机器学习在恶意样本检测方面的实践之路

东巽科技（北京）有限公司

www.dongxuntech.com

来自恶意样本的挑战

每天不断新增的可疑样本，分析和判别是两回事！

判定规则，之外还有什么？

机器学习是人工智能的核心，也是大数据分析的基石。

我们目前的成果

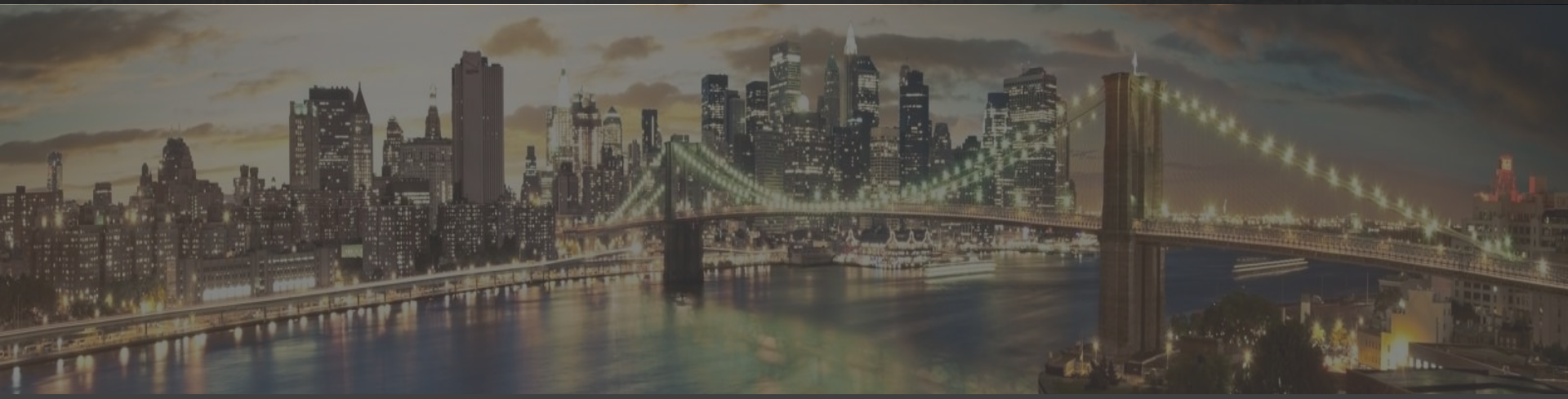
- 样本不均衡解决方法：过采样
- 样本训练集：重复正常样本数据，使得正常样本与恶意样本近似1:4(17288个样本集，包含正常样本3208个，恶意样本14080个)
- 特征：APIs
- 算法：RandomForest
- 样本预测集：约15万
- 识别率：98.84%

