UCLOUD

悟有所值 Ucan下午茶 Ucan Afternoon Tea

AutoML人工智能自动化模型设计与进化算法实现

钱广锐 博士

探智立方

探智立方

Intelligence Qubic

专业的AutoML公司

人工智能定制化服务伙伴

人工智能在企业应用现状

人工智能落地三要素：数据、算力和模型

实际场景中对人工智能“单一功能模型”和“复杂模型”的需求都普遍存在

人工智能在企业落地和市场发展趋势

悟有所值 Ucan Afternoon Tea

行业应用场景+AI模型建立

- 单一目的，简单应用场景
- 行业或专业数据导向
- 专业建模，高度端到端包装
- 定制化能力低

参考场景参考：

- 人脸识别 ：安防，银行，客服
- 自动驾驶 ：无人驾驶辅助
- 文本分析 ：法律服务，自动翻译
- 病理分析：单一病种的医疗
- 精准营销 ：零售行业
- 智能客服 ：智能机器人

注重端到端解决方案

赢家通吃

满足业务需求的解决方案

高数据安全、低成本

企业应用人工智能的痛点

模型难、数据烦、没人

企业生产流程+AI模型建立

- 复杂应用场景
- 企业数据导向
- 追求建模时效，与企业生产对接
- 高度定制化建模要求

参考场景参考：

- 投资决策
- 智能风控
- 智能质检
- 医疗辅助诊断
- IOT 制造流程优化
- 新零售、物流

注重与企业数据& 应用场景整合

高度定制

人工智能解决方案上在企业的成功之道：

自主，可控的定制化人工智能解决方案,满足企业未来的转型及成长目标

AutoML -- 人工智能游戏规则的改变者

悟有所值 Ucan Afternoon Tea

自动化机器学习（AutoML）

它是什么

开发机器学习模型需要一个耗时、专家驱动的工作流程，这个流程包括数据准备、特征选择、模型或技术选择、训练以及调优等。**AutoML**使用许多不同的统计和深度学习技术，旨在使这个工作流程实现自动化。

为什么很重要

AutoML是**AI**工具大众化的一部分，让商业用户能够在编程方面没有扎实背景的情况下开发机器学习模型。缩短数据科学家用来创建模型的时间。

pwc

Top 10 AI technology trends for 2018

Deep learning theory
The information bottleneck principle explains how a deep neural network learns.

Capsule networks
New type of deep neural network that learns with fewer errors and less data, by preserving key hierarchical relationships.

Deep reinforcement learning
This technique combines reinforcement learning with deep neural networks to learn by interacting with the environment.

Generative adversarial networks
A type of unsupervised deep learning system, implemented as two competing neural networks, enabling machine learning with less human intervention.

Lean and augmented data learning
Different techniques that enable a model to learn from less data or synthetic data.

Probabilistic programming
A high-level language that makes it easy for developers to define probability models.

Hybrid learning models
Approach that combines different types of deep neural networks with probabilistic approaches to model uncertainty.

Automated machine learning
Technique for automating the standard workflow of machine learning.

Digital twin
A virtual model used to facilitate detailed analysis and monitoring of physical or psychological systems.

Explainable artificial intelligence
Machine learning techniques that produce more explainable models while maintaining high performance.

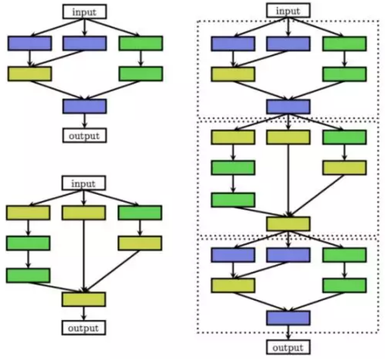
AutoML是“算法”和“工程”的集大成

“无人干预、全自动化”是AutoML的重要特征

AutoML技术发展方向和最新进展

- **AutoML解决的问题**
  - 数据特征提取和扩增
  - 机器学习和深度学习模型生成
  - 模型超参调优

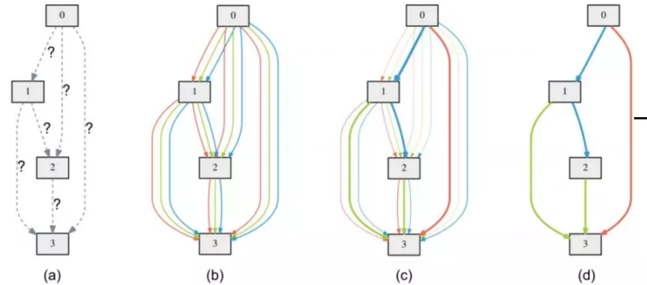
- **AutoML的主要实现方式**
  - 序贯模型优化(Sequential model-based optimization)
  - 迁移学习(Transfer Learning)
  - 强化学习(Reinforce Learning)
  - 元学习(Meta Learning)
  - 进化算法(Evolution Algorithm)
  - DARTS（基于连续假设的梯度求导方法）



Neural Architecture Search: A Survey

Thomas Elsken[1,2], Jan Hendrik Metzen[1], and Frank Hutter[2]

[1] Bosch Center for Artificial Intelligence, Robert Bosch GmbH
[2] University of Freiburg



DARTS: Differentiable Architecture Search

Hanxiao Liu
CMU
hanxiaol@cs.cmu.edu

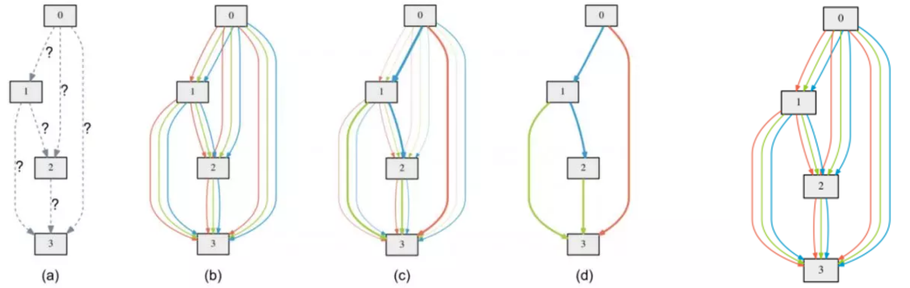
Karen Simonyan
DeepMind
simonyan@google.com

