常用的数据分析工具及方法

常用工具

EXCEL 电子表格软件，常用于数据整理和基本分析。

SQL 一种用于查询和操作关系型数据库的语言。

Python 一种通用编程语言，具有强大的数据分析库，如Pandas、Numpy、Matplotlib、Sklearn。

Tableau 一种数据可视化工具，可以将数据转换为交互式可共享的仪表板。

数据分析方法：机器学习模型

描述性分析：对数据进行总结和描述，以了解其基本特征和分布情况。

- 频率分布
- 集中趋势度量（均值、中位数、众数）
- 离散程度度量（极差、方差、标准差）
- 相关性分析
- 统计图表（条形图、饼图、散点图，箱线图）
- 相关矩阵

推断性分析：基于样本推断总体，以确定观察到的特征是否具有统计学意义。

- 假设检验
- t-检验
- 卡方检验
- 相关性检验
- 方差分析
- 贝叶斯推断
- 置信区间估计
- 回归分析
- 多元分析

预测性分析：使用统计模型和机器学习算法，基于历史数据，来预测未来事件。

- 时间序列分析
- 回归分析（线性回归、多项式回归、**逻辑回归**）
- 支持向量机
- **决策树**（CART）
- 集成学习（随机森林、**XGBoost**、LightGBM
- 神经网络（DNN、CNN、RNN）
- 聚类分析
- 异常监测

生成式AI：以大模型为代表的，通过模型创建新的内容和数据。

- 生成对抗网络
- 变分自编码器
- LSTM
- Transformer类模型
- ChatGPT
- Stable Diffusion
- Dall-E 2
- Sora
- 文心一言
- 通义千问
- Kimi
- ChatABC

工具一：SQL与数据库简介

SQL是一种用于查询和操作关系型数据库的语言。

关系数据库模型——典型代表：Oracle、数据库、Teradata

- **在用户观点下，关系模型中数据的逻辑结构是一张二维表，它由行和列组成。**

学生登记表

学　号	姓　名	年　龄	性　别	系　名	年　级
95004	王小明	19	女	社会学	95
95006	黄大鹏	20	男	商品学	95
95008	张文斌	18	女	法律学	95
…	…	…	…	…	…

数据库的应用

- 超市收银员扫描条码，就能调出商品价格，便于快速结账。
- 火车售票员录入出发地和目的就能调出车次、价格及车票剩余数量，利于快速售票。
- 到营业厅输入手机号和时间段就能打印出通话记录单 。
- 录入你的游戏账号和密码就能调出玩家的信息 。
- 网站发布的新闻、可转载的网络小说、网络视频、博客文章。

都在使用数据库

数据库优点

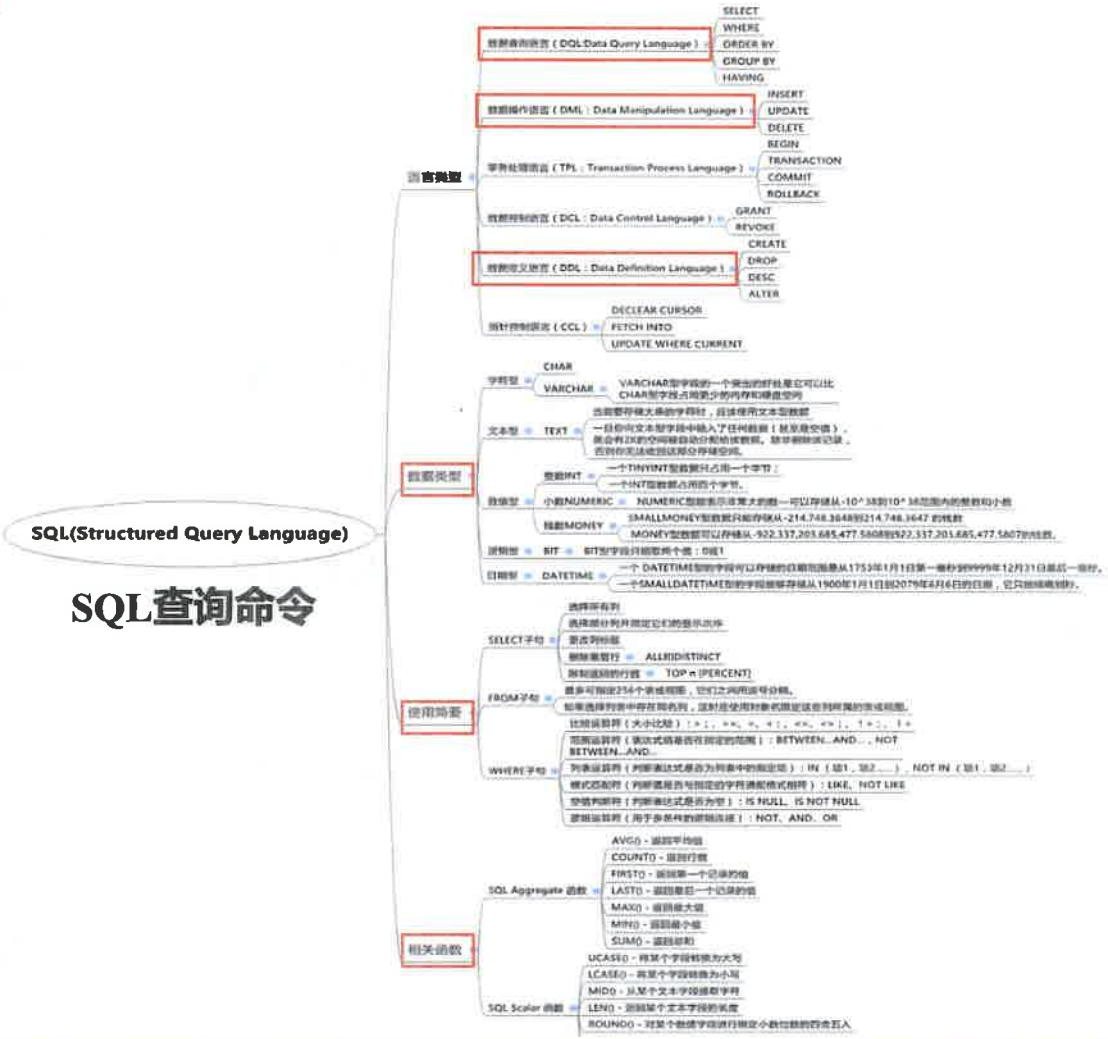
- 存储大量数据，方便检索和访问
- 保持数据信息的一致、完整
- 共享和安全
- 通过组合分析，产生新的有用信息

