北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

Data Analytics with R
语言数据分析

既是世间法、自当有分别

艾新波 / 2018 • 北京

课程体系

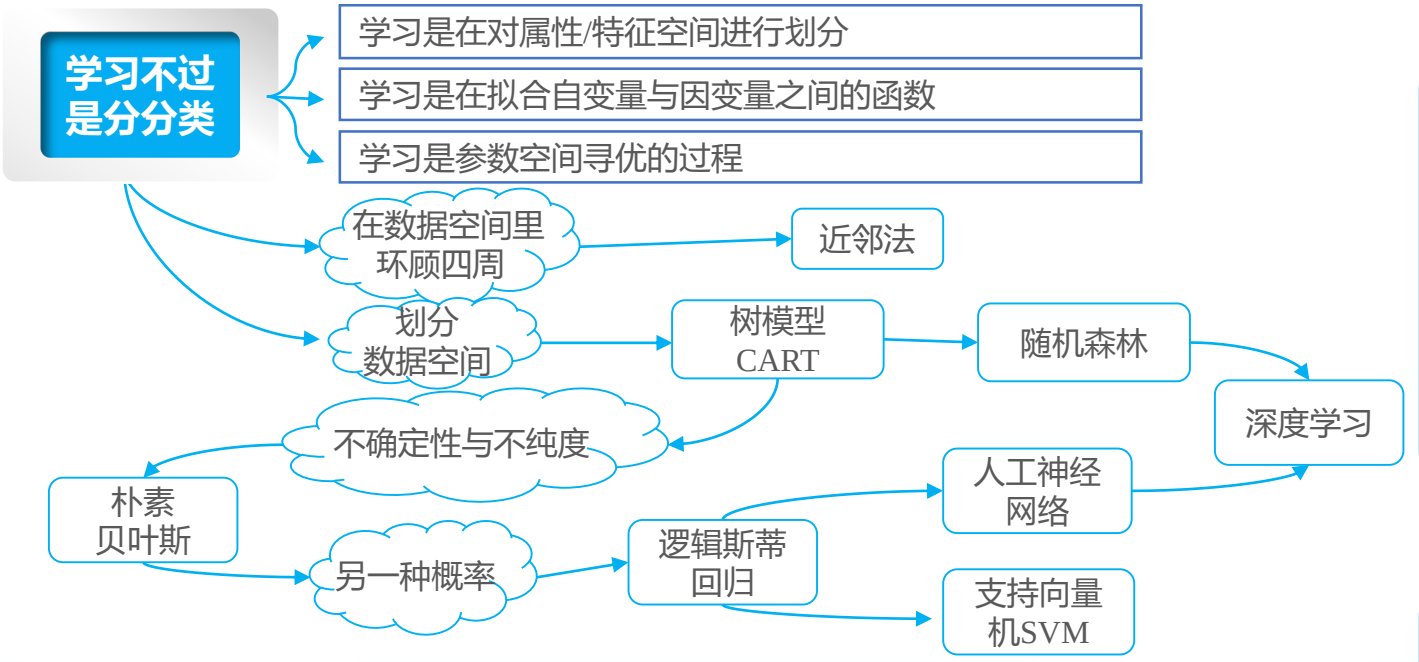
- R语言数据分析
  - 上部：论道 
    - 第1章 气象万千、数以等观
    - 第2章 所谓学习、归类而已
    - 第3章 格言联璧话学习
    - 第4章 源于数学、归于工程
  - 中部：执具 
    - 第5章 工欲善其事必先利其器
    - 第6章 基础编程
    - 第7章 数据对象
  -
    - 第8章 人人都爱tidyverse
    - 第9章 最美不过数据框
  - 下部 博术 
    - 第10章 观数以形
    - 第11章 相随相伴、谓之关联
    - 第12章 既是世间法、自当有分别
    - 第13章 方以类聚、物以群分
 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

分类与回归模型

分类回归模型，有数十种。各类改进的模型，更是数以百计

```
available_models <- modelLookup()
unique(available_models$model)
#> [1] "ada"                    "AdaBag"
#> [3] "adaboost"               "AdaBoost.M1"
.....
#> [235] "xgbLinear"              "xgbTree"
#> [237] "xyf"
length(unique(available_models$model))
#> [1] 237
```

算法模型


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
人工神经网络
另一种概率
逻辑斯蒂回归
支持向量机SVM