Yiming Yang
CMU
yiming@cs.cmu.edu

AutoML目前已成为“学术届”和“工业界”都成为大家追逐的热点

神经网络可微架构搜索 - DARTS

- 目前主流的神经网络架构搜索中，进化学习（evolution）和强化学习（Reinforcement）比较主流，两种方法的搜索空间都是不可微的。
- DARTS提出了一种可微的方法，可以用梯度下降来解决架构搜索的问题，所以效率可以比之前不可微的方法快几个数量级。
- 每两个节点之间都连着所有的边。点和点之间的所有的链接的权重为alpha（加权平均，和softmax类似）。alpha称作一个权值矩阵，通过梯度下降优化alpha矩阵。



$$x^{(i)} = \sum_{j<i} o^{(i,j)}(x^{(j)})$$

$$\bar{o}^{(i,j)}(x) = \sum_{o \in \mathcal{O}} \frac{\exp(\alpha_o^{(i,j)})}{\sum_{o' \in \mathcal{O}} \exp(\alpha_{o'}^{(i,j)})} o(x)$$

DARTS: Differentiable Architecture Search

Hanxiao Liu
CMU
hanxiaol@cs.cmu.edu

Karen Simonyan
DeepMind
simonyan@google.com

Yiming Yang
CMU
yiming@cs.cmu.edu

神经网络结构搜索方法的“质变”

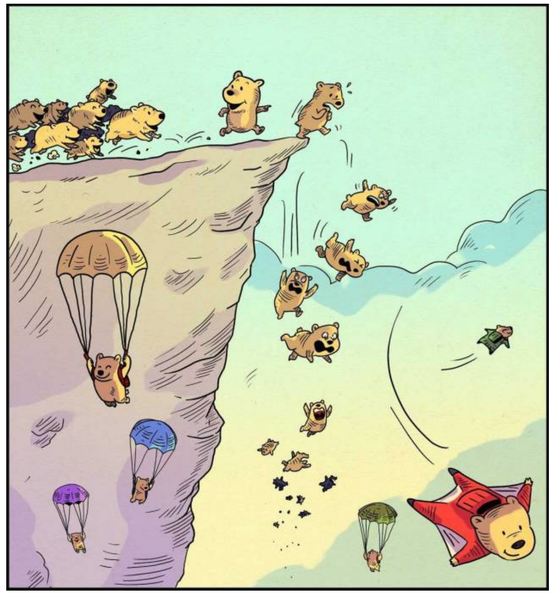
人工智能模型自动设计平台

- 以AutoML为理念，为客户打造的**企业级**人工智能平台
- 从**“零”**自动设计符合场景的模型
- 从数据准备、模型设计、到生产对接的**全生命周期管理**
- 支持不同领域**特征提取、机器学习、深度学习**业务
- 可部署于企业客户的私有云，保护数据安全

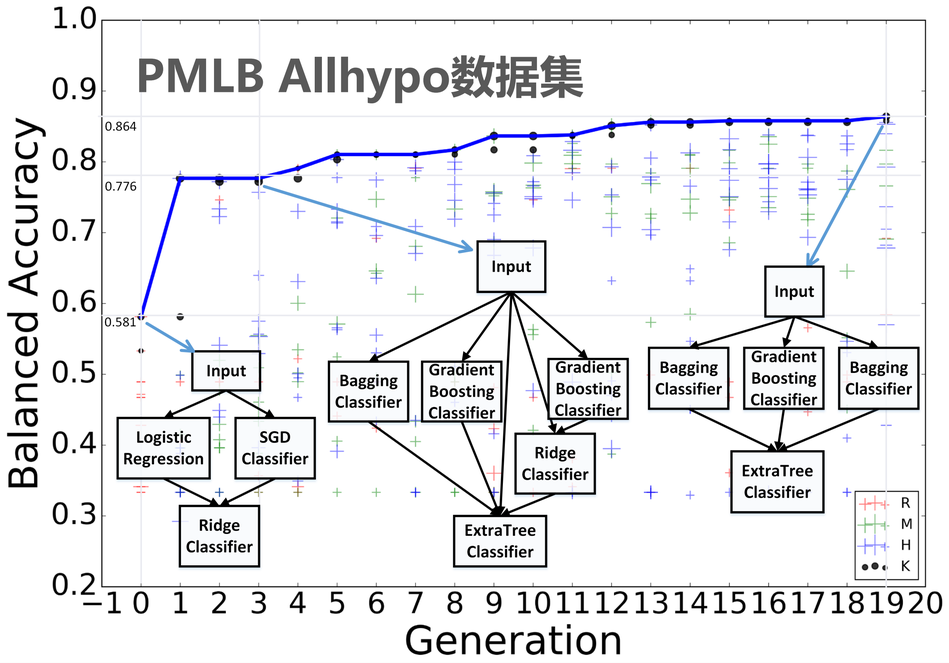
适者生存、定向淘汰

成为人工智能领域的“Windows”

