DataScience Note| 数据科学学习笔记

记录自己数据科学的学习过程，主要分为数据获取、数据处理、可视化、练手项目等。

因为在工作中经常与数据打交道，同时也对火热的机器学习、大数据等新技术感兴趣，但摸索了很久，一直没什么进步。一直处于兴头来了学一点，遇到不会的就Google，遇到较难的就退缩的状态。

后来看到一句话“**老老实实把基础打牢，否则你有解不完的问题**”，深以为然。我决心系统性地梳理一下，东西比较散乱，也是自己摸索的过程，分享出来希望能逼着自己不断去整理知识体系。

目录

一、数据获取

1.1 开放数据

1.2 爬虫相关知识

分类	内容	链接	状态	更新时间
基本知识	HTTP	HTTP简介	文稿 已完成	2021-07-01
		Http headers简介	文稿 已完成	2021-07-07

		http 重定向Redirections简介		
	HTML	HTML简介	文稿 已完成	2021-07-02
		CSS选择器		
		xpath选择器		
		Beautiful Soup 解析HTML		
	正则表达式			
服务端渲染	urllib3、pycurl、requests、hyper等：	urllib3简介	文稿 已完成 代码 已验证	2021-07-02
		PycURL简介	文稿 已完成 代码 已验证	2021-07-04
		一文了解requests基本和高级用法	文稿 已完成 代码 已验证	2021-07-06
客户端渲染	寻找ajax接口	chrome开发者工具		
		Fiddler/Charles 设置代理抓包		
	模拟浏览器执行	- selenium - splinter - spynner / Ghost.py - pyppeteer - PhantomJS - Splash - requests-html		
	直接提取JavaScript	正则表达式		
	模拟执行JavaScript	- selenium：使用execute_script方法执行，return可获取执行结果 - PyExecJS - js2py - PyV8		
爬虫框架	scrapy	scrapy		
爬取APP	普通接口（接口无加密）：	Charles / fiddler / mitmproxy 直接代理抓HTTP/ HTTPS包		
	加密参数接口	Fiddler：对接C#实时处理脚本处理		