CONTENTS

初识机器学习

入门级简单实践

工程化的那些坑

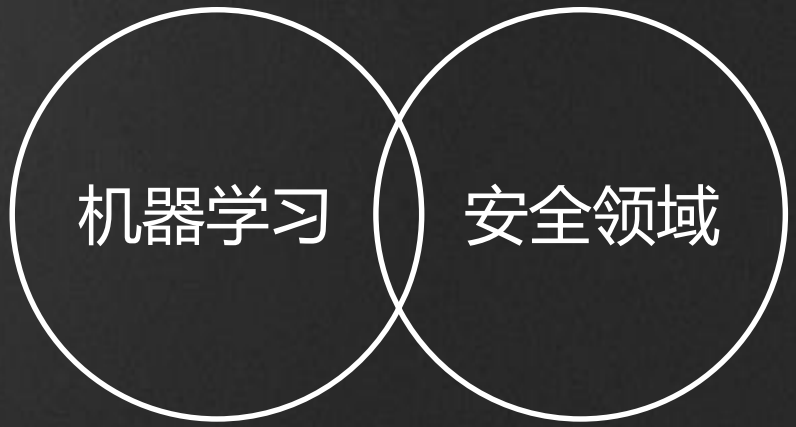


Technology Way

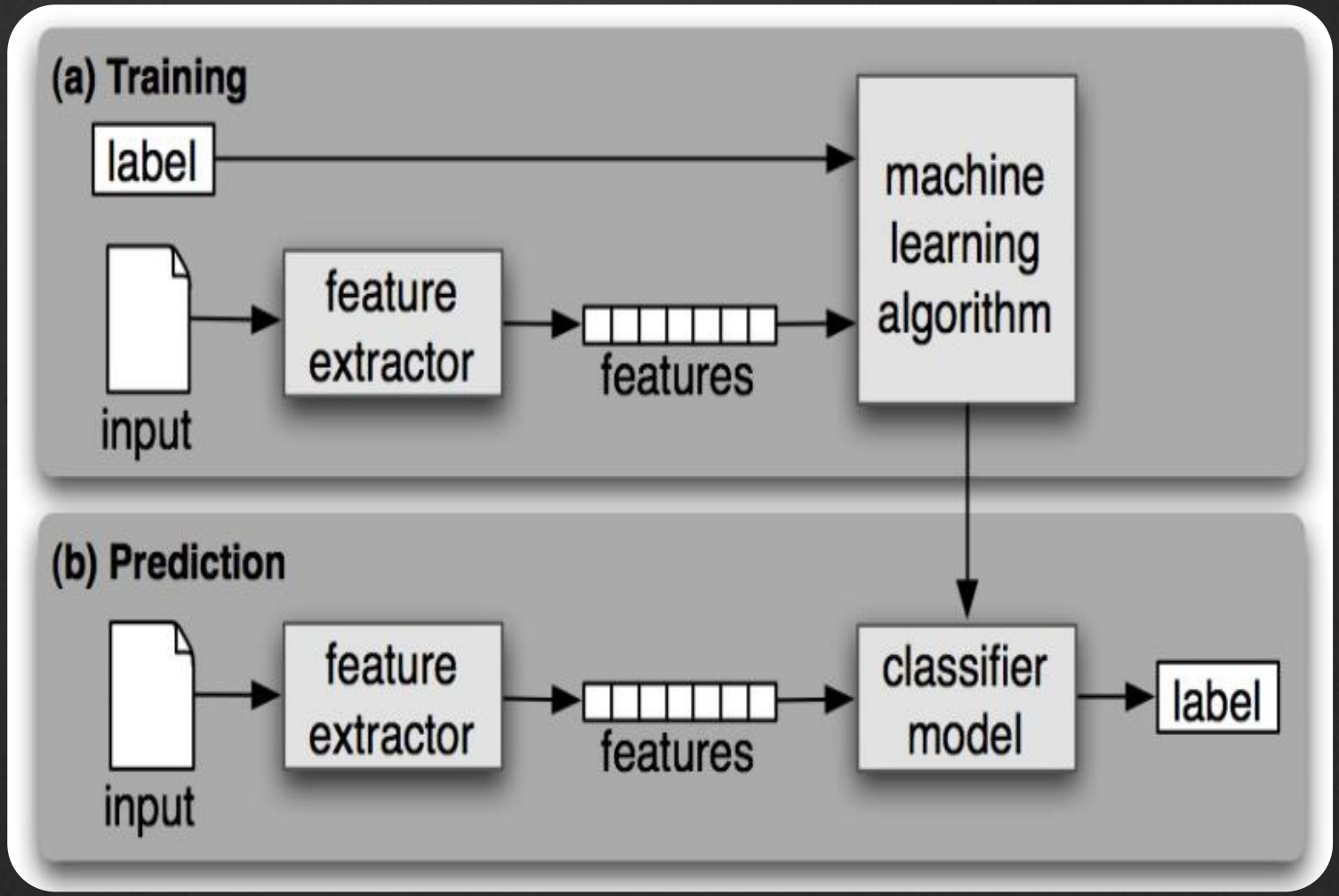
Tools

两个学科的结合

- 懂机器学习算法的，缺乏领域里的专业知识
- 有领域专业知识的，不懂机器学习算法
- 二者各自领域都存在比较高的门槛



机器学习的经典流程：训练和预测



怎么落地？

收集输入

如何选择要输入什么内容？
输入的是到底什么形式的数据？
输入的数据如何产生，从哪来呢？
输入怎么区分定义？

特征抽取

什么是特征，怎么分类？
多维度特征是什么意思？
如何选择更有效特征?
对机器学习而言，特征的选取很关键！

机器学习算法

机器学习算法五花八门，看得眼花缭乱，该如何选择算法来做训练好呢？采用聚类呢？还是分类算法呢？

评价衡量算法

要如何衡量训练的结果模型的好坏？
如何衡量是哪个因素导致的模型结果的好坏？

CONTENTS PAGE

02 入门级简单实践

输入数据

