

法律声明

□ 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：大数据分析挖掘

■ 新浪微博：ChinaHadoop



数据分析基本知识复习

目录

数据分析的基本概念

数据可视化

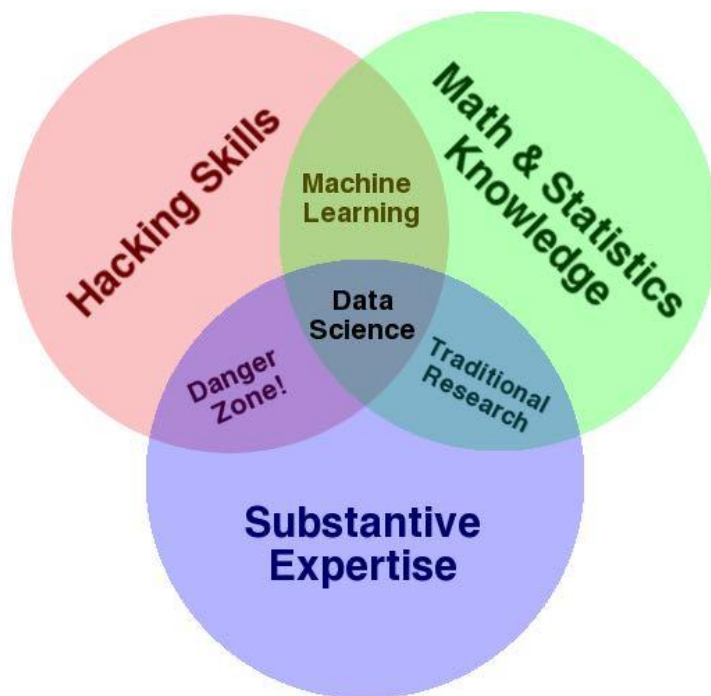
数据分析的常用模型

数据分析的常用工具

数据分析的基本概念

□ 数据分析的基本概念

针对特定的问题，用适当的学科知识从数据中提炼信息，形成结论



- 数理知识基础
- 数据获取、加工能力
- 行业知识

数据分析的基本概念

数据科学家的修炼途径



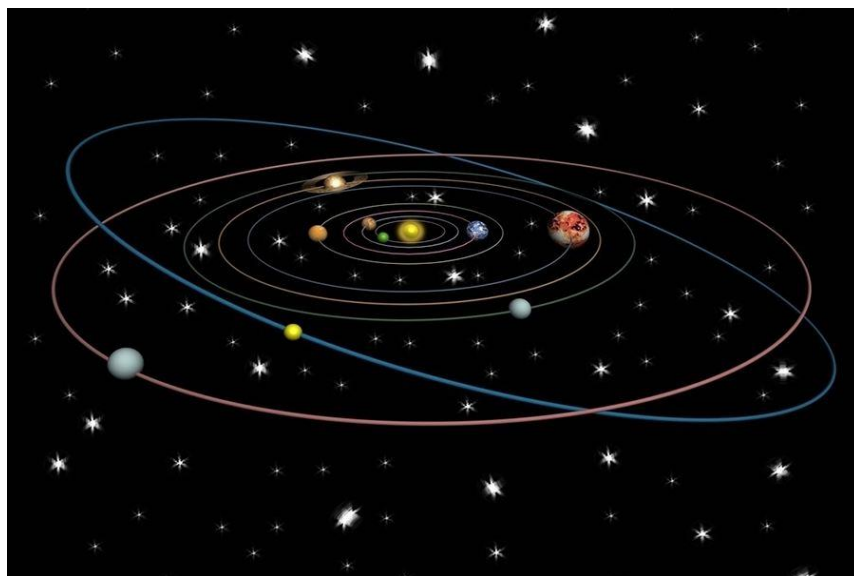
数据分析的基本概念

- 数据分析在具体业务中的使用环境
 - 业务逻辑清晰、指标明确
 - 可以转换成恰当的数据 / 数学 / 统计问题
 - 有足够的支撑数据
 - 熟悉模型 / 分析方法的局限性
 - 从场景中来，到场景中去