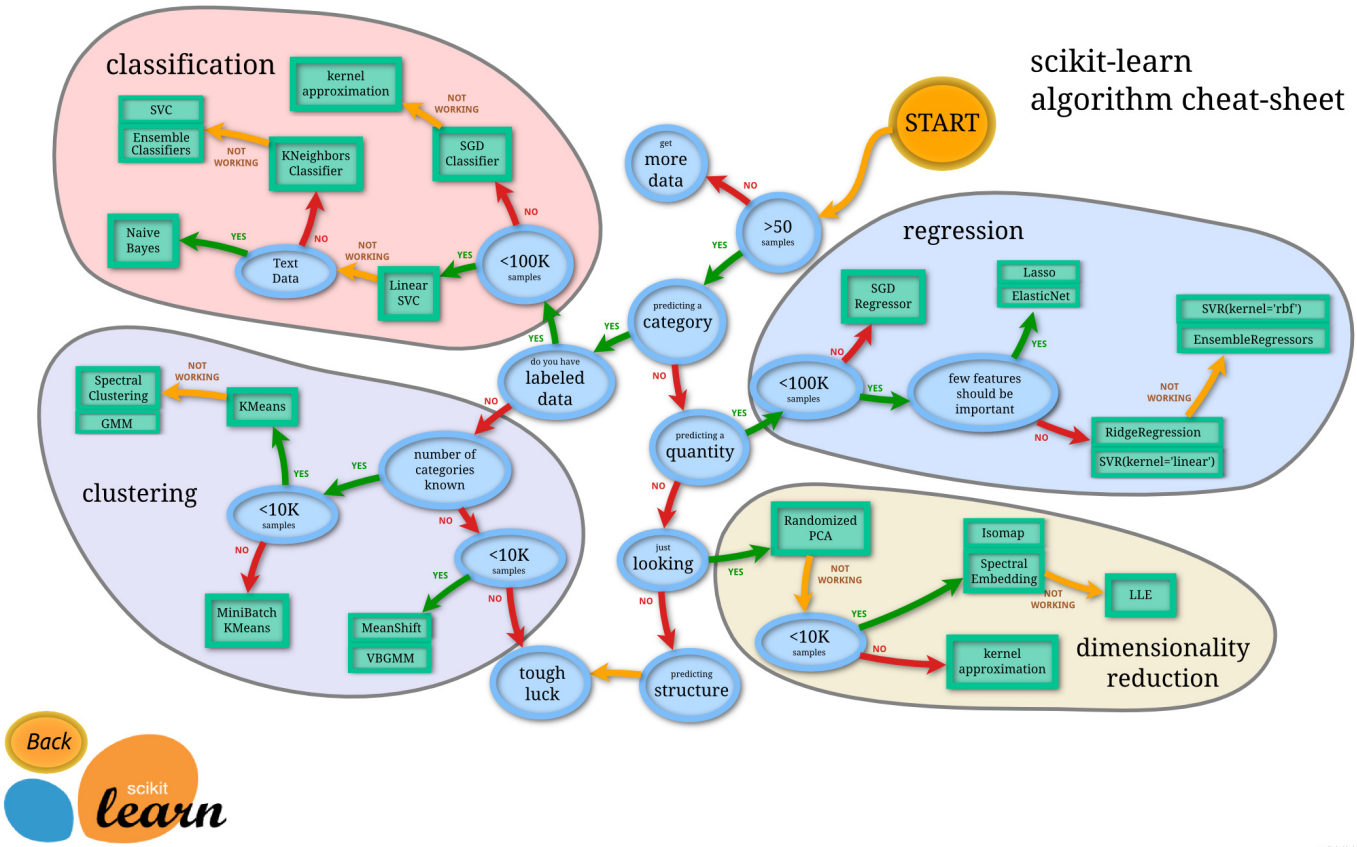
		mitmdump：对接Python脚本实时处理		
		Xposed：使用hook来直接获取结果		
		直接破解：直接破解加密参数构造规则		
	加密内容接口	Appium：类似selenium，可见即可爬		
		Xposed：使用hook来直接获取结果		
		反编译：找出加密算法，然后直接模拟		
		改写手机底层：直接修改操作系统源码		
非常规协议	wireshark：抓取所有协议的包			
	TCPdump：抓取TCP数据包			
防反爬		- 手机站点或APP站点：反爬较弱 - 免费代理：爬取免费代理使用，可用率极低 - 付费代理：可用率高，推荐：讯代理、阿布云代理、多贝云代理、芝麻代理 - 维护代理池：使用免费或付费代理自己维护代理池 - ADSL代理：使用ADSL拨号主机搭建代理池，推荐：云立方 - Tor代理：暗网代理，速度慢 - Socks代理：速度较快		

二、数据分析

2.1 数学基础

分类	内容	状态	更新时间
统计学			
线性代数	机器学习相关线性代数简介	文稿 进行中	
概率论			

2.2 scikit-learn学习笔记



分类	内容	状态	更新时间
基本知识	Scikit-Learn简介	文稿 已完成 代码 已验证	2021-07-08
分类（classification）	Logistic Regression		
	Support Vector Machine		
	Naive Bayes (Gaussian, Multinomial)		
	Stochastic Gradient Descent Classifier		
	KNN (k-nearest neighbor)		
	Decision Tree		
	Random Forest		
	Gradient Boosting Classifier		
	LGBM Classifier		
	XGBoost Classifier		
回归（regression）	Linear Regression		
	LGBM Regressor		
	XGBoost Regressor		
	CatBoost Regressor		
	Stochastic Gradient Descent Regression		
	Kernel Ridge Regression		
	Elastic Net Regression		
	Bayesian Ridge Regression		
	Gradient Boosting Regression		
	Support Vector Machine		
聚类（clustering）			
降维（dimensionality reduction）			

三、练手小项目

类别	名称	数据	源码	视频地址
分类｜奇异值检测	除了今日头条，还有哪些APP会被通报整改?	✅	✅	YouTube｜B站