工具一：SQL语言发展历史

SQL语言发展历史

- 1970年，IBM的研究员E.F.Codd博士在刊物《Communication of the ACM》上发表了一篇名为“A Relational Model of Data for Large Shared Data Banks”的论文，提出了关系模型的概念，奠定了关系模型的理论基础。
- 1972年，IBM公司开始研制实验型关系数据库管理系统SYSTEM R，其配备的查询语言称为SQUARE(Specifying Queries As RelationalExpression）语言，语言中使用了较多的数学符号。
- ·1974年，Boyce和Chamberlin把SQUARE修改为SEQUEL(Structured English Query Language）语言。后来SEQUEL简称为SQL(Structured Query Language），即“结构式查询语言”，SQL的发音仍“sequel”。现在SQL已经成为一个标准。
- 数据库语言的美国标准，I SO随后也通过这一标准使得SQL成为数据库领域的主流语言。

工具一：SQL语言命令简介



工具一：SQL-Select命令简介

SQL SELECT语句的功能

- 列选择：你能够使用SELECT语句的列选择功能选择表中的列，这些列是你想要用查询返回的。当你查询时，你能够选择你查询的表中指定的列。

- 行选择：你能够使用SELECT语句的行选择功能选择表中的行，这些行是你想要用查询返回的。你能够使用不同的标准限制你看见的行。

- 连接：你能够使用SELECT语句的连接功能来集合数据，这些数据被存储在不同的表中，在它们之间可以创建连接。在后面的课程中你将学到更多关于连接的内容。

```
%%sql
select APSDACTNO as '账号',
       APSDAACNO as '对方账号',
       apsdtrdat as '交易日期',
       APSDTRTM as '交易时间',
       apsdtramt as '交易金额'
from dadmr.p_00abis_aps
where apsdtrdat = '20231231' and apsdtramt > 1000
limit 2
```

活期交易明细表

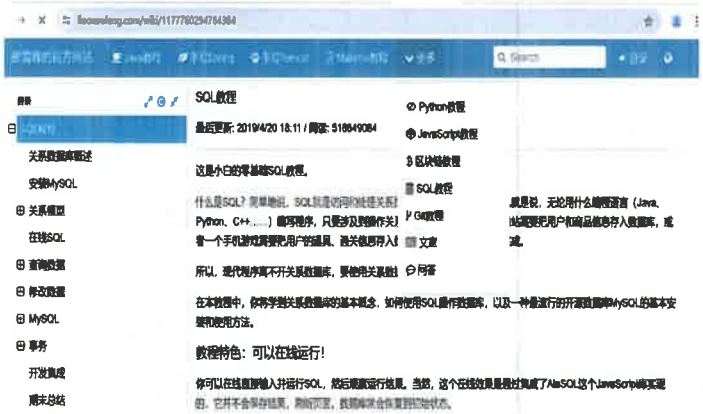
```
Last executed at 2024-05-22 16:09:58 in 4.27s
您的SQL已进入GBase集群，请等待执行结果

2 rows affected.
```