DarwinML人工智能模型自动设计平台



适者生存、定向淘汰



PMLB数据集： Penn Machine Learning Benchmark data

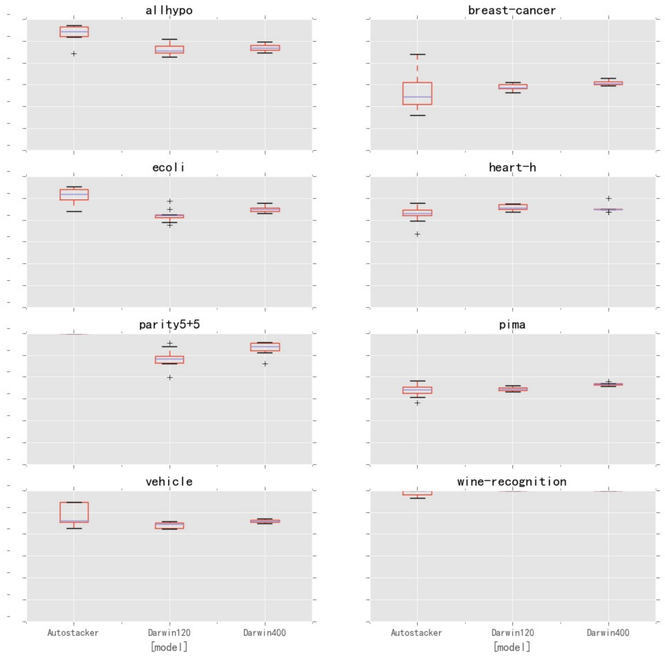
DarwinML机器学习性能结果比较

与传统的单模型、树搜索算法、Stacking方法相比：

- 算法健壮性好
- 算法准确度高

Datasets	RandomForest	Autosklearn	TPOT	Autostacker	DML120	DML400	DML400+BHO
monk1	0.98±0.009	1±0	1±0	1±0	1±0	1±0	**1±0**
parity5	0.02±0.053	0.87±0.209	0.81±0.21	0.94±0.138	1±0	1±0	**1±0**
parity5+5	0.6±0.05	1±0	1±0	1±0	0.88±0.044	0.93±0.03	**1±0**
pima	0.73±0.033	0.72±0.04	0.73±0.05	0.74±0.023	0.74±0.009	0.77±0.006	**0.79±0.009**
prnn_crabs	0.95±0.027	0.99±0.019	1±0.008	1±0	1±0	1±0	**1±0**
allhypo	0.79±0.021	0.89±0.029	0.95±0.046	0.94±0.026	0.86±0.025	0.87±0.015	**0.97±0.003**
spect	0.68±0.068	0.71±0.046	0.81±0.031	0.82±0.04	0.83±0.024	0.85±0.013	**0.86±0.01**
vehicle	0.83±0.021	**0.9±0.017**	0.82±0.039	0.89±0.044	0.84±0.012	0.86±0.007	0.85±0.005
wine-recognition	0.99±0.015	0.97±0.021	0.98±0.018	0.99±0.012	1±0	1±0	**1±0**
breast-cancer	0.59±0.058	0.59±0.059	0.67±0.09	0.66±0.08	0.69±0.015	0.71±0.012	**0.77±0.017**
cars	0.91±0.034	0.97±0.013	0.96±0.036	0.98±0.014	0.93±0.022	0.96±0.018	**0.99±0.007**
dis	0.55±0.042	0.68±0.069	0.76±0.061	0.79±0.046	0.79±0.033	0.82±0.022	**0.9±0.003**
Hill_Valley	0.56±0.027	**1±0.003**	0.96±0.043	0.98±0.015	0.98±0.013	0.97±0.01	0.97±0.013
ecoli	0.91±0.03	0.89±0.062	0.86±0.043	0.92±0.029	0.82±0.03	0.85±0.016	**0.95±0.013**
heart-h	0.79±0.036	0.79±0.042	0.81±0.047	0.83±0.022	0.86±0.013	0.85±0.018	**0.86±0.008**

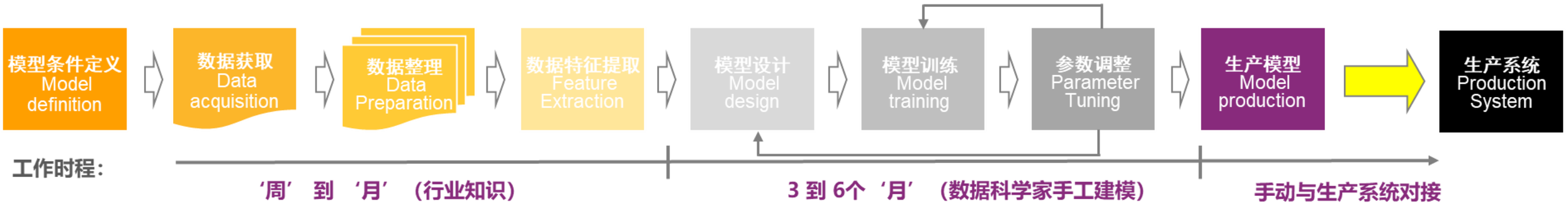
Table 1: Test Accuracy Comparison. Results on same 15 PMLB Datasets in Autostacker.



使用过程DarwinML前的模型设计

悟有所值 Ucan 下午茶 Ucan Afternoon Tea

传统机器学习模型设计流程



'烦' '难' '专业'

今天的人工智能建模门槛 '高' 、而且 '不方便'

DarwinML缩短人工智能模型设计难度和周期

- 针对客户数据，**自动生成和优化**符合业务场景的最优人工智能模型
- 涵盖**机器学习、深度学习**领域，满足企业级客户复杂模型的需求
- 平台具备**自我进化学习能力**，与时俱进加快模型设计收敛速度
- 可**多模型同时并行开发**