数据分析的基本概念

□ 数据分析的目的

从数据中挖掘规律、验证猜想、进行预测



案例：开普勒通过分析行星角位置的观测数据，找出了行星运动规律

数据分析的基本概念

数据获取的途径

公开
信息

外部
数据库

自有
数据库

调查
问卷

客户
数据

数据分析的基本概念

数据清洗

数据的可读性

数据的完整性

数据的唯一性

数据的权威性

数据的合法性

数据分析的基本概念

数据描述

类别型变量	取值的种类
	每种取值的分布
数值型变量	极值和分位点
	均值和标准差
	变量间相关性
通用	缺失度
	重复性

目录

数据分析的基本概念

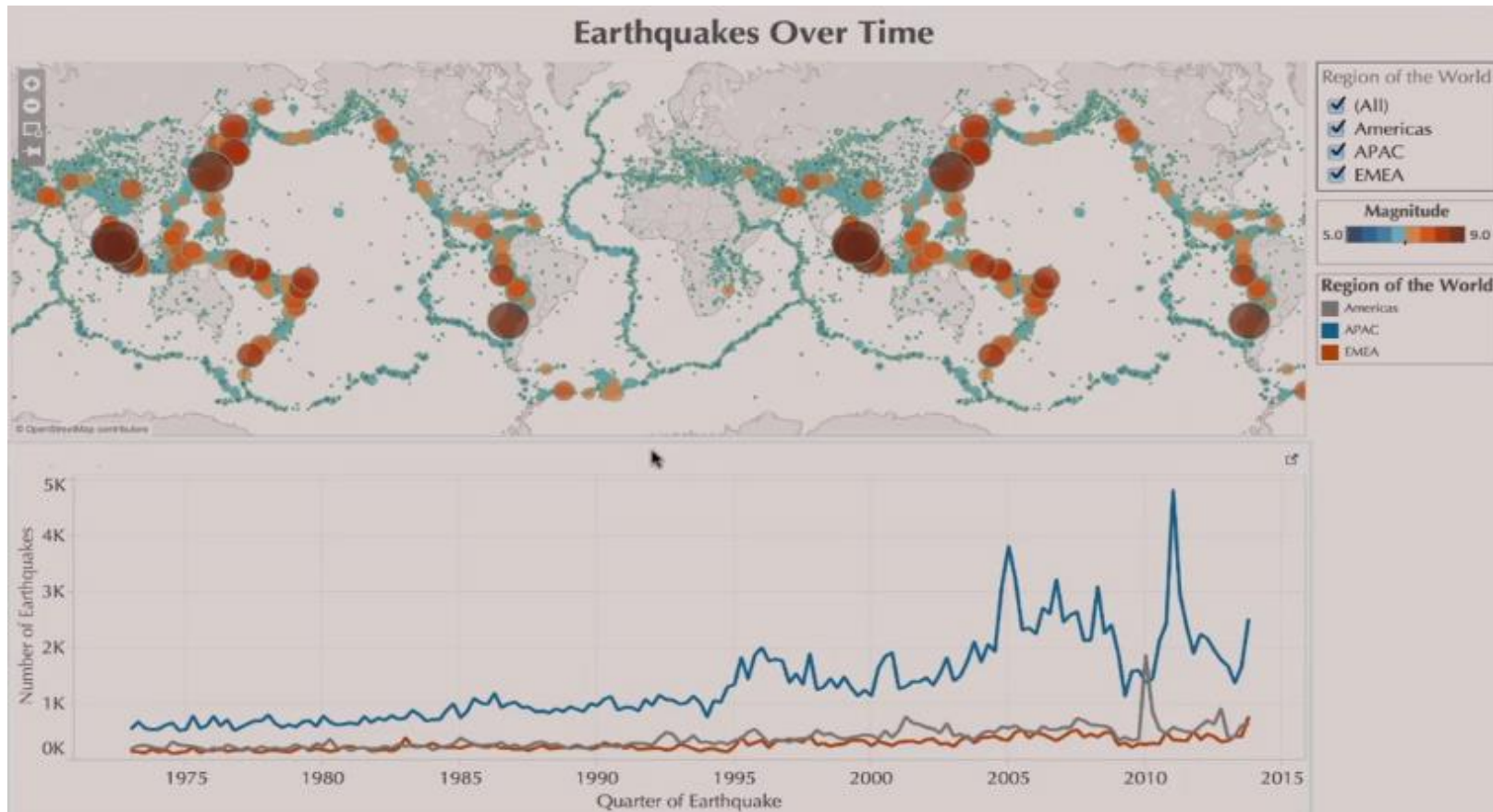
数据可视化

数据分析的常用模型

数据分析的常用工具

数据可视化

以视觉的形式展现数据的方式



数据可视化

专业工具

Tableau

Plotly

通用工具

EXCEL

R

Python

目录

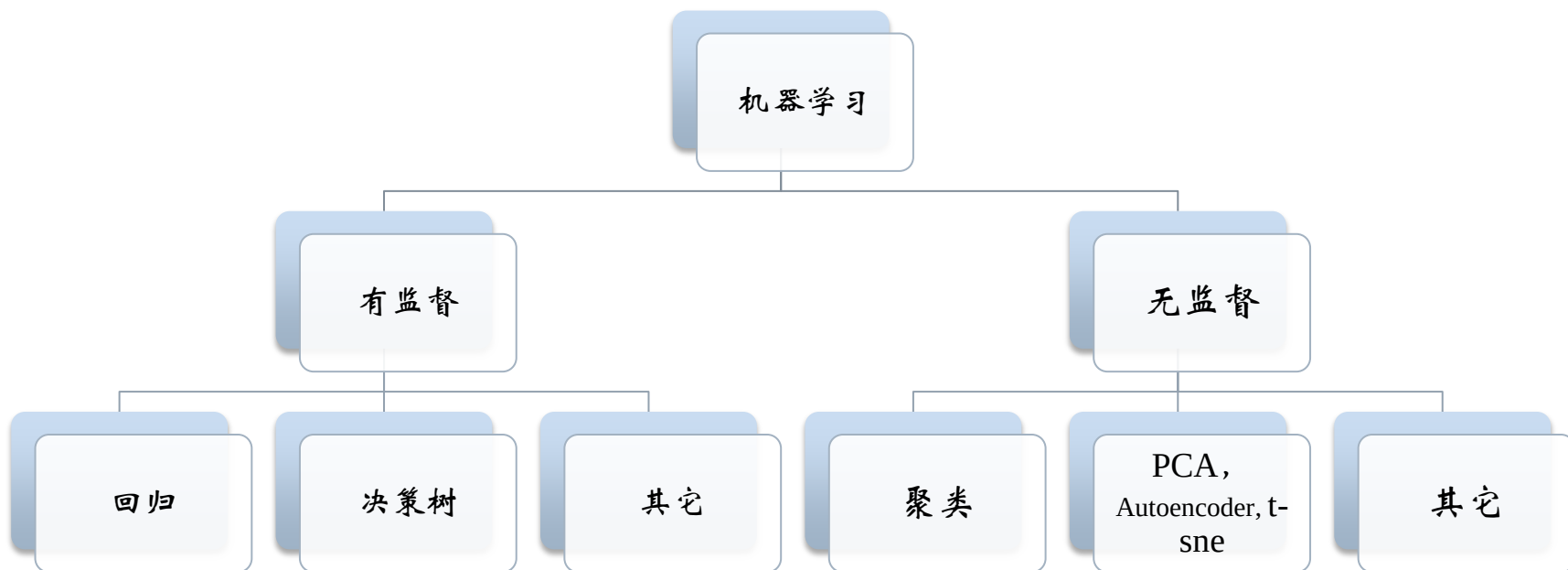
数据分析的基本概念

数据可视化

数据分析的常用模型

数据分析的常用工具

数据分析的常用模型



数据分析的常用模型

线性回归模型

$$y = \beta_0 + \sum_i \beta_i x_i + \epsilon$$

进一步推广：

$$E(y) = \mu = z(x)' \beta$$

拓展：

广义线性回归：

$$E(y) = \mu = h(z(x)' \beta)$$

其中 $h^{-1}()$ 成为连接函数， $y \sim$ 指数分布族

数据分析的常用模型

决策树模型

一个树结构（二叉树或非二叉树）。其每个非叶节点表示一个特征属性上的测试，每个分支代表这个特征属性在某个值域上的输出，而每个叶节点存放一个类别。

使用决策树进行决策的过程就是从根节点开始，测试待分类项中相应的特征属性，并按照其值选择输出分支，直到到达叶子节点，将叶子节点存放的类别作为决策结果

随机森林

- 基于树模型的集成模型
- 从待选属性集中随机选取子集，再进一步挑选最优分裂属性
- 加权平均／投票
- 泛化能力强

数据分析的常用模型