一切可以收集到、真实准确的信息都可以尝试拿来作为机器学习的输入数据。

- 样本静态分析

IDA Pro、OllyDbg、LordPE、OllyDump

- 样本动态分析

ProcessMonitor、Wireshark、CWSandbox、Cuckoo

特征抽取

东巽铁穹产品中的沙箱分析引擎每天都在产生大量的样本动态行为报告日志

铁穹沙箱分析引擎

样本动态分析报告

- 系统资源操作行为
- **系统API调用行为**
- 函数调用行为
- 字符资源调用行为
- 线程操作行为
- 数据流处理行为
- ……

➢ 样本静态报告信息：关键汇编代码段、动态库导入、可打印字符、函数长度、控制流图……

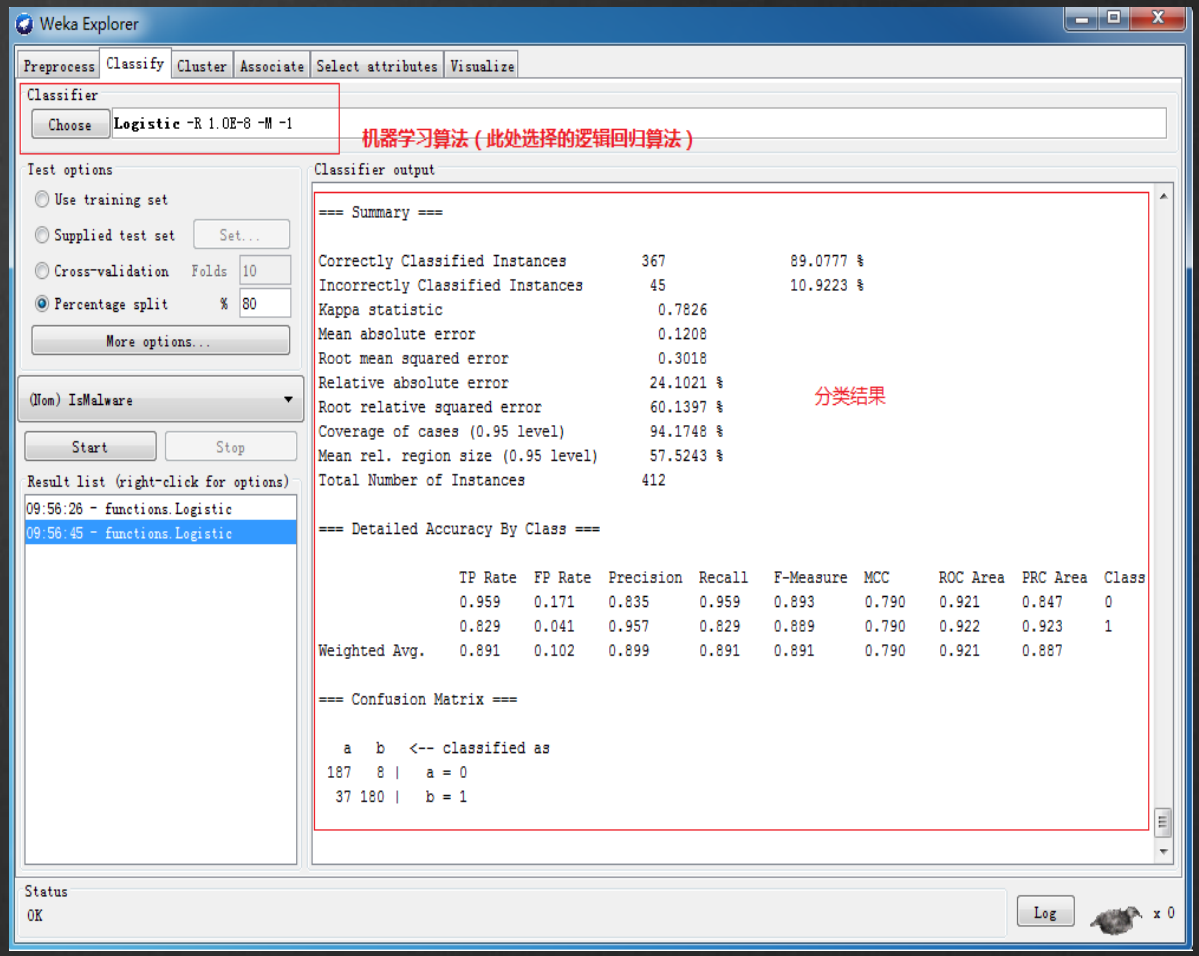
机器学习算法的选取

聚类 or 分类？

基于业务需求考虑，还是从监督模式的分类算法起步，把可疑样本分成恶意样本和非恶意样本两类

WEKA

WEKA作为一个公开的数据挖掘工作平台，集合了大量能承担数据挖掘任务的机器学习算法，包括对数据进行预处理，分类，回归、聚类、关联规则以及在新的交互式界面上的可视化。



PS:WEKA存储数据的格式是ARFF，同时WEKA也支持json、csv等格式的数据文件

评价和衡量算法优劣

- 正确率，即被分对的样本数除以所有的样本数，accuracy = （TP+TN）/(P+N)
- 错误率，也叫误差，error rate = (FP+FN)/(P+N)=1- accuracy
- 精度，即被分为正例的示例中实际为正例的比例， precision=TP/（TP+FP）
- 召回率，是覆盖面的度量，度量有多个正例被分为正例，recall=TP/(TP+FN)=TP/P

实际类别	预测类别			
		Yes	No	总计
	Yes	TP	FN	P（实际为Yes）
	No	FP	TN	N（实际为No）
	总计	P′（被分为Yes）	N′（被分为No）	P+N

简单实践的结果

- 训练样本信息： 样本数据共7099个32位PE可执行程序，其中包含恶意样本数据4000条，非恶意样本数据3099条。
- 训练样本信息路径：\analysis-datas\datamining\000001
- 算法选择： 逻辑回归
- 训练模式： 100%，即样本均作为训练样本
- 训练结果： 正确率：92.6328 % 错误率：7.3672 %