	账号		对方账号	交易日期	交易时间	交易金额
621	3781	621	0900	20231231	112848	2800.00
620	1901	620	0001	20231231	211332	9900.00

工具一：SQL学习资料

书籍： SQL必知必会

网站： 廖雪峰的官方网站-SQL教程
http://www.liaoxuefeng.com



视频： B站-MySQL基础+进阶

目录

工具二：Python简介

Python 是一种解释型、面向对象、动态数据类型的高级程序设计语言，具有强大的数据分析库。首个公开发行版诞生于1991年。
常用版本：Python 3.X，最新已发布了3.12。官网：https://www.python.org

1. 易于学习：关键字相对较少，结构简单，语法定义简单，学习相对容易。
2. 易于阅读：代码定义清晰，便于上手。
3. 易于维护：源代码容易维护。
4. 广泛的标准库：丰富的内置库，可以处理各种工作，跨平台兼容性很好。
5. 支持互动模式：可从终端输入代码并获得结果，实现代码的互动调试。
6. 可移植：由于源代码开放，Python已被移植到许多平台。
7. 可扩展：可以灵活调用C语言程序实现关键模块。
8. 提供数据库接口，可以调用数据库中的数据。

目前是最受欢迎的编程语言之一，广泛应用于科学计算、人工智能、数据分析等领域。

工具二：Python应用场景

数据获取（爬虫）

各种好用的爬虫框架

形成数据分析报告

数据分析：快速做数据的汇总统计。
绘图工具包：方便自定义绘制图表。

数据建模

数据探查及处理阶段：Pandas等类数据库的包，可以做数据表关联及增删改查，处理数据方便快捷。
建模阶段：成熟的机器学习、深度学习算法包可直接调用，提供从模型构建、调参到评估全流程的函数，能快速验证建模思路。

工具二：Python下载与安装

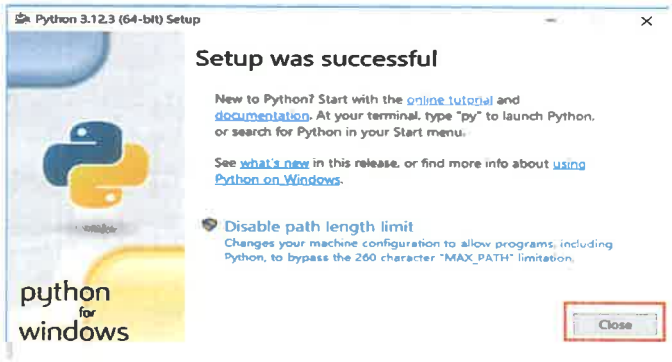
进入python官网https://www.python.org，选择合适的安装包，点击下载。

python-3.1
2.3-amd64.e

Files

Version	Operating System	Description	MD5 Sum	File Size	GPG	Signature	SBOM
Gzipped source tarball	Source release		3c5498a34d5226c9b746b1199f0bf2d9	25.9 MB	SIG	signature	SPDX
XZ compressed source tarball	Source release		8defb33f0c37ae4bdd3a38be52abde4e	19.7 MB	SIG	signature	SPDX
macOS 64-bit universal2 installer	macOS	for macOS 10.9 and later	611e3bbf6188f21eb3bfbeb885be1bf	43.8 MB	SIG	signature	
Windows installer (64-bit)	Windows	Recommended	c8694971f0e04f1e085db870290819489	25.8 MB	SIG	signature	
Windows installer (ARM64)	Windows	Experimental	ef016521b5a147f3ded730801d36e350	24.7 MB	SIG	signature	
Windows embeddable package (64-bit)	Windows		38cca2bf3b4de76db19a31f4e07206e	10.5 MB	SIG	signature	
Windows embeddable package (32-bit)	Windows		85d873e723cb66d6746e9872af5e715e	9.4 MB	SIG	signature	
Windows embeddable package (ARM64)	Windows		3229271cac55ff913aad1bb46ebc9931	9.6 MB	SIG	signature	
Windows installer (32-bit)	Windows		a95c4fbdce0b6a22ce7cfb450f37c010	24.2 MB	SIG	signature	