DarwinML全生命周期自动化流程

机器学习案例：风险控制（实时反欺诈）商户模型

悟有所值 Ucan Afternoon Tea

场景：基于商户行为的欺诈商户分类，应对实际场景中的诸多挑战。

目标：

从2万不同商户的不同时间段的统计信息，识别出8种不同的欺诈行为，并且达到每个分类识别的准确率都在95%以上

数据信息：

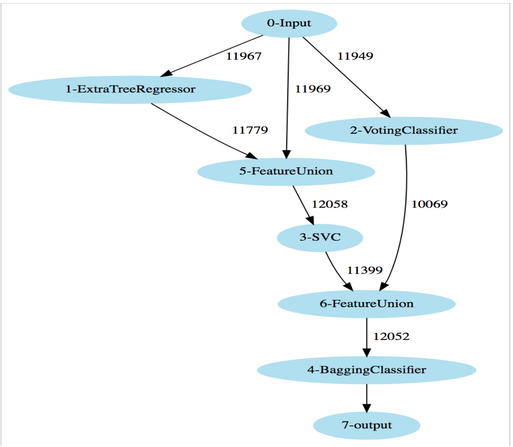
一共50509条交易记录，最少的一类欺诈商户交易统计记录仅112条

模型生成方式：

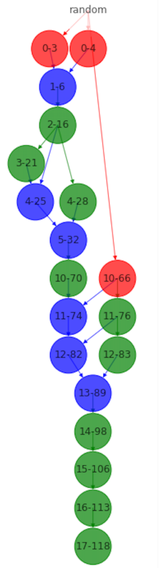
将数据总表清洗并整理后导入DarwinML平台自动生成人工智能模型

结果：

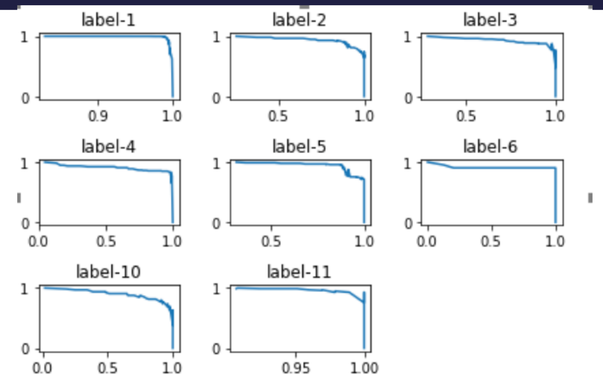
1.5天完成可用模型设计，1天模型自动超参优化



最优模型



第十七代118号模型演化轨迹



	1	2	3	4	5	6	10	11
1	2987	1	3	8	2	0	2	2
2	1	53	0	0	2	0	1	1
3	7	2	92	2	2	0	0	1
4	23	0	0	148	2	0	0	0
5	3	0	0	1	93	0	1	0
6	1	0	0	0	1	19	0	0
10	6	2	1	0	4	0	52	0
11	0	2	1	0	2	0	0	93

第十七代模型统计信息(混淆矩阵) : Accuracy 0.975

- 从第六代开始模型开始有好的收敛
- 一代运行时间为2个小时（随着机器增加，可以线性扩展）
- 进入到第十七代，在best列表中的模型都超过客户期望95%，启动Early Stop，总运行时间为1.5天

机器学习案例：风险控制（实时反欺诈）商户模型

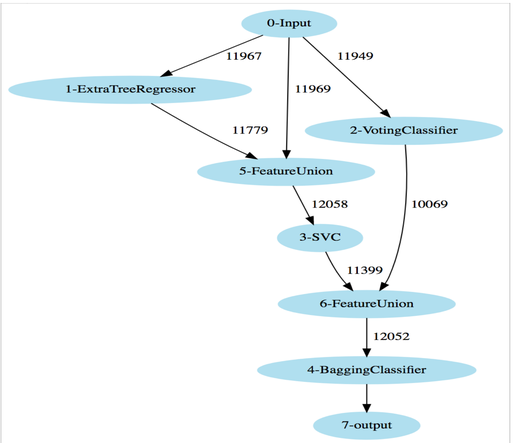
悟有所值 Ucan Afternoon Tea

时间差异

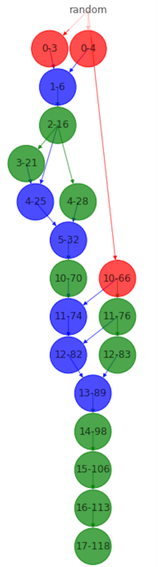
通常一个风控的模型从数据分析到模型建立到优化模型需要2个数据科学家“月”的时间来完成，而使用DarwinML只需要“3小时”数据预处理+“3天”模型训练时间

方式差异

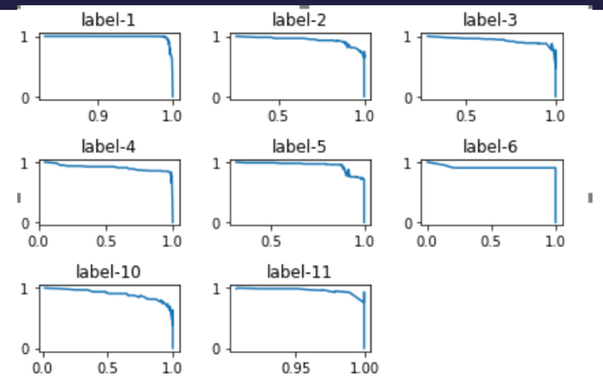
通常数据科学家需要了解业务，然后通过数据分析手段确定哪些数据特征适合用于训练，哪些需要剔除，哪些需要扩增。而DarwinML完全释放这部分工作，数据特征提和扩征取通过多层Assembling的方式来实现，降低模型设计门槛，使得普通业务人员也有机会设计模型。而且目前对于商户建模，由于商户的交易有掩盖欺诈行为的不同方式，普通规则定义很难准确找到欺诈商户



最优模型



第十七代118号模型演化轨迹



	1	2	3	4	5	6	10	11
1	2987	1	3	8	2	0	2	2
2	1	53	0	0	2	0	1	1
3	7	2	92	2	2	0	0	1
4	23	0	0	148	2	0	0	0
5	3	0	0	1	93	0	1	0
6	1	0	0	0	1	19	0	0
10	6	2	1	0	4	0	52	0
11	0	2	1	0	2	0	0	93

第十七代模型统计信息(混淆矩阵) : Accuracy 0.975

快速验证、快速落地是人工智能普及的一个重要前提

深度学习案例：工业制造缺陷检测

悟有所值 Ucan Afternoon Tea

场景： 以工业制造数据集测试DarwinML的自动模型开发深度能力

目标：

- DarwinML自动设计神经网络，并在效率和准确率上达到业界标准的性能水平
- 设计模型集群控制在24块GPU以内，设计时间在2天之内

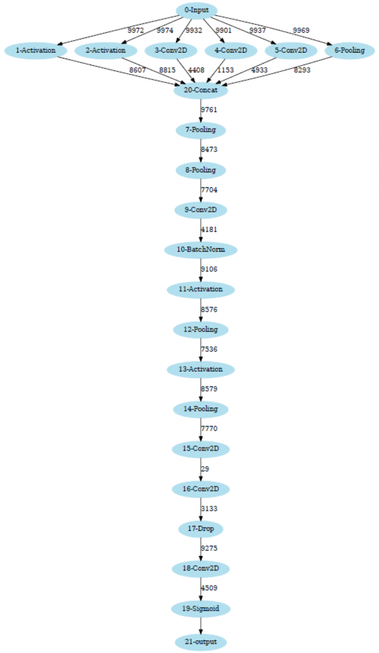
数据信息:

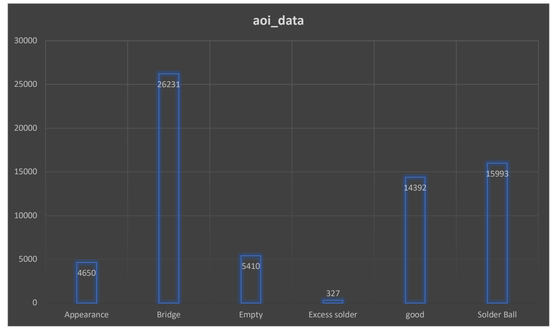
50,000张200x200 6分类图片，数据大小～800MB。

模型生成方式：从‘零’自动演化生成深度学习神经网络

结果：2天内最终准确率96%。

DarwinML在设计深度学习模型时从数据特点出发，设计出准确率和效率都十分优异的模型，无需用户了解深奥的数学建模、数据预处理等专业知识；快速建模，快速验证。Resnet50模型准确率是97% (模型文件180MB)，Darwin设计的模型更小(模型文件大小5MB）。



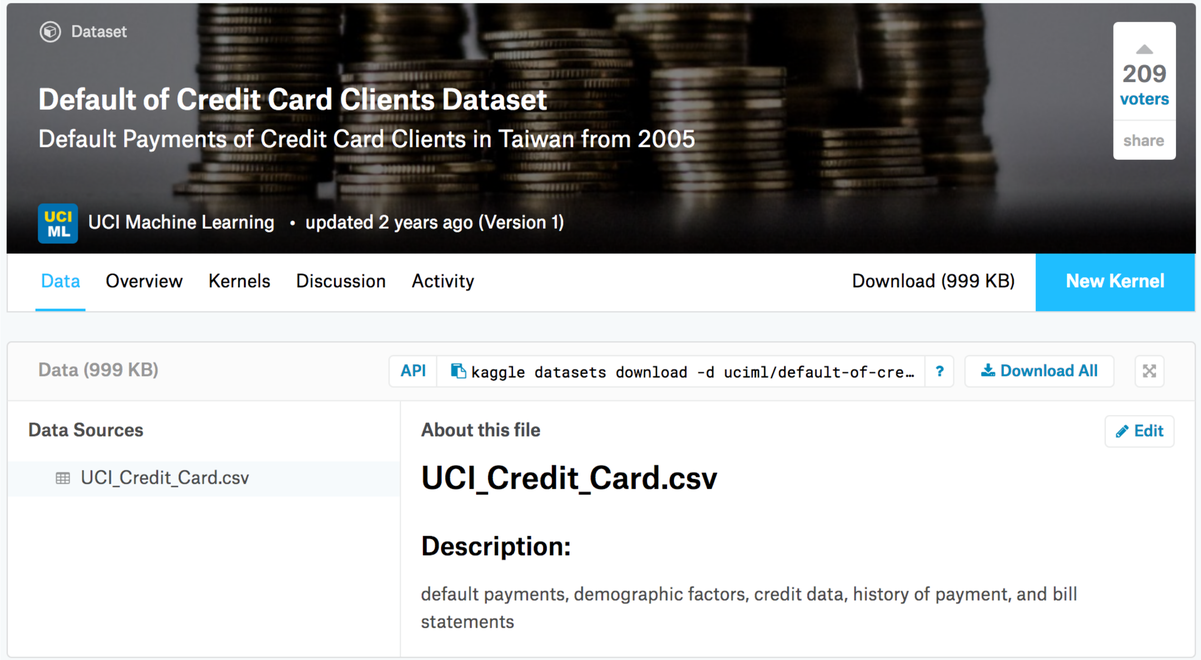


accuracy	0.947164
Precisio	0.960969
recall	0.947754

confusion_matrix		Predicted label							
		Appearan	Bridge	Empty	Excess_s	Solder_B	good	total	recall
ground true	Appearanc	416	41	6	0	0	2	465	0.894624
	Bridge	1	2615	2	1	1	3	2623	0.99695
	Empty	0	11	528	0	0	2	541	0.97597
	Excess_s	0	0	0	33	0	0	33	1
	Solder_B	7	134	5	0	1364	89	1599	0.853033
	good	3	31	2	0	13	1390	1439	0.965949
	total	427	2832	543	34	1378	1486	6700	
	Precisio	0.974239	0.923376	0.972376	0.970588	0.98984	0.935397		

DarwinML设计出的模型层数少、准确率高、模型文件小

机器学习案例：信用卡欺诈检测模型



- 数据：共30000条，按8:1:1导入，其中80%作为模型设计中的训练集、10%作为模型设计中的验证集、另外10%作为最后最初模型评估的数据；数据在平台中均分为10份，每份都是8:1:1
- 模型设计：分别按照不同的优化目标来进行模型设计：F1、Accuracy、Recall、AUC
- 模型设计参数：共10代、每代15个网络、每个网络限制在30分钟内完成、每个网络内存不超过4GB
- 模型最终输出数量：3个

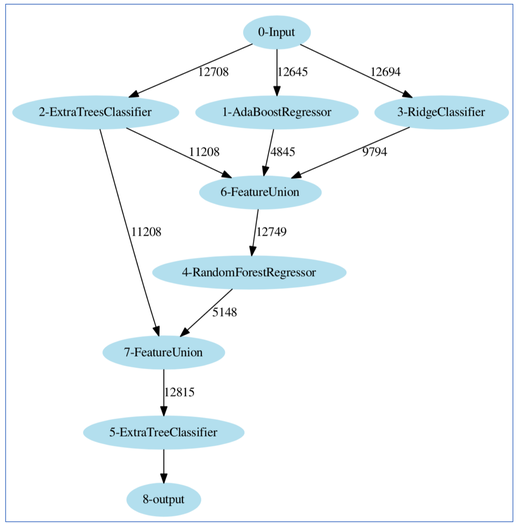
- 模型评估：根据设计出来的网络，训练出10个模型，每个模型的训练数据是9份数据，训练出来的模型对剩余的一份数据做评估，即Cross-Validation的评估方法
- 硬件环境：1台物理8core、128G的机器
- 每类目标的模型设计时间：4-5小时完成
- 每类目标的模型超参优化时间：3-4小时