- 测试样本信息： 测试样本数据一共为1000条，均为32位恶意PE可执行程序。
- 测试结果：（正确率和错误率）正确率：88.1 % 错误率：11.9 %

CONTENTS PAGE

03 工程化的那些坑

算法模型调优

在理解算法参数的基础上，对每种算法参数的调优，并加快迭代收敛。然后再通过对算法作横向对比，对于每种算法的优劣、适用场景会有更深的认识。

测试算法	漏报	误报
逻辑回归	24%	20%
SVM	22%	43%
随机森林	12.8%	10%

该实验结果表明，随机森林算法优于逻辑回归及SVM算法

小实验：分别使用不同的算法训练生成模型，再对另一批恶意及非恶意样本进行测试。

尝试引入新的特征向量

- 不同层面
- 不同维度
- 不同颗粒度

多角度观察数据的特征

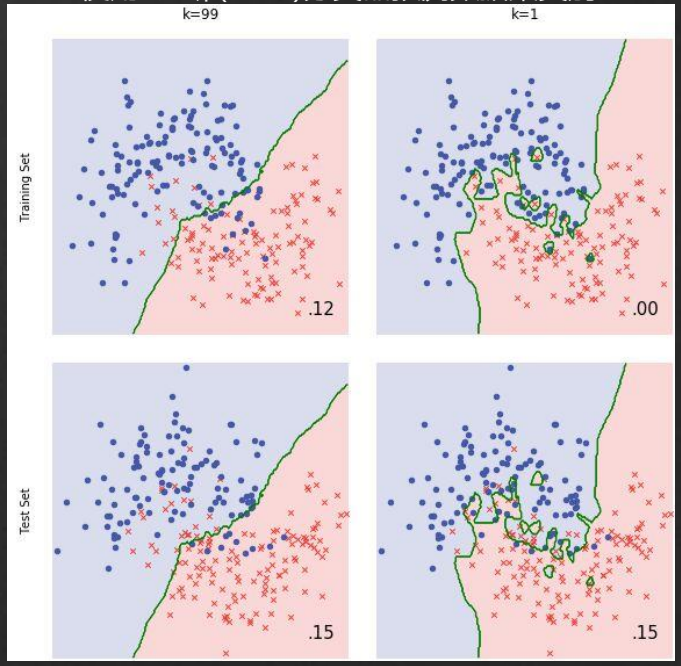
欠拟合和过拟合现象

对于一个监督学习模型来说， 过小的特征集合使得模型过于简单，过大的特征集合使得模型过于复杂，需要找到平衡之道。

- ✓ 对于特征集过小的情况，称之为 欠拟合（ underfitting ）
- ✓ 对于特征集过大的情况，称之为 过拟合（ overfitting ）

右图中，k=99的模型对捕获月牙形数据特征方面表现不是很好(这是欠拟合),而k=1的模型是对噪声严重的过拟合，过拟合的特点是良好的训练表现和糟糕的测试表现。

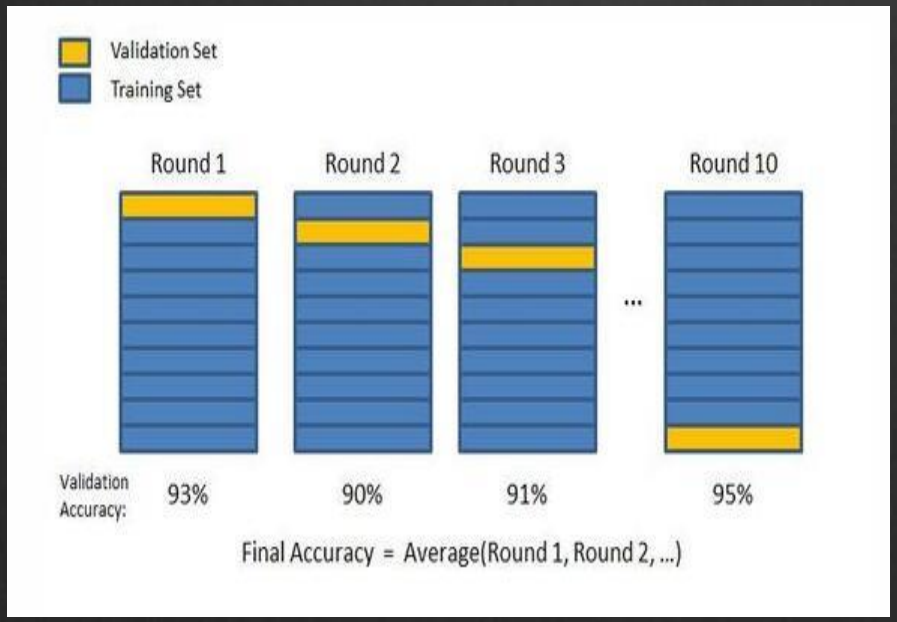
使用K近邻(KNN)分类器预测数据集类别



作者：Natasha Latysheva;Charles Ravarani

交叉验证

要解决过拟合现象就需要进一步研究算法调优，考量输入数据（样本）对算法结果的影响。 可采用K层交叉检验的方式进行试验，同时还要逐步加大测试的数据量。



交叉检验是一种打造模型的方法，通过去除数据库中K层中的一层，训练所有K减1层中的数据，然后用剩下的第K层来进行测验。然后，再将这个过程重复K次，每一次使用不同层中的数据测试，将错误结果在一个整合模型中结合和平均起来。这样做的目的是生成最精确的预测模型。