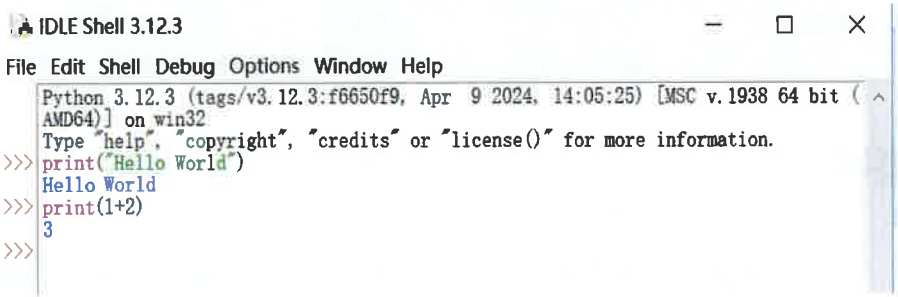
下载后，双击下载包，进入 Python 安装向导，安装非常简单，勾选Add python.exe to PATH后，只需要使用默认的设置一直点击"下一步"直到安装完成即可。



工具二：Python的使用

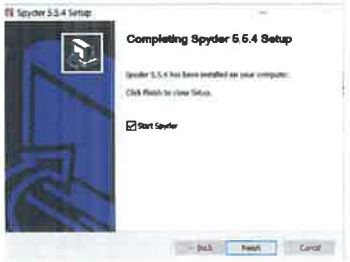
有两种方式可以运行Python代码
一是终端交互式解释器：可以通过命令行窗口进入 Python，并在交互式解释器中开始编写 Python 代码。
二是集成开发环境（IDE）：PyCharm(需要激活码)、Anaconda(要商用了)，**spyder**，**Jupyter notebook**

