北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

Data Analytics with R
语言数据分析

源于数学、归于工程

艾新波 / 2018 • 北京

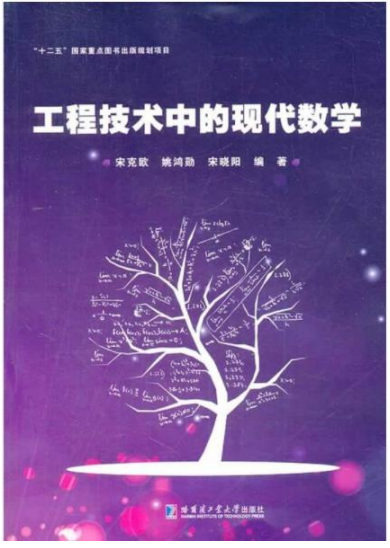
课程体系

- 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

数学是宇宙的语言

[日] 大栗博司 著，尤斌斌 译.用数学的语言看世界. 人民邮电出版社, 2017.pp.152.

数学是宇宙的语言



数学教育，特别是理工科各专业的数学教育的重要性是不言而喻的，它占据学校教育的时间最长。更重要的是，**科技工作者要终生和数学打交道**，现在大家正逐渐认识到“**被如此称颂的高技术本质上是一种数学技术**”

梦想很丰满，现实很骨感

$$\sqrt{2}$$

从小到大我们所学习到的数学

《教育部·全日制普通高级中学数学教学大纲》：数学是研究空间形式和数量关系的科学。数学能够处理数据、观测资料，进行计算、推理和证明，可提供自然现象、社会系统的数学模型。

《数学与统计学教学指导委员会·工科类本科数学基础课程教学基本要求》：数学是研究客观世界数量关系和空间形式的科学。随着现代科学技术和数学科学的发展，“数量关系”和“空间形式”具备了更丰富的内涵和更广泛的外延……

数学与机器学习存在着天然的联系

Mathematics is often defined as the science of space and number, as the discipline rooted in geometry and arithmetic. ... **Mathematics is the science of patterns**. The mathematician seeks patterns in number, in space, in science, in computers, and in imagination. Mathematical theories explain the relations among patterns; functions and maps, operators and morphisms bind one type of pattern to another to yield lasting mathematical structures. Applications of mathematics use these patterns to explain and predict natural phenomena that fit the patterns.

Articles

The Science of Patterns

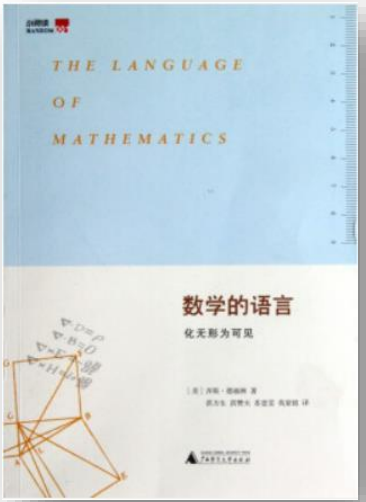
LYNN ARTHUR STEEN

The rapid growth of computing and applications has helped cross-fertilize the mathematical sciences, yielding an unprecedented abundance of new methods, theories, and models. Examples from statistical science, core mathematics, and applied mathematics illustrate these changes, which have both broadened and enriched the relation between mathematics and science. No longer just the study of number and space, mathematical science has become the science of patterns, with theory built on relations among patterns and on applications derived from the fit between pattern and observation.

Forces for Change

Lynn Arthur Steen. “The Science of Patterns, ” Science 1988,240: 611-616.

数学与机器学习存在着天然的联系

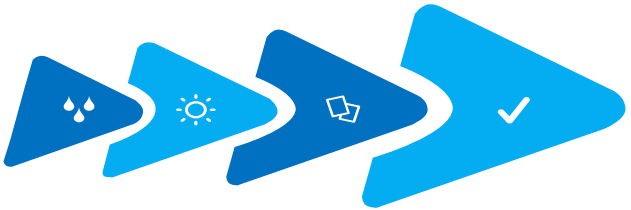


在最近大约三十年间，一个为大部分数学家所同意的有关数学的定义，才终于出现了：**数学是研究模式的科学**(science of patterns)。数学家的所作所为，就是去检视抽象的模式——数值模式、形状的模式、运动的模式、行为的模式、全国人口的投票模式、重复机会事件的模式等。

(美）德福林著；洪万生等译.数学的语言：化无形为可见.桂林：广西师范大学出版社，2013.pp.3.