一般使用 k=5 or k=10

不平衡数据集问题

样本比例不均衡导致机器学习算法在不平衡数据集上表现不佳

过采样法

重复样本少的那类样本数据，使之达到能和样本多的那类样本数据均衡（合适的比例范围需要根据不同的算法去尝试）。

欠采样法

主要是对大类进行处理，减少大类的数据来使得数据比例平衡，同时辅以2种实现方法：

- 简易集成法

就是把大类分成多份和小类数据比例均衡的数据，然后混合大类分出的小份数据和小类数据，使用它们可以训练生成多个分类器模型。

- 平衡级联法

先生成多个分类器，再基于一定规则系统的筛选出哪些大类样本应当被保留。

这才刚刚开始

刚上路而已...路还很长...

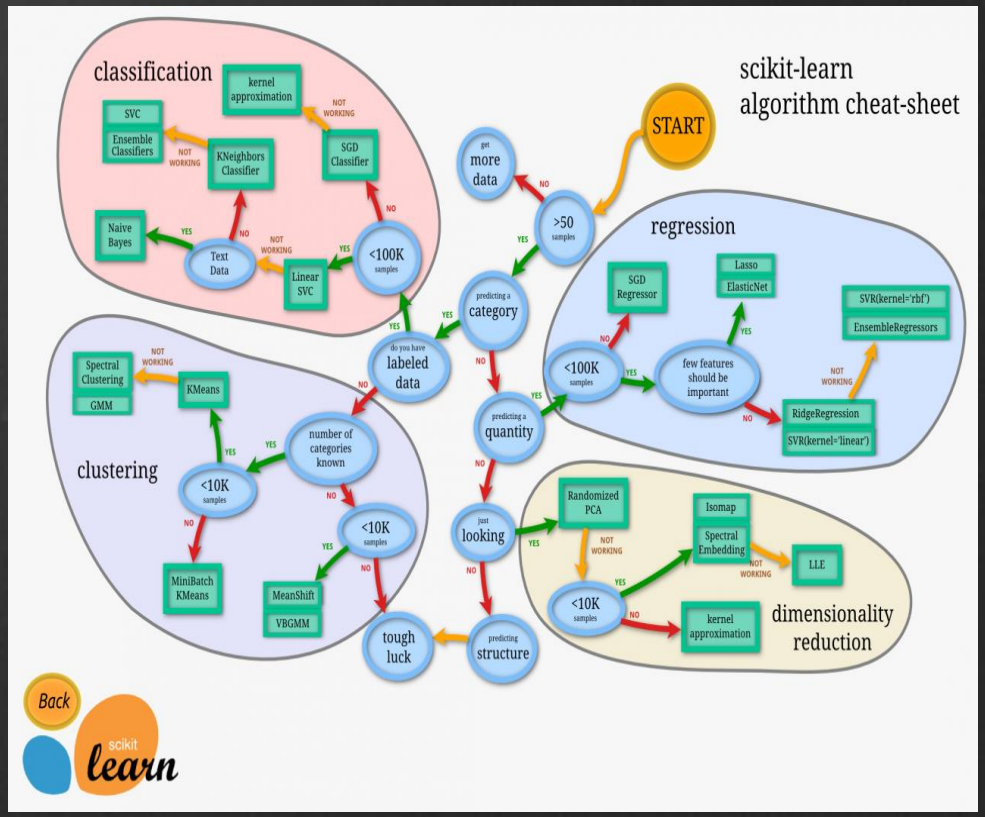
路上还有山有水有河流...