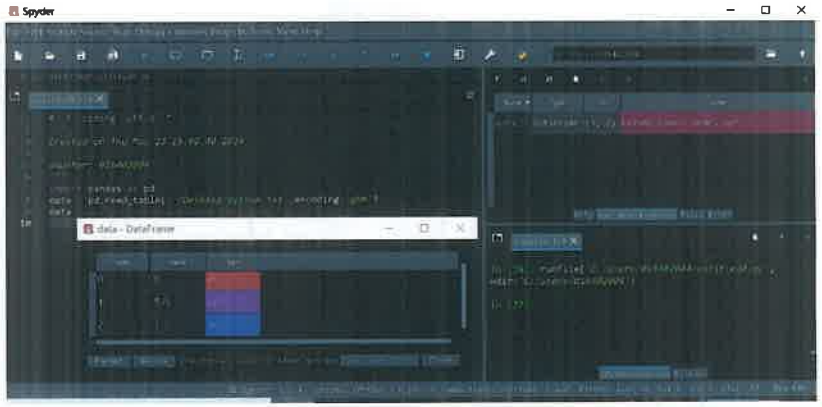
方式一：终端交互式解释器



方式二：spyder

下载后，双击进入 spyder安装向导，一直点击"下一步"直到安装完成即可。





工具二：Python语法内容简介

Python 标识符：为对于变量、常量、函数、语句块起的名字，由字母、数字、下划线“_”组成；不能以数字开头；区分大小写。

数据类型：数值、字符串、列表、字典、元组

逻辑控制语句：条件判断if、For循环、While循环

常用的数据包：

Pandas（最常用）：提供了快速、灵活以及表达能力强的DataFrame数据结构，使得数据操作简单直观，主要用于数据清洗和分析工作。

Numpy：是一个科学计算库，提供对多维数组进行高效率的矩阵运算支持。

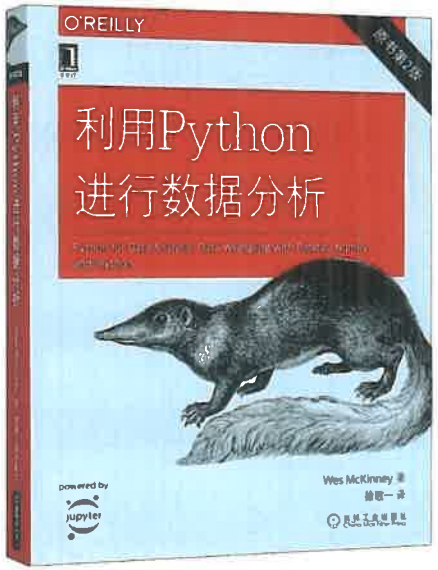
Matplotlib：绘制2D的静态、交互式以及动画图表，常用于数据可视化。

Seaborn：是一个基于matplotlib的高级接口，使得绘图更简单，绘出的图更漂亮。

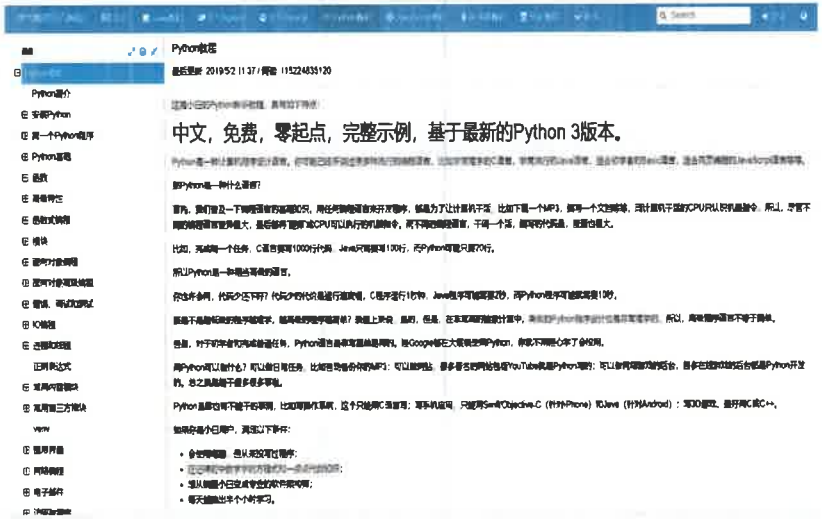
Scikit-learn：是一个广泛使用的机器学习库，提供了简单有效的工具，使用各类机器学习算法，进行数据挖掘和数据分析。