支持向量机

- 低维映射到高维
- 最优分割平面

对偶问题

- 核函数

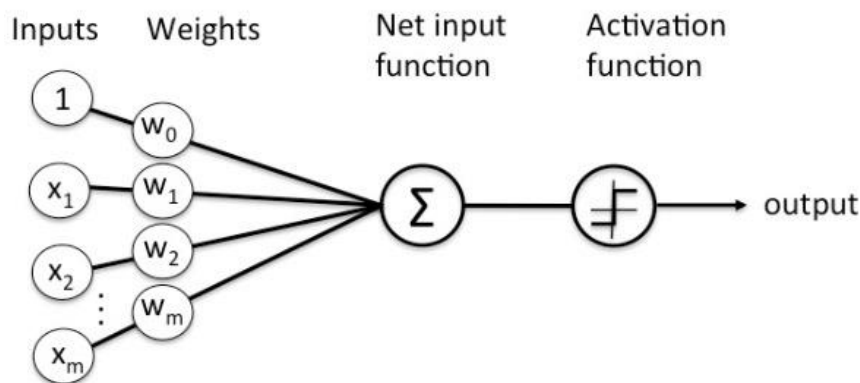
数据分析的常用模型

神经网络

一个包含了很多参数的、由若干个(简单)函数互相(嵌套)代入的模型

最简单的单层神经网络：感知机

$$y = f\left(\sum w_i x_i - \theta\right)$$



数据分析的常用模型

聚类分析

- 对无标记训练样本的学习来揭示数据的内在性质及规律，为进一步的数据分析提供基础
- 机器学习领域中新算法出现最多、最快的领域，因其不存在客观标准

类型

- 原型聚类，k-means
- 密度聚类，DBSCAN
- 层次聚类，AGNES

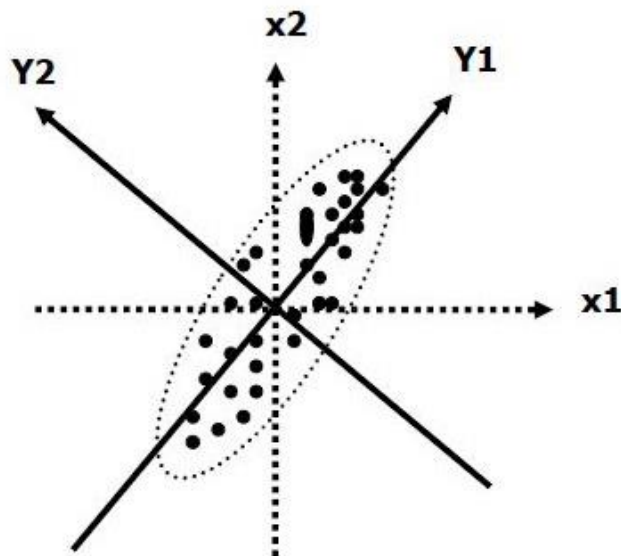
聚类集成

对多个聚类学习器进行集成，能有效地降低聚类假设与真实聚类结构不符、聚类过程中的随机性等因素带来的不利影响

数据分析的常用模型

主成分分析

将高维向量 x 通过特征向量矩阵投影到低维空间，表征为低维向量 y ，仅仅损失一点点信息



数据分析的常用模型

因子分析

能找出少数几个互不相关的综合变量来尽可能地反映原来数据所含有的绝大部分信息

假设 $x_1, x_2, \dots, x_n$ 是搜集反映信用卡用户消费类的指标, 则有:

$$x_i = a_{i,1}F_1 + a_{i,2}F_2 + a_{i,3}F_3$$

其中 F_1 表示持卡人的消费能力, F_2 表示持卡人的忠诚度, F_3 表示商家促销和发卡行优惠活动, $\{F_k\}$ 彼此不相关

在合理的数据预处理的条件下, 基于合理的假设求出 $\{a_{i,k}\}$

数据分析的常用模型

半监督学习

- 训练样本部分有标签，剩余没有
- 所有样本独立采样于同一数据源
- 利用学习器不依赖于外部、自动地利用未标记样本来提升学习性能

数据分析的常用模型

半监督支持向量机(Semi-Supervised SVM, S3VM)

TSVM(Transductive SVM)

带标签的数据集 $D_l = \{(x_1, y_1), \{x_2, y_2\}, \dots, \{x_l, y_l\}\}$

