北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

Data Analytics with R
语言数据分析

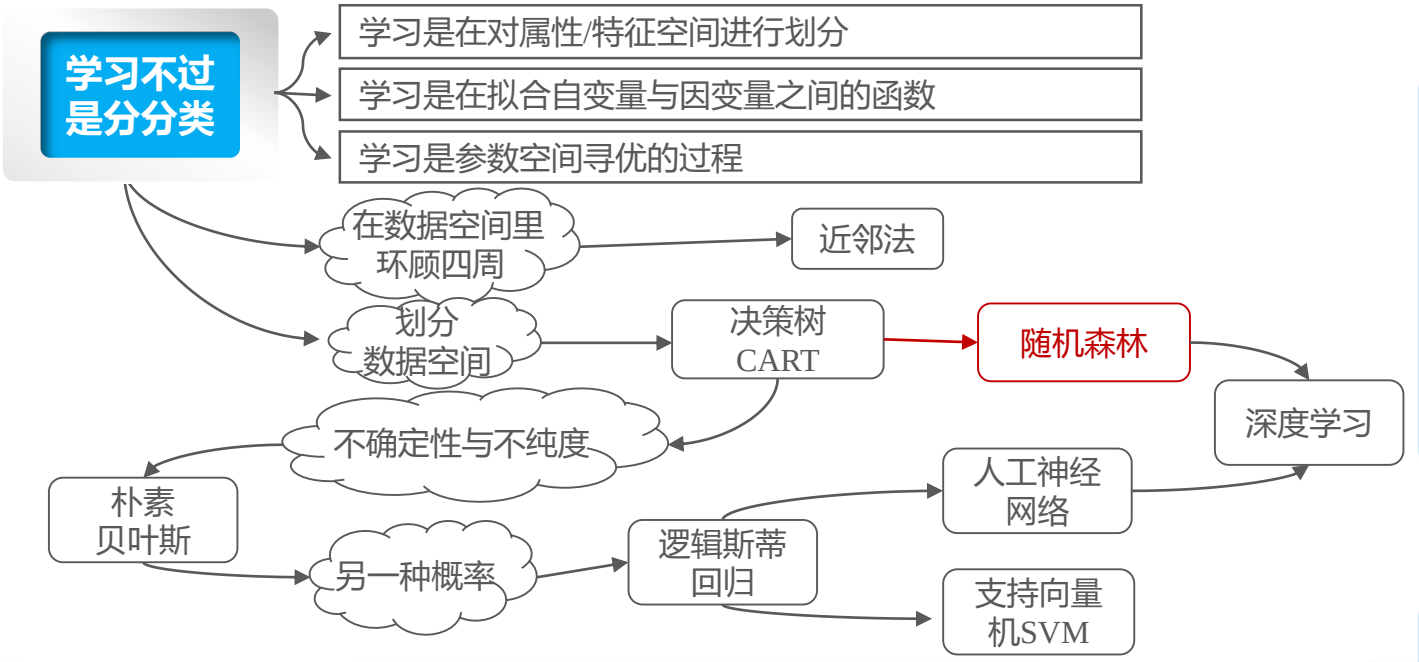
既是世间法、自当有分别

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

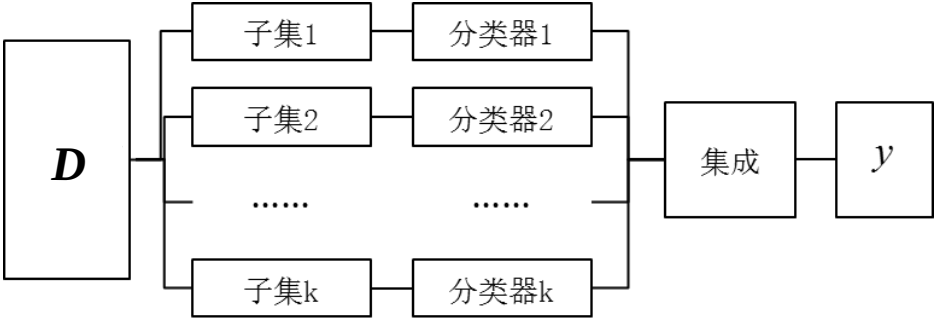
算法模型


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
决策树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

集成学习

集成学习，Ensemble Learning ，又称分类器组合学习

三个臭皮匠，顶个诸葛亮



将多个弱分类器集成为强分类器

主要算法包括boosting，bagging，随机森林，选择性集成等

回到问题的原点

Sometimes an idea or concept seems very complicated because of all the complicated maths you have to use, but at the core there is usually a very simple idea. …. In the sense that it's trying to ask the elementary questions and to reduce everything to the simplest questions you can imagine.

Yan LeCun

https://www.rsipvision.com/ComputerVisionNews-2018November/10/

回到原点：看看Breiman最初的灵感与直觉

我发现线性回归中的子集选取问题非常不稳定。如果你稍微扰动一下数据，回归的五个最好的变量就会变成另外的五个……我们可以通过给数据加扰动把它稳定下来，得到五个最好的预测变量。然后再次扰动数据，再拿到五个最好的变量，最后把这些五变量的预测器作一个平均……这样做的效果确实很好……