工具二：Python学习资料

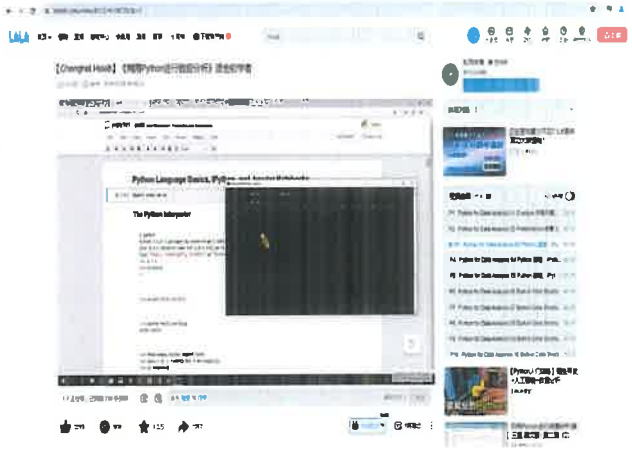
书籍：利用Python进行数据分析



网站：廖雪峰的官方网站-Python教程
http://www.liaoxuefeng.com



视频：利用python进行数据分析



工具三：机器学习是什么？

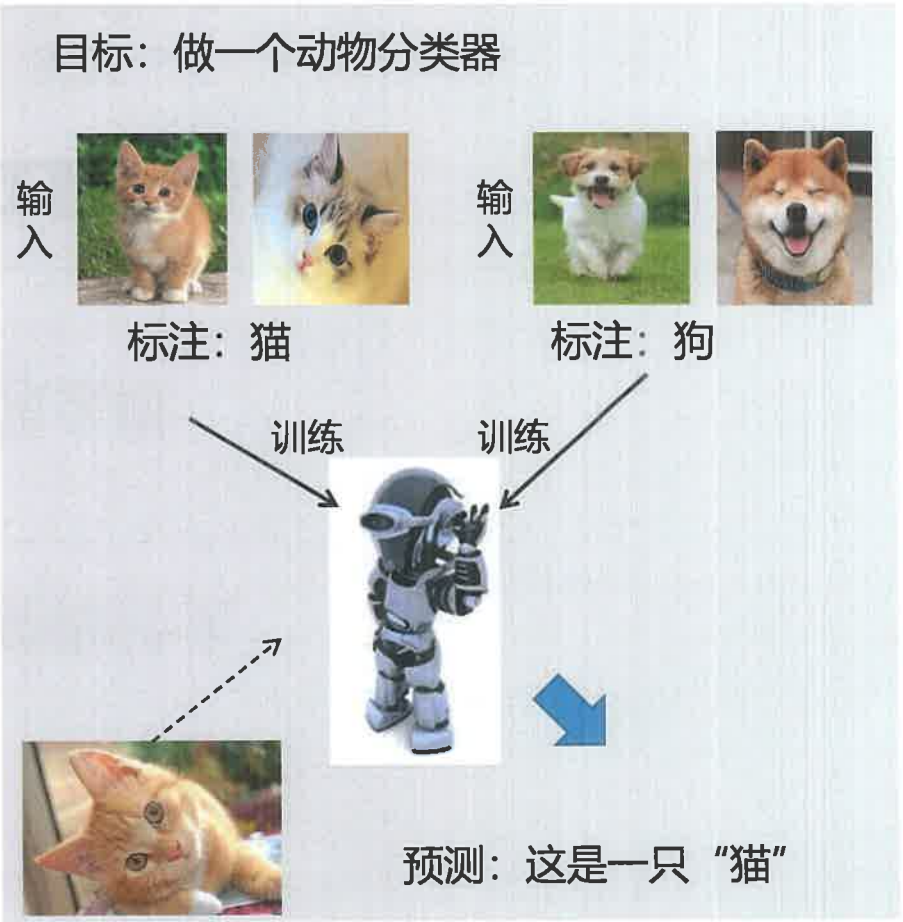
机器学习是人工智能的一个分支，

我们设计一组**计算机程序任务**，→ **机器学习模型**

使它能够根据提供的**训练数据**按照一定的方式来学习，→ **样本**

随着训练次数的增加，该程序可以在**性能**上不断改进，→ **评价指标**

通过参数优化的学习模型，能够用于**预测**相关问题→ **模型应用**



工具三：机器学习模型分类

机器学习模型按照数据标签的情况分为监督学习、无监督学习、半监督学习、强化学习等。

监督学习是指：输入数据集包含特征和**相应的标签**，即所有样本均有特征X和标签y。根据标签是连续值还是离散值，分为回归问题和分类问题。

无监督学习是指：输入数据集只包含特征，**没有相应的标签**，模型的目标是发现数据中的模式、结构或分布，自动寻找样本的潜在规律，比如聚类、异常检测和降维。

回归

标签为连续值

示例：

① 某个住宅房价预测

② PM2.5指数预测

分类

标签为离散值

示例：

① 贵宾客户是否流失

② 客户是否涉及欺诈

③ 图像主题打标（多分类）

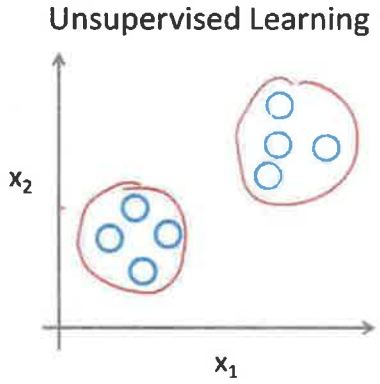
典型应用包括：

① 新闻分类

② 社交网络分群

③ 入侵检测

④ 反欺诈



工具三：机器学习基本术语

特征

```
df.sort_values(by='乘客年龄',ascending=False).head(10)
```

标签

	乘客编号	是否生还	船票舱位等级	乘客姓名	乘客性别	乘客年龄	乘客兄弟姐妹/配偶的个数	乘客父母/孩子的个数	船票号码	船票费用	所在船舱	登船港口
630	631	1	1	Barkworth, Mr. Algernon Henry Wilson	male	80.0	0	0	27042	30.0000	A23	S
851	852	0	3	Svensson, Mr. Johan	male	74.0	0	0	347060	7.7750	NaN	S
493	494	0	1	Artagaveytia, Mr. Ramon	male	71.0	0	0	PC 17609	49.5042	NaN	C
96	97	0	1	Goldschmidt, Mr. George B	male	71.0	0	0	PC 17754	34.6542	A5	C
116	117	0	3	Connors, Mr. Patrick	male	70.5	0	0	370369	7.7500	NaN	Q
672	673	0	2	Mitchell, Mr. Henry Michael	male	70.0	0	0	C.A. 24580	10.5000	NaN	S
745	746	0	1	Crosby, Capt. Edward Gifford	male	70.0	1	1	WE/P 5735	71.0000	B22	S
33	34	0	2	Wheadon, Mr. Edward H	male	66.0	0	0	C.A. 24579	10.5000	NaN	S
54	55	0	1	Ostby, Mr. Engelhart Cornelius	male	65.0	0	1	113509	61.9792	B30	C
280	281	0	3	Duane, Mr. Frank	male	65.0	0	0	336439	7.7500	NaN	Q