机器学习：让数学情境化

解题



解决实际科学/工程问题

- 用数据来实现抽象数学理论的物化：情境化
- 四则运算、函数、内积、随机变量、条件概率、……

机器学习的两大数学视角：概率与几何

参数估计

贝叶斯

信息增益

密度估计

条件概率

极大似然估计

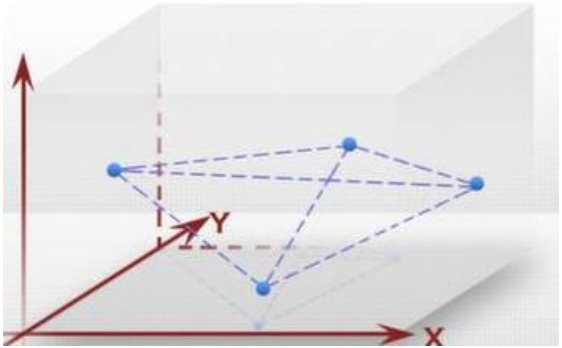
高斯混合模型

机器学习的两大数学视角：概率与几何

特征空间

数据空间密度

距离度量



空间投影

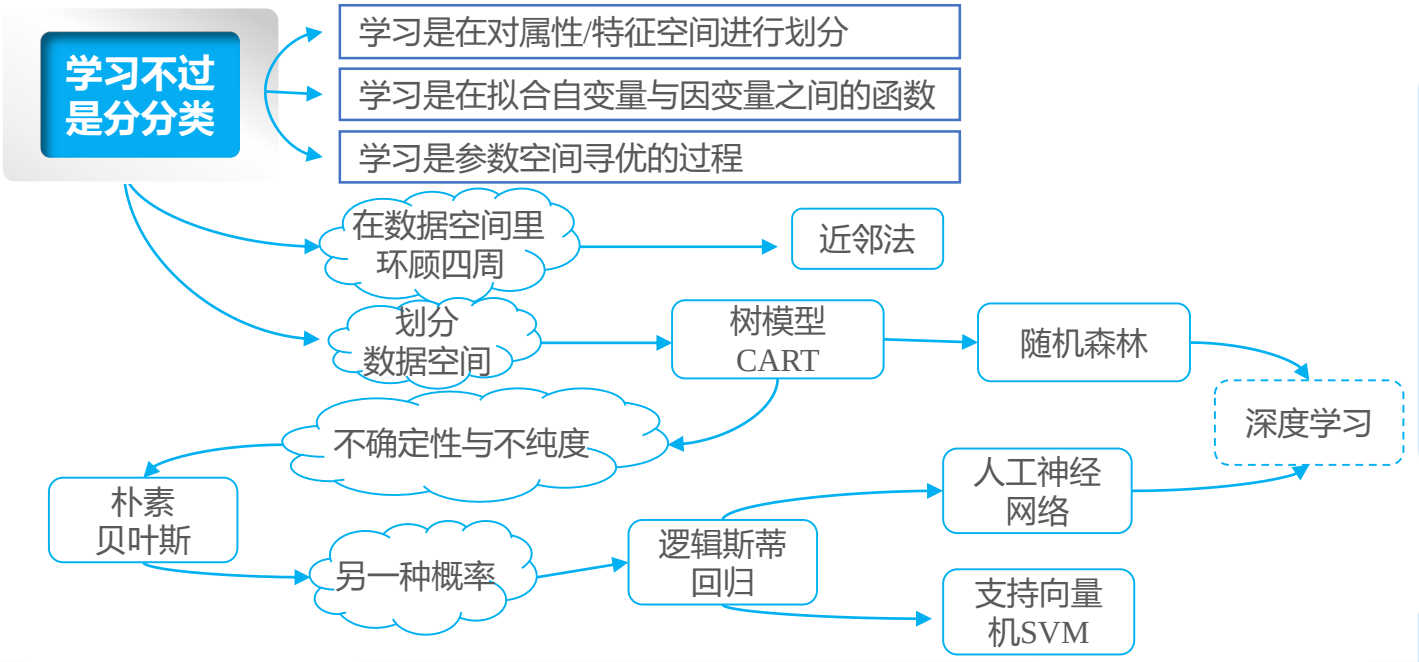
梯度下降

线性可分

类间距离

分类超平面

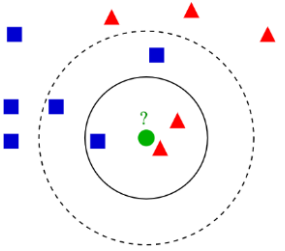
机器学习中最基本的数学视角：概率与几何



机器学习中最基本的数学视角：概率与几何

学习不过是分分类

- 学习是在对属性/特征空间进行划分
- 学习是在拟合自变量与因变量之间的函数
- 学习是参数空间寻优的过程



在数据空间里环顾四周 → 近邻法

划分数据空间 → 树模型CART → 随机森林 → 深度学习

不确定性与不纯度

朴素贝叶斯

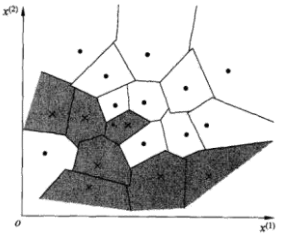
另一种概率 → 逻辑斯蒂回归 → 人工神经网络 → 深度学习

逻辑斯蒂回归 → 支持向量机SVM

机器学习中最基本的数学视角：概率与几何

学习不过是分分类

- 学习是在对属性/特征空间进行划分
- 学习是在拟合自变量与因变量之间的函数
- 学习是参数空间寻优的过程



在数据空间里环顾四周 → 近邻法

划分数据空间 → 树模型CART → 随机森林 → 深度学习

不确定性与不纯度 → 朴素贝叶斯

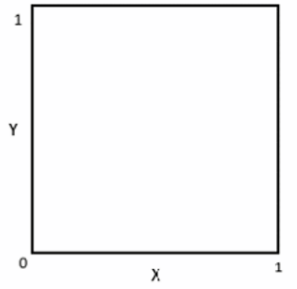
