北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

Data Analytics with R
语言数据分析

所谓学习，归类而已

艾新波 / 2018 • 北京

课程体系

- R语言数据分析
  - 上部：论道 
    - 第1章 气象万千、数以等观
    - 第2章 所谓学习、归类而已
    - 第3章 格言联璧话学习
    - 第4章 源于数学、归于工程
  - 中部：执具 
    - 第5章 工欲善其事必先利其器
    - 第6章 基础编程
    - 第7章 数据对象
  -
    - 第8章 人人都爱tidyverse
    - 第9章 最美不过数据框
  - 下部 博术 
    - 第10章 观数以形
    - 第11章 相随相伴、谓之关联
    - 第12章 既是世间法、自当有分别
    - 第13章 方以类聚、物以群分
    - 第14章 庐山烟雨浙江潮

从数据中学习规律/模式/模型/知识

韩家炜等.数据挖掘:概念与技术. 北京:机械工业出版社,P6

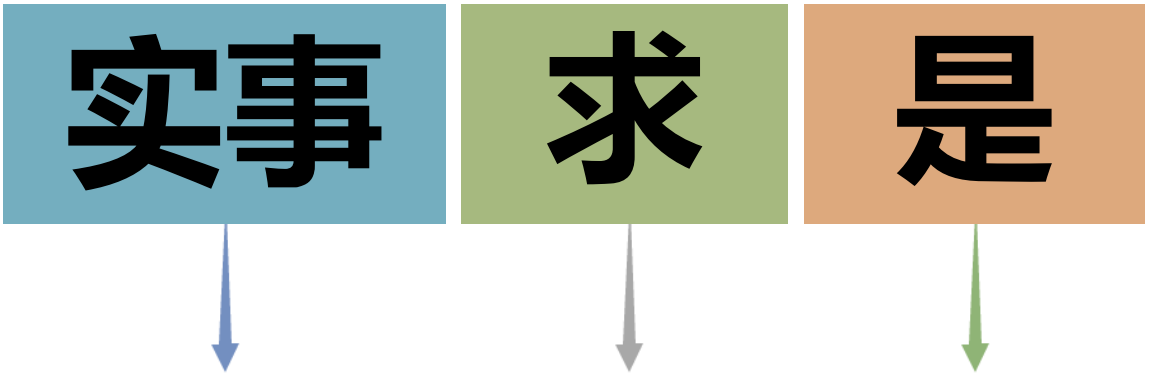
机器学习的一般过程：大智若愚

实事　求　是

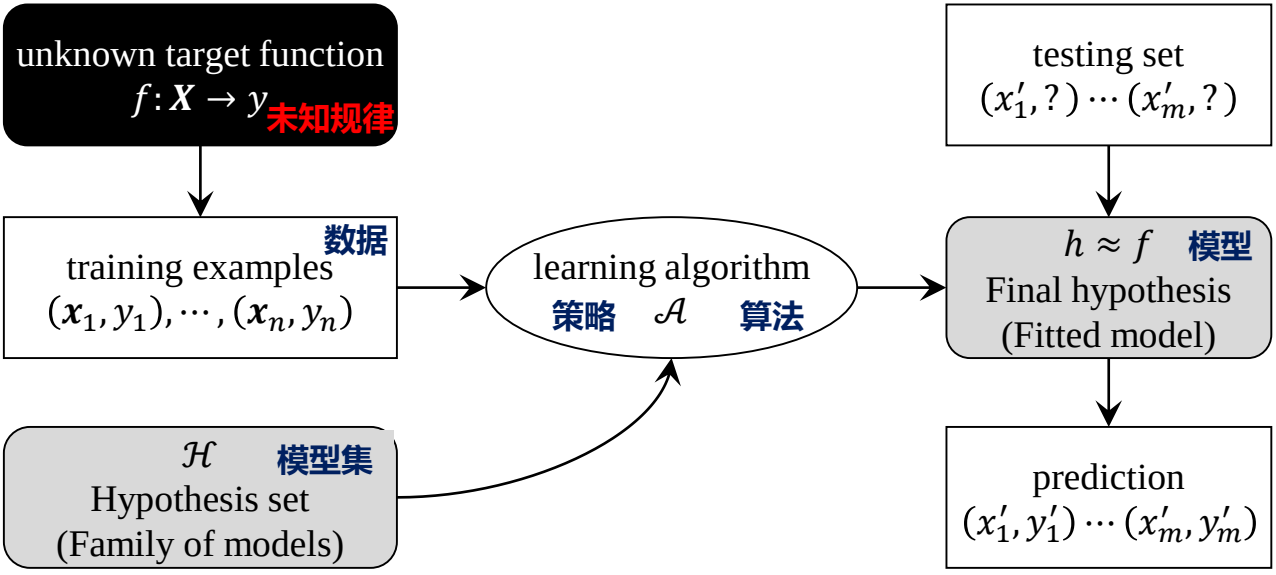
毛泽东：

实事就是客观存在着的一切事物，是就是客观事物的内部联系，即规律性，求就是我们去研究

机器学习的一般过程：大智若愚



机器学习的一般过程：大智若愚



与数学化归的思想如出一辙：化未知f为已知h，用已知逼近未知$h \approx f$

机器学习三要素

所谓的机器学习

通过**算法**从模型集里选出一个最贴近观察记录的**模型**，用来表示我们想要的关系结构

要刻画“贴近”的程度，需要一个量化标准，我们称之为**策略**

机器学习：作为一个搜索问题

容易看出，机器学习是真正在做：

research

机器学习：作为一个优化问题

学习策略与控制规律如出一辙：无非是“利用偏差，消除偏差”而已

经验风险最小化ERM： Empirical Risk Minimization

$$\min_{h\in\mathcal{H}}\left\{\frac{1}{|D|}\sum_{i=1}^{|D|}L\left(y_i,h\left(x_i\right)\right)\right\}$$

$$L\left(y_i,h\left(x_i\right)\right)=\left(y_i-h\left(x_i\right)\right)^2$$

$$L\left(y_i,h\left(x_i\right)\right)=\left|y_i-h\left(x_i\right)\right|$$

结构风险最小化SRM：Structural Risk Minimization

$$\min_{h\in\mathcal{H}}\left\{\left(\frac{1}{|D|}\sum_{i=1}^{|D|}L\left(y_i,h\left(x_i\right)\right)\right)+\lambda J\left(h\right)\right\}$$

何谓建模

模型真的是

一砖一瓦垒建而成的么

模型：构建还是选择?

数学建模

OR

数学选模

模型
不在于构建
而在于选择

从这个意义上讲，常说的“建模”二字，并不准确，毕竟模型从来就不是建构出来的^-^

建模就好比是量体裁衣

所谓建模，好比是量体裁衣

建模就好比是量体裁衣

衣服每年都有新款

不同类型的模型集（Model Family）也层出不穷

建模就好比是量体裁衣

款式（模型类）选定之后，显然要做的事情就是确定其参数

量好各个项参数至关重要（确定好参数便拟合好了模型）

机器学习/数据挖掘是从大量的数据中归纳出（先前未知的）有用或有趣关系结构（模式、模型、知识、规律、…）的过程

谢谢聆听

Thank you

教师个人联系方式

艾新波

手机：13641159546

QQ： 23127789

微信：13641159546

E-mail: 13641159546@126.com

axb@bupt.edu.cn

地址：北京邮电大学科研楼917室

课程网址：https://github.com/byaxb/RDataAnalytics