样本

泰坦尼克号乘客数据

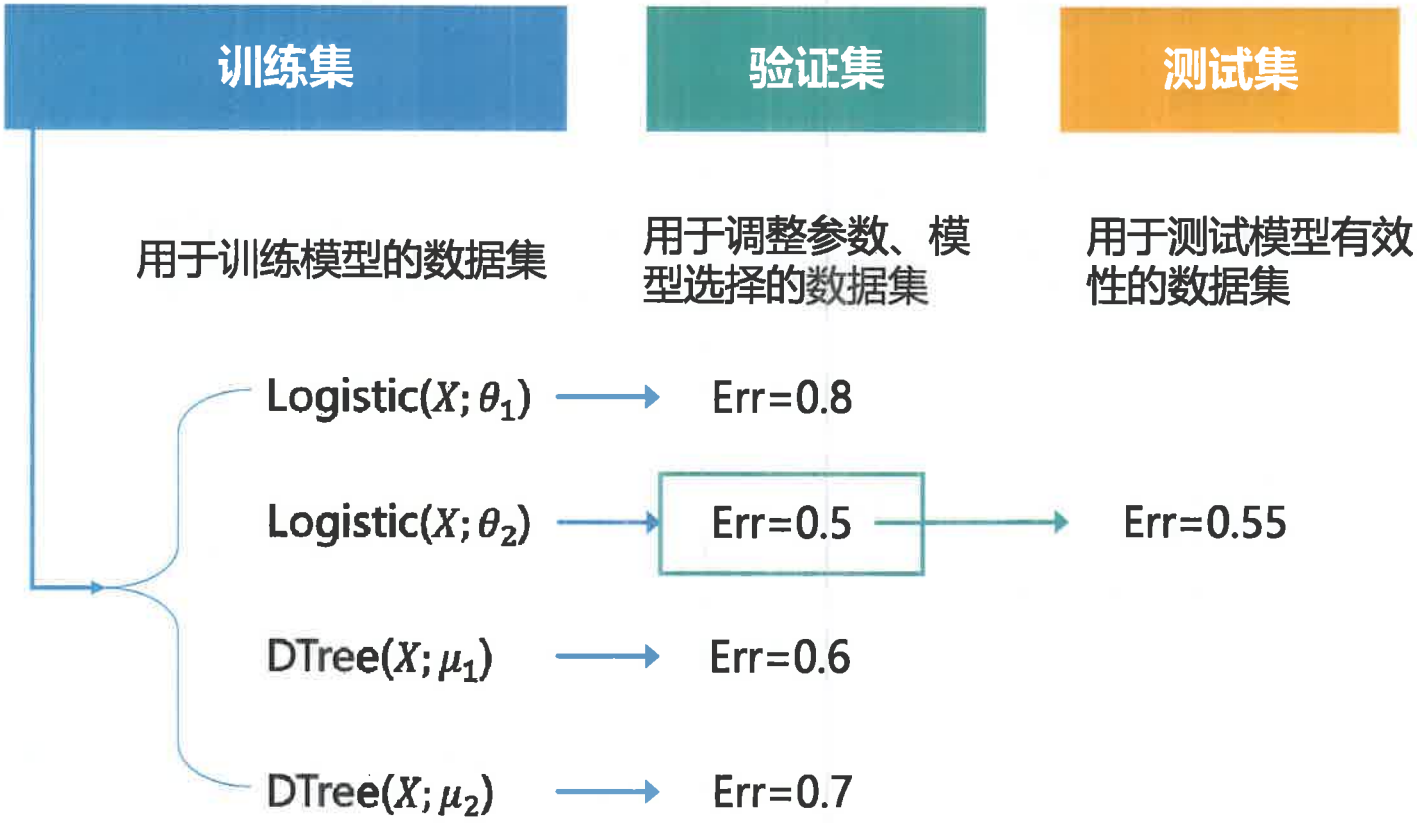
机器学习模型

定义特征与标签的关系，y=f(x)

✓训练过程：通过样本，创建或学习模型

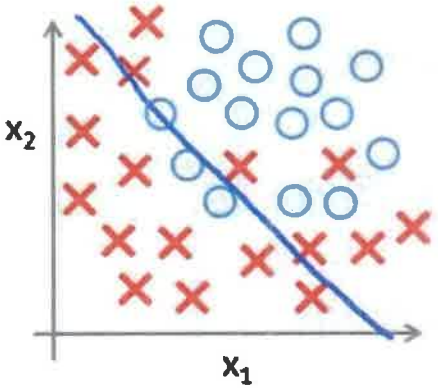
✓推断过程：将训练后的模型应用于新的样本进行预测，在新样本集上预测能力又称为**泛化**。

工具三：机器学习基本术语



工具三：机器学习基本术语

Example: Logistic regression

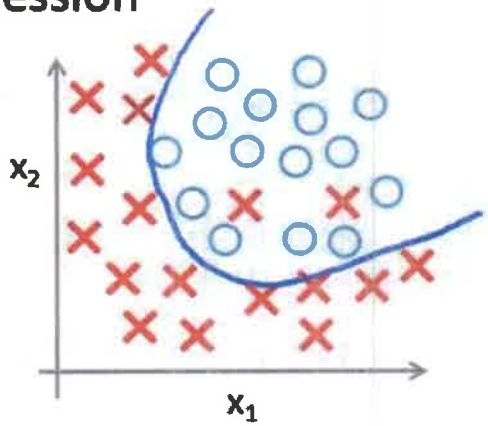


$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

(g = sigmoid function)