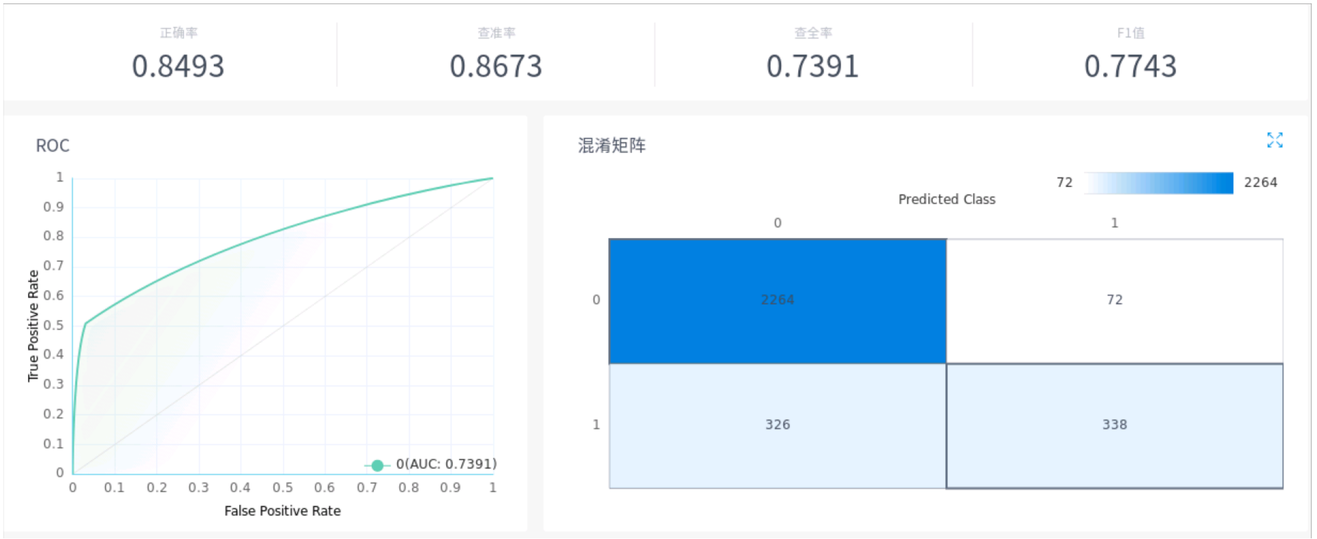
https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