SciKit-Learn

Scikit-learn是一个非常强大的python机器学习工具包，该库基于numpy，scipy以及matplotlib库。该学习工具包中包含了丰富的机器学习的过程包，比如：预处理，降维，分类，回归，聚类等。

分类

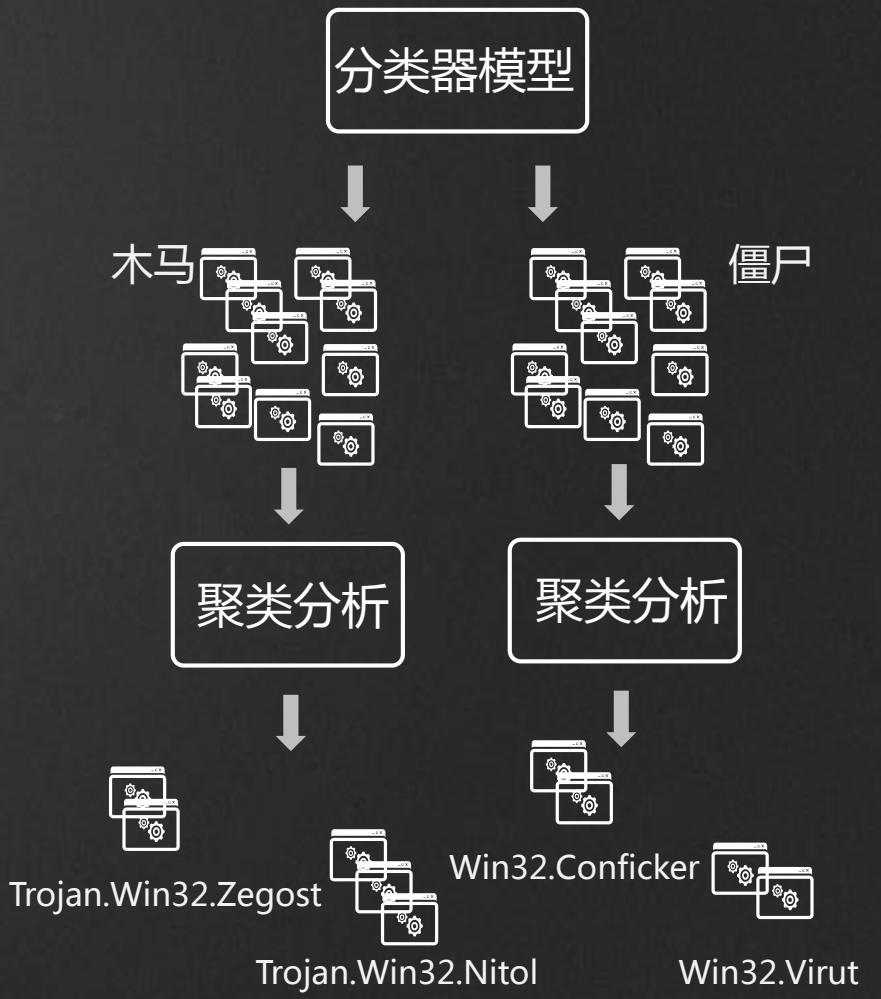
回归



聚类

降维

聚类在样本判别业务中的应用



聚类训练样本，自动标记输入数据

聚类检测样本，精细化描述检测结果

还有很多工作需要落地

优化输入的数据，尝试对输入数据进行更为细致的标签分类

引入特定行为检测规则作为抽取特征的一部分

降维去噪，做特征统计和人工分析，筛选掉一些效果不明显的特征

尝试不同算法&不同特征向量的搭配组合 $\sum_{n}^{1} x_i \times \text{分类器}_\text{i}$

增加静态分析的数据并抽取相应的高价值特征

提出一个好问题

新增样本的增量型模型问题，难道都要重新训练吗？


THANKS