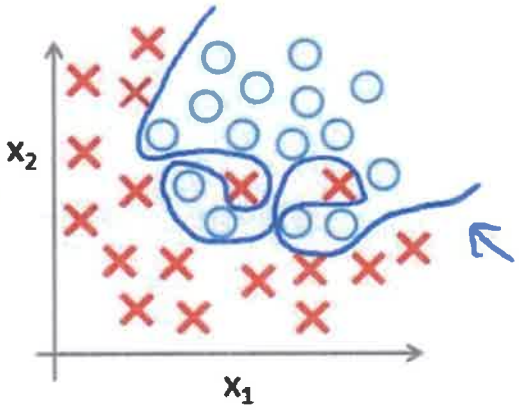
欠拟合

模型未能有效捕捉到特征，不能够很好地拟合数据



$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$

拟合



$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \dots)$

过拟合

模型过度拟合训练样本，无法对新的数据进行正确推断

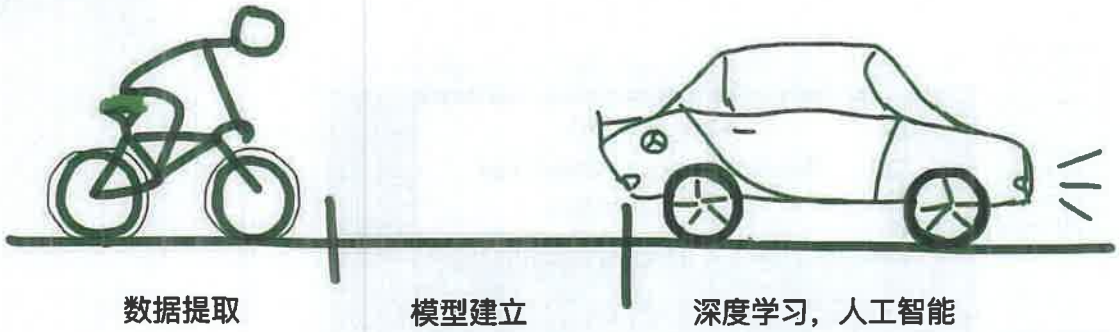
工具三：机器学习资料推荐

书籍：机器学习

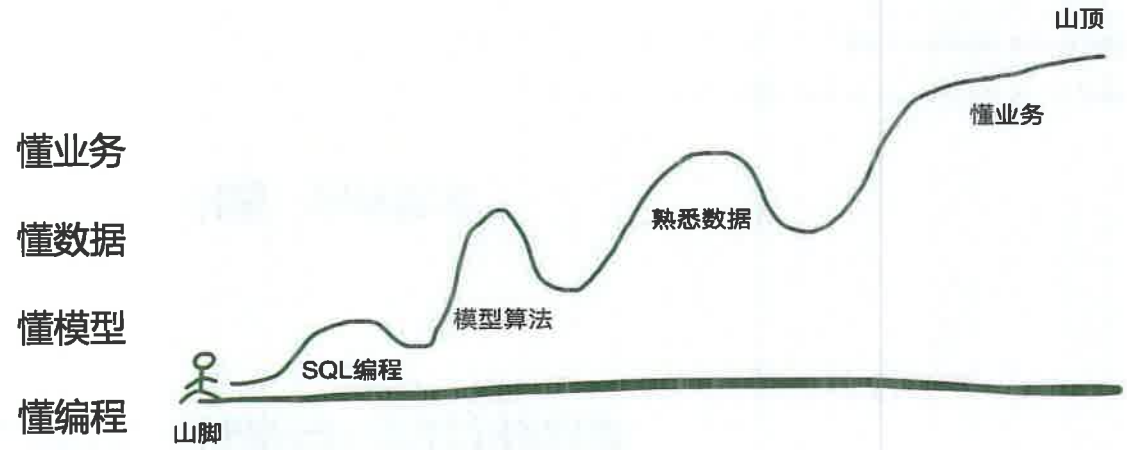
视频：B站-机器学习--李宏毅

数据分析师的成长如爬山

你之前听说的大数据分析：



真实的大数据分析：



工具	使用工具目的	目标
SQL	根据业务规则从数据库中提取大量数据，并计算指标。	会加工数据
Python	基于SQL提取的指标进行清洗和加工特征。	
机器学习模型	基于python加工好的特征，进行模型构建及评估应用。	建立分析模型
.....		

内部资料，严禁转载