另一种概率 → 逻辑斯蒂回归 → 人工神经网络

逻辑斯蒂回归 → 支持向量机SVM

机器学习中最基本的数学视角：概率与几何

学习不过是分分类

- 学习是在对属性/特征空间进行划分
- 学习是在拟合自变量与因变量之间的函数
- 学习是参数空间寻优的过程



在数据空间里环顾四周 → 近邻法

划分数据空间 → 树模型CART → 随机森林 → 深度学习

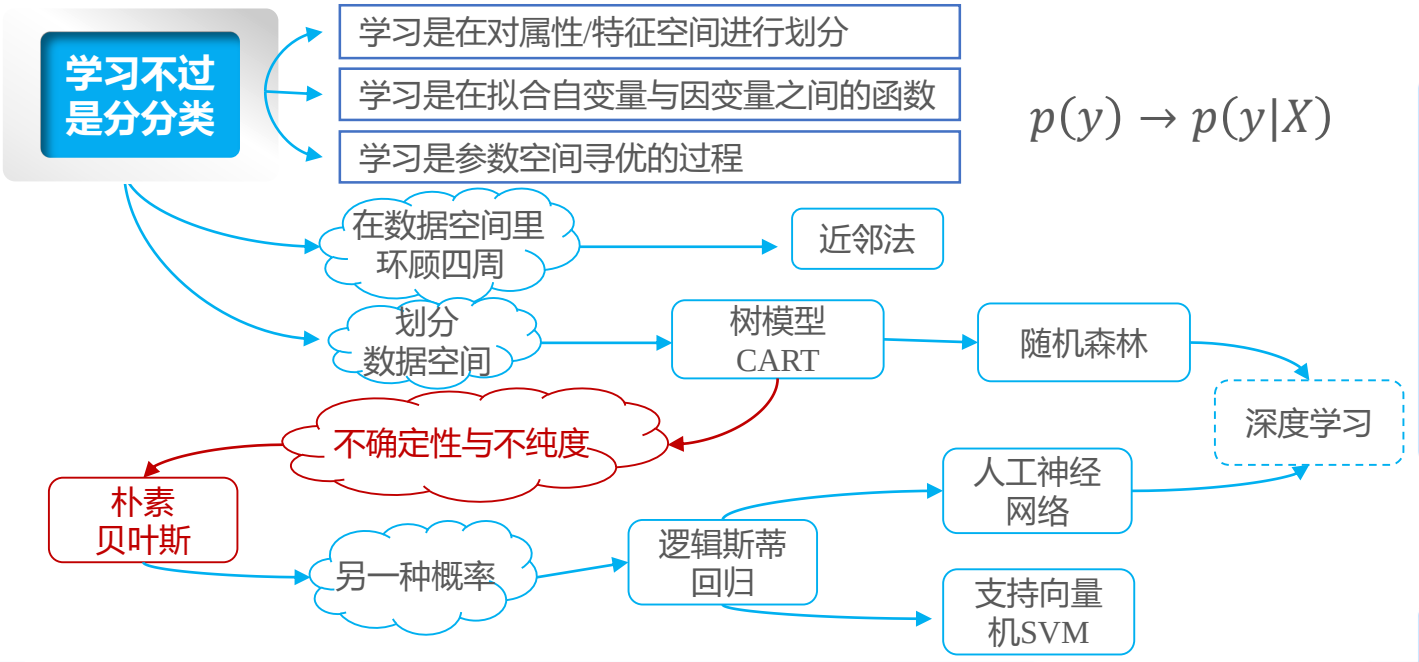
不确定性与不纯度

朴素贝叶斯

另一种概率 → 逻辑斯蒂回归 → 人工神经网络 → 深度学习

逻辑斯蒂回归 → 支持向量机SVM

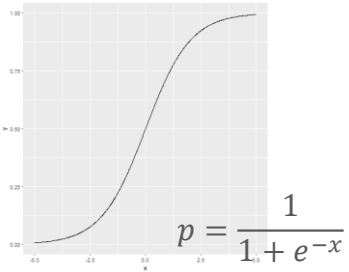
机器学习中最基本的数学视角：概率与几何



机器学习中最基本的数学视角：概率与几何

学习不过是分分类

- 学习是在对属性/特征空间进行划分
- 学习是在拟合自变量与因变量之间的函数
- 学习是参数空间寻优的过程



在数据空间里环顾四周 → 近邻法

划分数据空间 → 树模型CART → 随机森林 → 深度学习

不确定性与不纯度

朴素贝叶斯

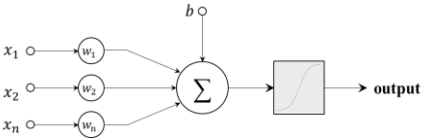
另一种概率 → 逻辑斯蒂回归 → 人工神经网络 → 深度学习

逻辑斯蒂回归 → 支持向量机SVM

机器学习中最基本的数学视角：概率与几何

学习不过是分分类

- 学习是在对属性/特征空间进行划分
- 学习是在拟合自变量与因变量之间的函数
- 学习是参数空间寻优的过程



$$O_j = \frac{1}{1+e^{-I_j}} = \frac{1}{1+e^{-(w^T x+b)}}$$

在数据空间里环顾四周 → 近邻法

划分数据空间 → 树模型 CART → 随机森林 → 深度学习

不确定性与不纯度 → 朴素贝叶斯