机器学习案例：“F1”最优检测模型设计

悟有所值 Ucan Afternoon Tea

以指标**F1**作为模型演化的目标，**F1**最高**0.7743**



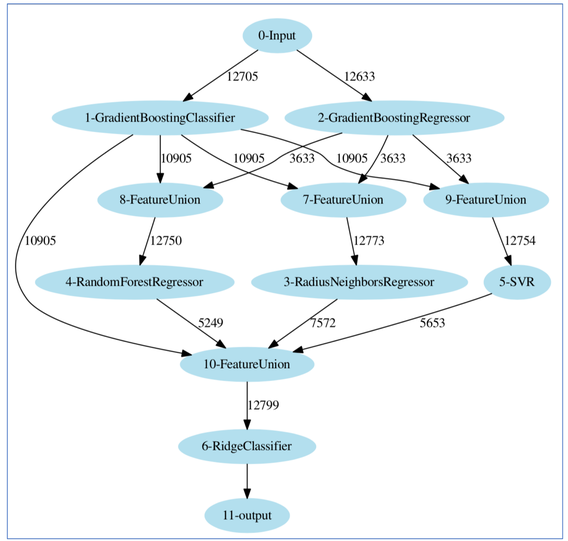


https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

机器学习案例：“正确率”最优检测模型设计

悟有所值 Ucan 下午茶 Ucan Afternoon Tea

以指标**Accuracy**作为模型演化的目标，**Accuracy**最高**0.8398**

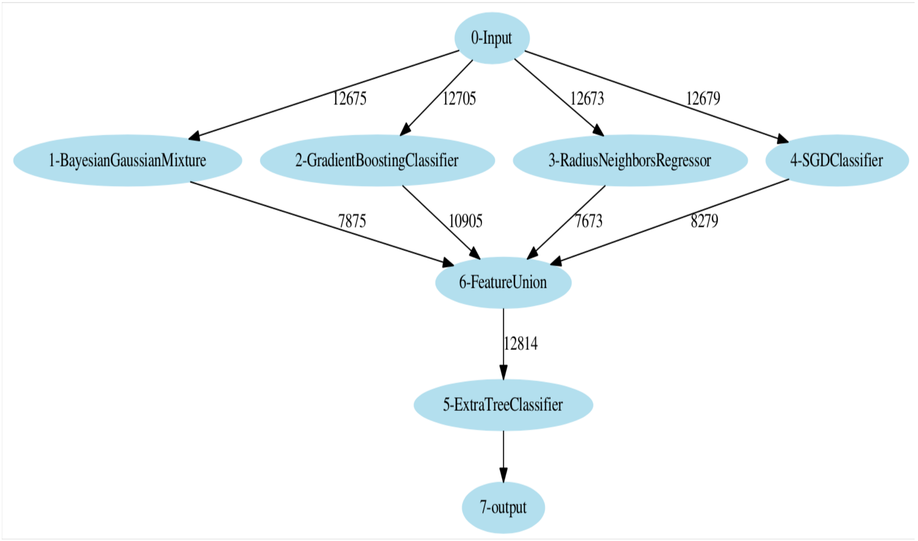


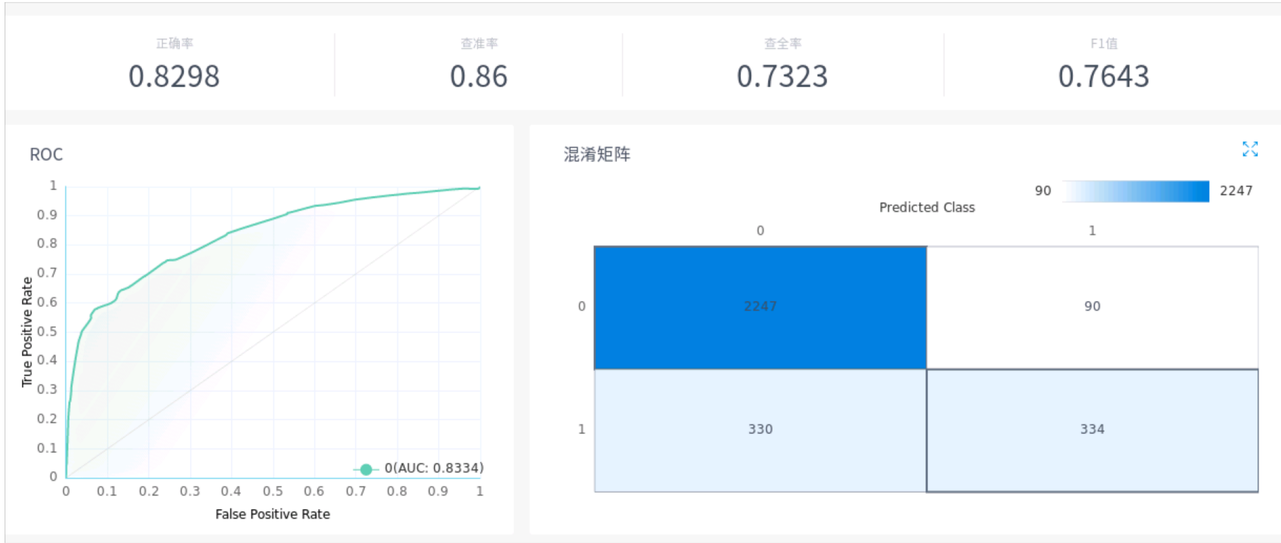
https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

机器学习案例：“查全率”最优检测模型设计

悟有所值 Ucan Afternoon Tea

以指标**Recall**作为模型演化的目标：**recall** 最高**0.7323**



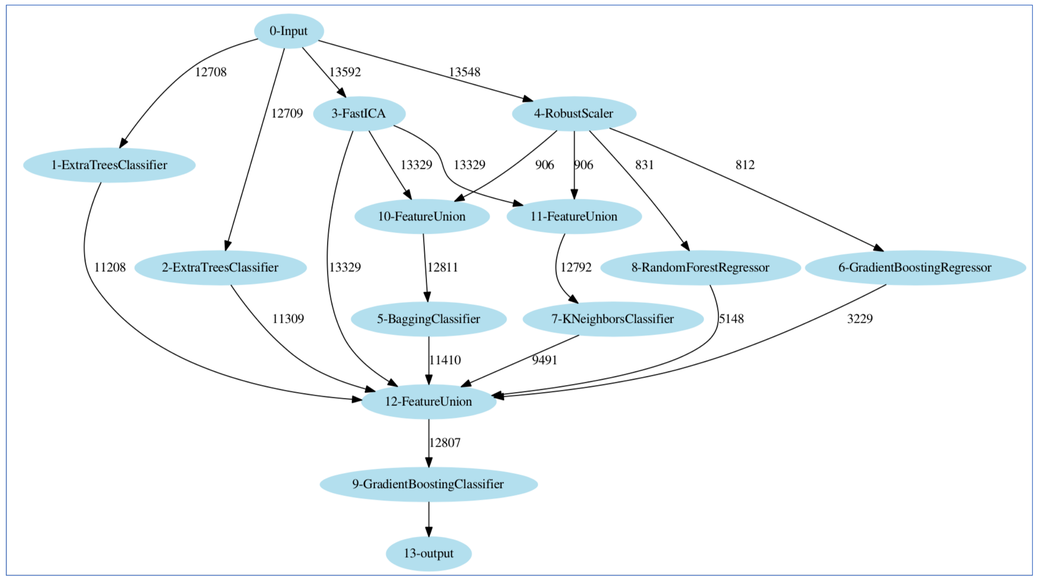


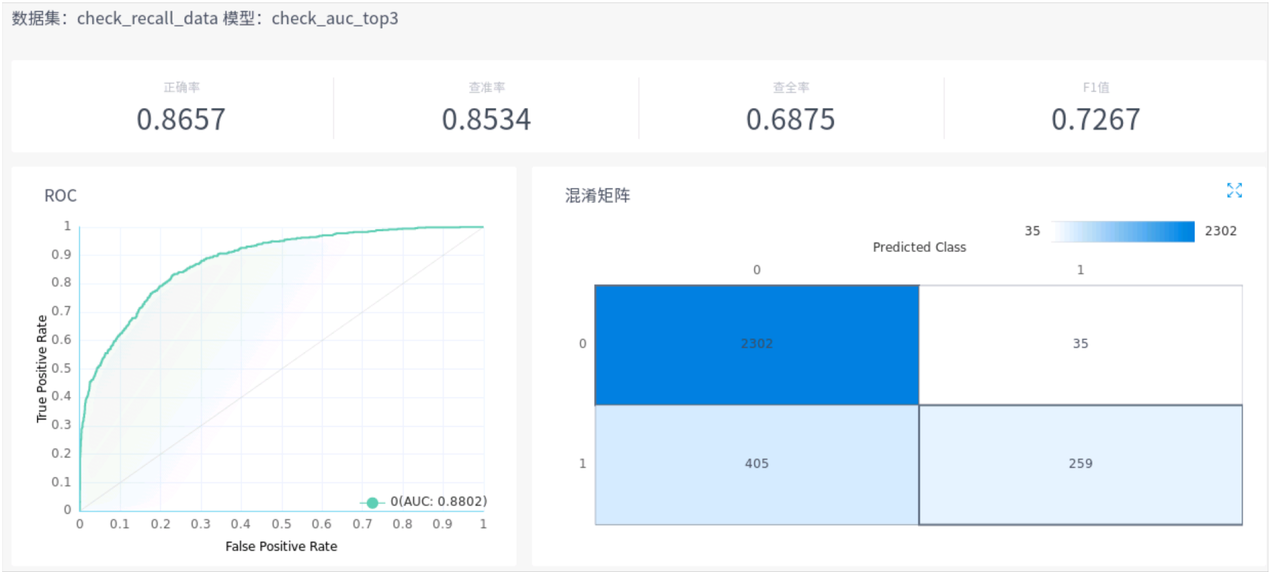
https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