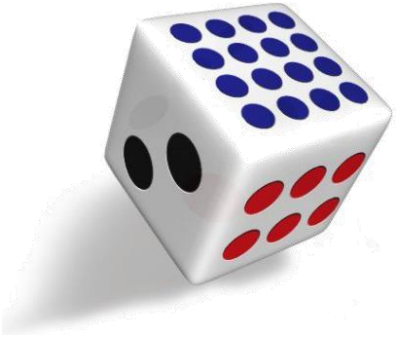
有生于无：想想那些没有模型的岁月

艾新波曾经就读的沙洲中学

总共有六个班

请你预测一下

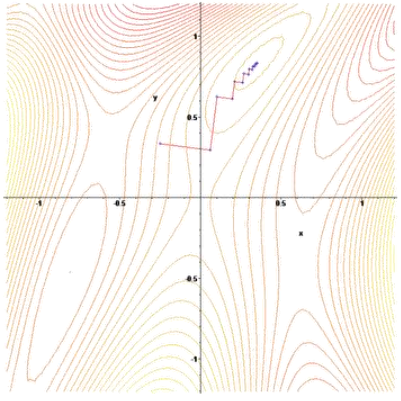
那个曾经的青涩少年属于几班？

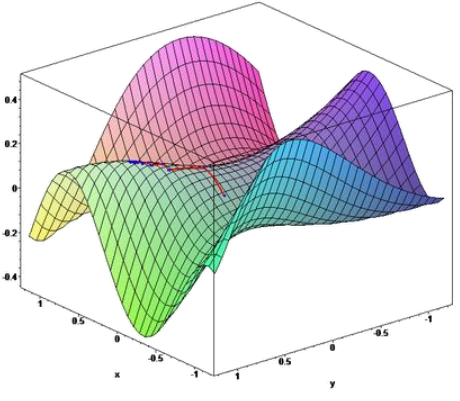


有生于无：

当你无从下手时，基本上也就是掷色子——一切只好**随缘**了

有生于无：想想那些没有模型的岁月





有生于无：

梯度下降亦然——千里之行、始于随机（缘）

有生于无：想想那些没有模型的岁月

有生于无：

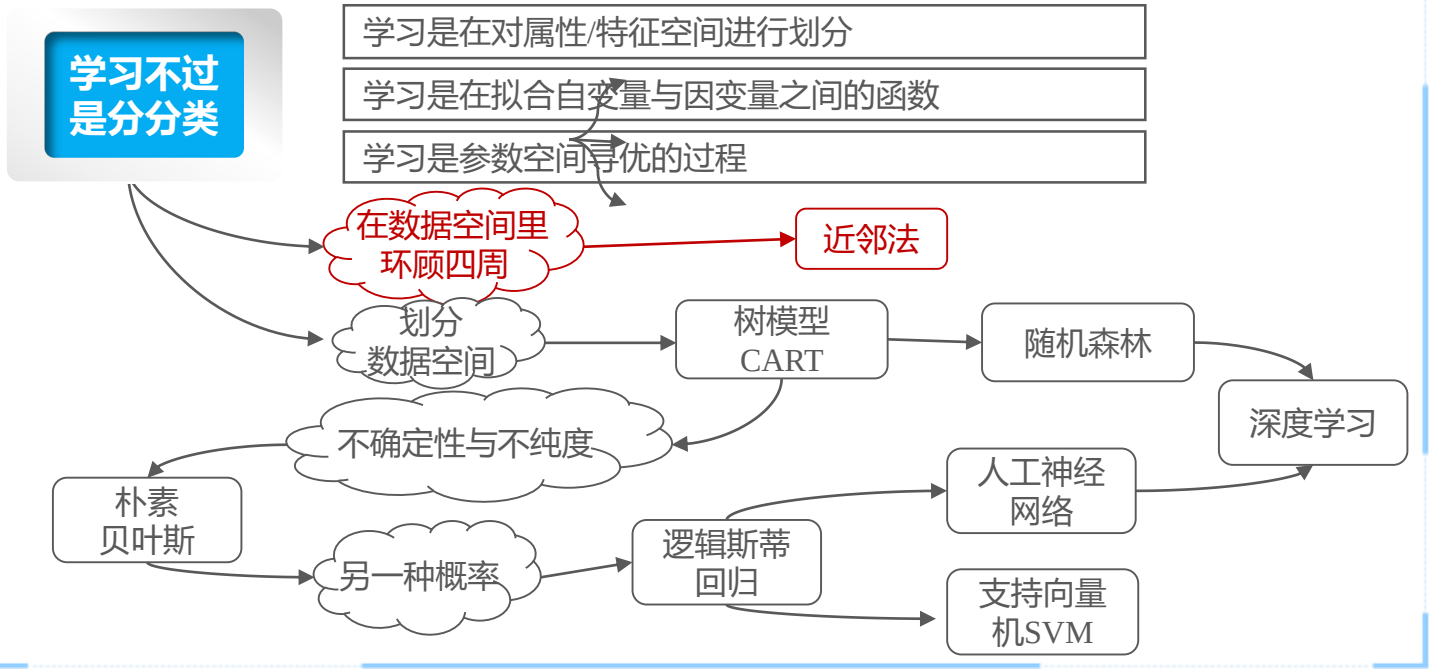
在没有模型的时候，用来做预测的基本上就是均值和众数

有生于无：想想那些没有模型的岁月

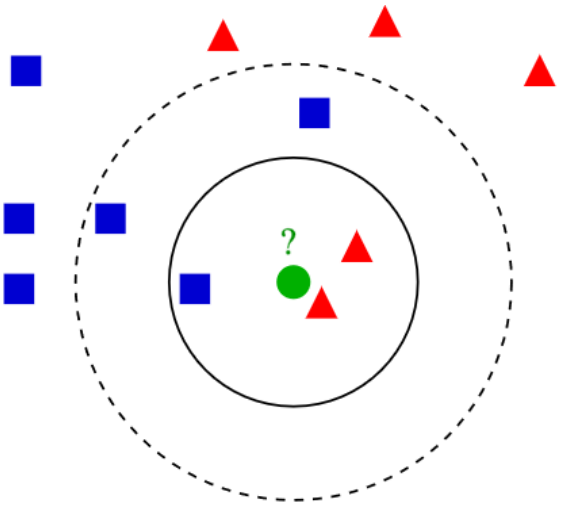
有生于无：

在后续课程将会发现，**在诸多模型中，都仰仗均值和众数的预测功能**

算法模型


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
支持向量机SVM

k-最近邻分类



有生于无：kNN也没好得了多少

在数据空间里环顾四周，看看周边小伙伴额头大多贴着什么标签

k-最近邻分类

听其言、观其行

发现这只动物走路也像鸭子、叫起来也像鸭子，那么……

k-最近邻分类

输入：最近邻数目 k，训练集 S，测试集 T

输出：对测试集 T 中所有测试样本预测其类标号值

(1) for 每个测试样本 $z = (\boldsymbol{X}', y') \in T$

(2) 计算 z 和每个训练样本 $(\boldsymbol{X}, y) \in S$ 之间的距离 $d(\boldsymbol{X}, \boldsymbol{X}')$

