展能力便于在大量数据基础上，采用更复杂的方法进得各种实验。从而提高所得结果可信度，显著提高生产效率，并更高效地完成建模。

此外，利用内置容错内存管理，先进的工作流程可应用于数据，确保完成整个过程。

强大的数据处理和管理功能

利用同一分布式内存环境中强大的数据操作和管理功能准备分析数据。访问数据、组合表、拆分和过滤数据，生成最终机器学习项目的表单。

数据探查、特征工程和减维

利用描述性统计和强大图形编程功能探查数据。利用先进分析技术发现并解决数据问题。迅速确定潜在预测因子，减少大型数据集维度，轻松提炼出原始数据的新特征。

现代统计、数据挖掘和机器学习技术

强大的无监督和有监督学习算法应用于结构化和非结构化数据，如聚类、主成分分析、线性和非线性回归、逻辑回归、决策树，随机森林、梯度提升、神经网络及支持向量机，迅速确定最佳模型。利用矩阵

因子分解，可构建定制推荐系统。

通过自动调节复杂机器学习算法的场景，可迅速获得高度可信的结果。采用先进优化技术，一体化自动调节流程搜索可能的参数设置组合，给出最佳设置(参见第 1 页样例)。

主要特点(续)

- 并行处理支持大规模数据探查和汇总。
- 能够快速轻松地生成全面描述性数据统计。
- 变量测量和作用智能建议(分类、数字、间隔和 ID)。
- 采样：
  - 随机和统计：罕见事件过采样。
  - 建立便于处理跟踪采样记录的变量指标。

数据探查、特征工程和减维

- 大规模连续变量离散化。
- 以用户指定值、平均值、伪中值和无缺失值的随机值来填补缺失值。
- 大规模减少连续和分类变量的维度：
  - 减少结构化输入维度，选择最优特征子集，最大化提升监督模型的预测能力。
  - 实现无监督变量选择，识别出能最大化解释数据方差的一组变量 (协方差分析)。
- 大型主成分分析(PCA):
  - 提供特征值分解、NIPALS 和 ITERGS 算法。
  - 输出主成分得分。
  - 生成碎石图和均值轮廓图。
- 无监督聚类分析：
  - 对连续和名义变量的 K 均值聚类。
  - 多种的描述相似性的距离测量方法。
  - 自动估计最佳聚类数量。
  - 输出整个观测的聚类归属和距离测量值。

现代统计、数据挖掘和机器学习算法模型部署

- 线性回归和逻辑回归模型：
  - 支持任意程度交互作用嵌套效果、多项式和曲线效果。
  - 基于向前、向后、逐步、最小角回归和套索选择方法自动模型选择。
  - 提供丰富的模型诊断结果和自动模型评估。
- 广义线性模型：
  - 支持各种分布的响应变量，包括二项、正态、泊松和伽玛分布等。
  - 支持任意程度交互作用以及嵌套、多项式和曲线效果。
  - 基于向前、向后、逐步、最小角回归和套索选择方法自动模型选择。
  - 提供丰富的模型诊断结果和自动模型评估。
- 非线性回归：
  - 非线性回归模型适合采用最小二乘法或最大似然估计算法。
  - 支持标准分布的响应变量，如二项、泊松和正态分布等。
  - 对自定义分布特征的相应变了，支持编程语法。
  - 支持特定模型和参数表达式编程语法。
  - 提供各种参数估计优化方法。
- 决策树：
  - 支持包含分类和连续特征的分类树和回归树。
  - 提供成本复杂性、C4.5 和减少误差的修剪树方法。
  - 自动修剪并基于保留最优树。
  - 自动处理缺失值，包括代理规则。
  - 自动模型拟合度评估，包括基于模型的(重新代入)统计。
- 支持二分变量、名义变量和连续变量的随机森林：
  - 自动组合多个决策树预测单个目标。
  - 自动分配独立模型训练任务。
  - 自动智能调整参数设置确定最佳模型。

集成式文本分析

这一解决方案专门为大数据而设计，您可以检查 13 种语言的大型文本文档库。利用强大的文档预处理、自然语言处理、主题检测等功能，通过探查文本数据，可获得未知情况和系统新的洞察结果。集成式文本分析功能还便于数据科学家，利用隐藏在非结构化数据中的分析结果改进监督学习。

模型评估和评分

一次性测试不同建模方法，利用标准化测试对比多种监督学习算法的结果，快速确定最佳模型。分类模型可采用提升表、ROC 图表、协调统计和错误分类表进行评估。确定最佳模型后，在分布式传统环境中进行分析，自动生成 SAS 评分代码。

主要特点(续)

- 支持二分变量、名义变量和连续变量的梯度提升：
  - 自动迭代搜索选定标签变量相关的最优数据分割。
  - 自动生成最终监督模型的加权平均值。
  - 基于验证的数据评分自动终止标准，避免过度拟合。
- 支持二分变量、名义变量和连续变量的神经网络:
  - 提供智能默认的大部分神经网络参数，如激活和误差函数。
  - 定制神经网络结构和加权。
  - 可利用任意数量隐藏层支持深度学习。
  - 自动袋外数据验证提前终止，避免过度拟合。
  - 自动智能调整参数设置确定最佳模型。
- 支持二分变量的支持向量机模型：
  - 线性和多项式内核模型训练。
  - 可采用内点法和有效集法。
  - 支持模型验证数据分割。
  - 支持惩罚选择交叉验证。
- 因子分解机：
  - 基于用户 ID 和项目评分稀疏矩阵开发推荐系统。
  - 采用全两两相互作用张量分解。
  - 含时间戳、人口数据和上下文信息的增压模型。
  - 支持热启动，因此可用新交易更新模型，不必完全重新训练。

集成式文本分析

- 支持 13 种开箱即用本地语言(英语、德语、法语、意大利语、西班牙语、葡萄牙语、荷兰语、俄语、芬兰语、土耳其语、日语、汉语和韩语）。
- 自动识别语义分段 (系统定义 15 种以上)。
- 提取标准实体，如预定义选项的地点、时间、日期和地址。
- 检测名词组合和多项列表，创建处理的单一词语。
- 同义词检测自动查找词语变种。
- 使用默认起止列表管理解析和下游处理的特定词语。
- 机器学习主题以文档收集的结构化数字表达式提供词语文档矩阵生成的文本处理。

模型评估和评分

- 自动计算选定的有监督学习模型的评估结果。
- 创建区间和分类目标的提升度表。
- 创建分类目标的 ROC 表。
- 自动生成模型评分 SAS 数据步代码。
- 提供评分功能对训练、验证和新数据进行模型评分。

联系当地 SAS 机构，请访问：sas.com/offices