机器学习案例：“AUC”最优检测模型设计

悟有所值 Ucan Afternoon Tea

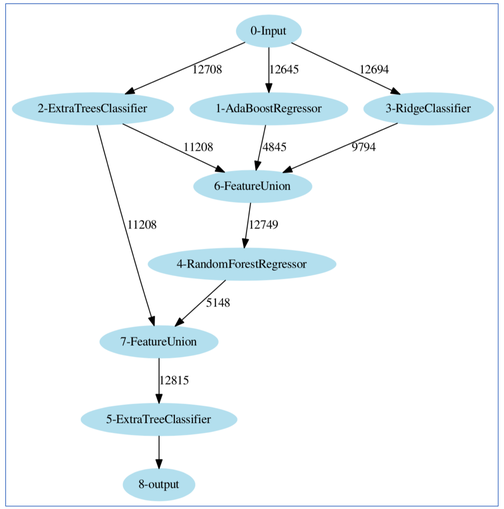
以指标AUC作为模型演化的目标, AUC最高0.8802



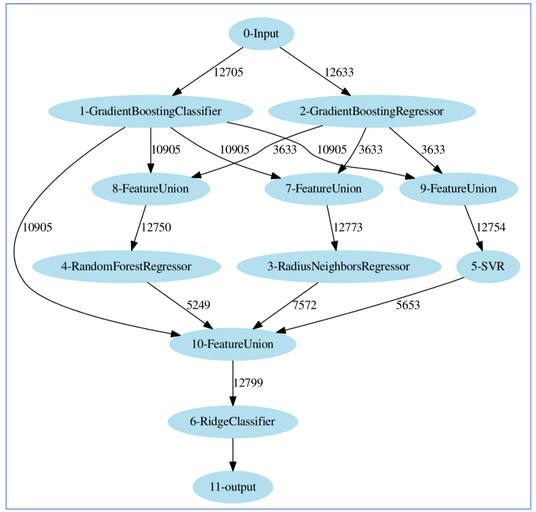


https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

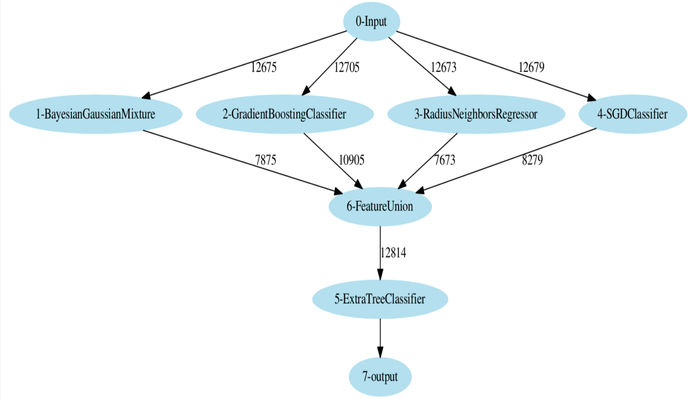
机器学习案例：信用卡欺诈检测模型



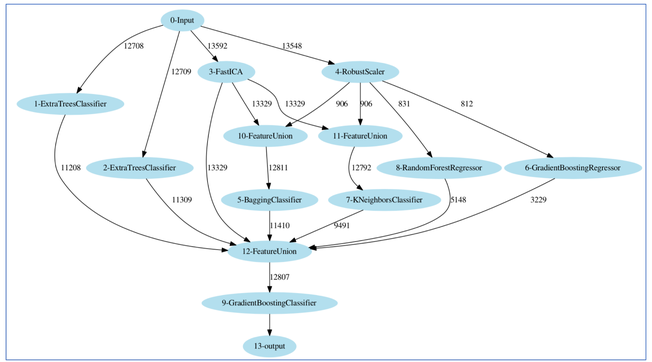
F1最高



Accuracy最高



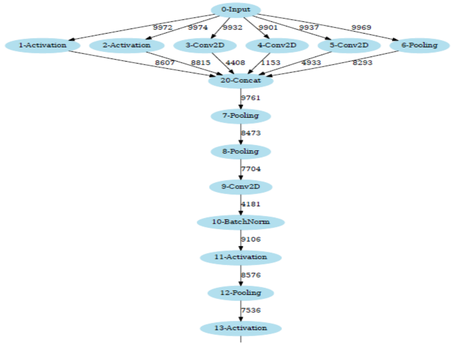
Recall最高



AUC最高

https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

DarwinIOT行业解决方案



模型设计和训练平台

- 定制化场景模型设计
- 数据标签和增广
- 模型微型化优化

Qubic4Lite

模型设计和优化

数据采集和模型反馈

TensorFlowLite　MACE

IOT Inference平台

- 用于移动设备和嵌入模型
- 数据采集

DarwinML产品路线图

图像分类

根据图像的语义信息将不同的类型的图像区分开来。

图像目标检测

检测出图像中指定的目标区域，支持多目标检测。

图像分割

将一副图像分成若干互不重叠的子区域，使得每个子区域相似，不同子区域相异。

图像检索

对给定查询图像，搜索与之在视觉或语义上相似的图像，即以图搜图。

文本分类

根据文本的语义信息将多文本划分类别，应用于情感分析等。

文本序列分析

解决序列到序列的有监督文本任务，如机器翻译、对话机器人、命名实体识别等。

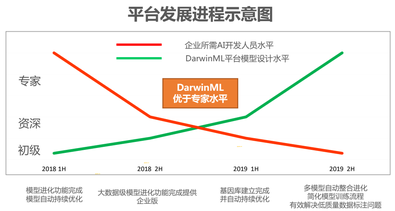
数值分类

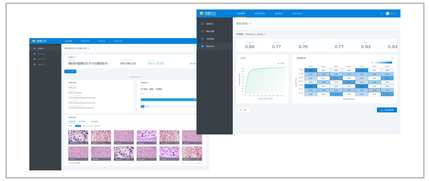
根据特征化的数值信息，对多条样本记录进行分类。

数值预测

根据有时间序列性且特征化的数值信息，预测未来时点或时段的目标数值。

做业界最好用的“人工智能”平台





v1.1
GUI支持
Sept 2018

v1.5
深度学习全面支持
Dec 2018

v2.0
数据处理支持
July 2019

半年一个小版本，一年一个大版本

DarwinML个人版本将在未来面向所有个人开发者发布

悟有所值
Ucan Afternoon Tea

THANKS