不带标签的数据集 $D_u = \{(x_{l+1}, y_{l+1}), \{x_{l+2}, y_{l+2}\}, \dots, \{x_m, y_m\}\}, y_i \in \{1, -1\}$

TSVM的学习目标是为 D_u 中的样本给出预测标记 $\hat{y} = \{\hat{y}_{l+1}, \hat{y}_{l+2}, \dots, \hat{y}_m\}$ 使得

$$\min_{w, b, \hat{y}, \xi} \frac{1}{2} \|w\|^2 + C_1 \sum_{i=1} \xi_i + C_2 \sum_{i=l+1} \xi_i$$

St

$$\begin{aligned} y_i(w^T x_i + b) &\geq 1 - \xi_i, i = 1, 2, \dots, l \\ \hat{y}_i(w^T x_i + b) &\geq 1 - \xi_i, i = l + 1, l + 2, \dots, m \\ \xi_i &\geq 0, i = 1, 2, \dots, m \end{aligned}$$

其中 (w, b) 确定了一个超平面, ξ 是松弛向量

目录

数据分析的基本概念

数据可视化

数据分析的常用模型

数据分析的常用工具

数据分析的常用工具

数据分析的利器

R

面向统计分析的编程语言，丰富的作图功能，开源

CRAN

Rstudio

`install.packages(), library()`

Python

自有软件，胶水语言，免费的MATLAB

`pip install yourPackage`

`import yourPackage as pkg`

`From yourPackage import yourFunction`



人生苦短，
我用python

疑问

□ 问题答疑：<http://www.xxwenda.com/>

■ 可邀请老师或者其他回答问题

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：大数据分析挖掘
- 新浪微博：ChinaHadoop