(3) 选择离 z 最近的 k 最近邻集合 $S_z \subseteq S$

(4) 多数表决 $y' = \underset{v}{\operatorname{argmax}} \sum_{(X_i, y_i) \in D_z} I(v = y_i)$

(5) end for

邻近性的度量

邻近性用距离度量

由于是基于距离的比较，因而距离的度量至关重要

欧几里得距离：$dist(X, X') = \sqrt{\sum_{i=1}^{n}\left(x_i - x_i'\right)^2}$

曼哈顿距离：$dist(X, X') = \sum_{i=1}^{n}\left|x_i - x_i'\right|$

Weighted kNN

尽管由 k 个最近邻点决定，但距离不同，这些点决定能力应有所区别

距离加权表决：

$$y' = \underset{v}{argmax} \sum_{(X_i, y_i) \in D_z} w_i \times I(v = y_i)$$

$$w_i = K\big(d(X', X_i)\big)$$

不同的核函数：

Gauss kernel: $\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{d^2}{2}\right)$

inversion kernel: $|d|^{-1}$

k的确定

可以通过实验来确定

从k=1开始，使用检验集估计分类器的错误率

重复该过程，每次 k 增加1，允许增加一个近邻

选取产生最小错误率的k

一般而言，训练元组越多，k 的值越大

需要进行多次训练，找到 k 值

k-近邻一般采用k为奇数

跟投票表决一样，避免因两种票数相等而难以决策

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