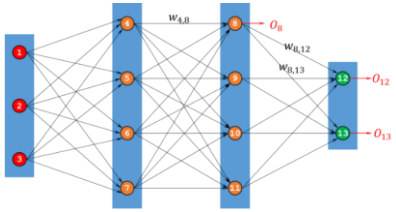
不确定性与不纯度 → 另一种概率 → 逻辑斯蒂回归 → 人工神经网络 → 深度学习

逻辑斯蒂回归 → 支持向量机SVM

机器学习中最基本的数学视角：概率与几何

学习不过是分分类

- 学习是在对属性/特征空间进行划分
- 学习是在拟合自变量与因变量之间的函数
- 学习是参数空间寻优的过程



在数据空间里环顾四周 → 近邻法

划分数据空间 → 树模型 CART → 随机森林 → 深度学习

不确定性与不纯度 → 朴素贝叶斯

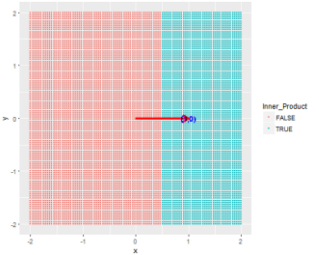
不确定性与不纯度 → 另一种概率 → 逻辑斯蒂回归 → 人工神经网络 → 深度学习

逻辑斯蒂回归 → 支持向量机SVM

机器学习中最基本的数学视角：概率与几何

学习不过是分分类

- 学习是在对属性/特征空间进行划分
- 学习是在拟合自变量与因变量之间的函数
- 学习是参数空间寻优的过程



在数据空间里环顾四周 → 近邻法

划分数据空间 → 树模型 CART → 随机森林 → 深度学习

不确定性与不纯度

朴素贝叶斯

另一种概率 → 逻辑斯蒂回归 → 人工神经网络 → 深度学习

逻辑斯蒂回归 → 支持向量机SVM

源于数学，归于工程

理论是灰色的，而工程之树长青

- 单凭语法不能激起诗意，光靠逻辑也不能产生思想
- 在机器学习/数据分析的领域，同样没有什么理论能让人自然地产生令人兴奋的idea
- 要找对感觉，最好的方式就是在具体的情境中实战
- 唯有实践才是完整的：**一个再小的工程，都有书本上的理论所覆盖不到的地方**
- 多一些工程思维，**由理性认识再次上升到感性认识**

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