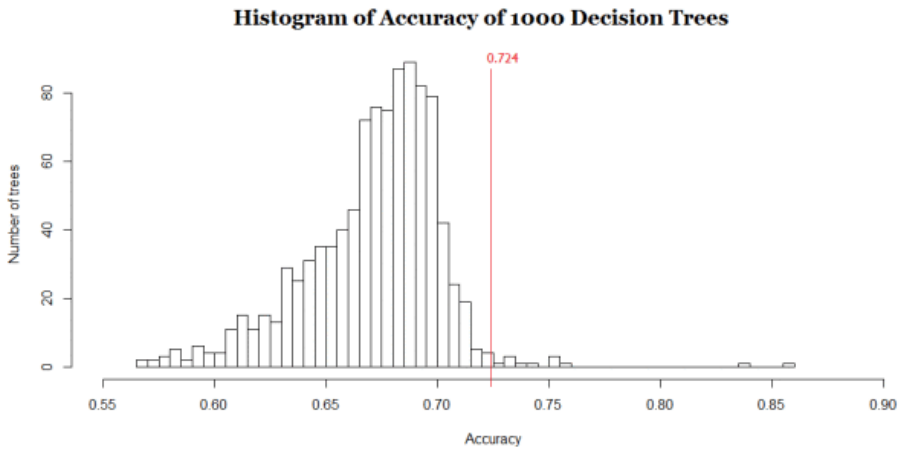
Richard Olshen. A conversation with Leo Breiman. Statistical Science 2001, 16(2): 184-198.

回到原点：看看Breiman最初的灵感与直觉

而CART也有同样的特性。你知道的，如果你稍微改变数据，你会得到一棵非常不同的树。所以我就想，“那好，为什么我不试试在CART上做同样的事情呢？如果我改变数据生成一棵树，然后再次改变数据，生成另一棵树，最后取它们的均值或者让它们投票选出票数最多的类，没准我能提高准确性……

Richard Olshen. A conversation with Leo Breiman. Statistical Science 2001, 16(2): 184-198.

随机森林



Histogram showing the accuracy of 1000 decision trees. While the average accuracy of decision trees is 67.1%, the random forest model has an accuracy of 72.4%, which is better than 99% of the decision trees

https://algobeans.com/2016/08/25/random-forest-tutorial/

随机森林

设：给定d个元组的训练集D，为组合分类器产生k棵决策树

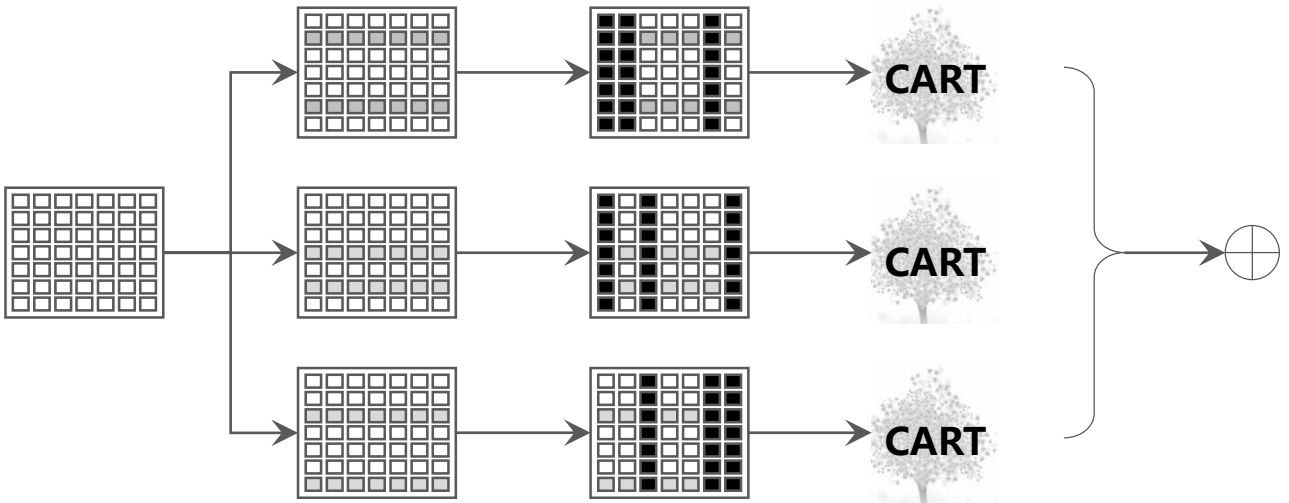
(1)使用有放回抽样生成训练集D_i，每个D_i都是D的一个自助样本，某些元组在D_i中出现多次，而某些元组不出现

(2)每个自助样本集生长为单棵分类树（随机选取分裂属性集）：设F是用来在每个节点决定划分的属性数，其中F远小于可用属性数。为构造决策树分类器M_i，在每个节点随机选择F个属性作为该节点划分的候选属性。使用CART方法增长树，增长到最大规模且不剪枝

有生于无
不厌其烦

(3) 采用简单多数投票法得到随机森林的结果

随机森林



行：有放回抽样　　列：选取少量特征　　完全生长　　多数表决

R语言实现

```
library(randomForest)
set.seed(2012)
imodel <- randomForest(wlfk~.,
                       ntree = 220,
                       data = cjb[train_set_idx, ])
predicted_train <- predict(imodel,
                           newdata = cjb[train_set_idx,],
                           type = "response")
Metrics::ce(cjb$wlfk[train_set_idx], predicted_train)
#>[1] 0
```

R语言实现

```
predicted_test <- predict(imodel,
                          newdata = cjb[-train_set_idx,],
                          type = "response")
Metrics::ce(cjb$wlfk[-train_set_idx], predicted_test)
#> [1] 0.1845494
```

R语言实现

```
library(randomForest)
sp <- Sys.time() #记录开始时间
cat("\n[Start at:", as.character(sp))
for (i in 1:length(kfolds)) {
  curr_fold <- kfolds[[i]] #当前这一折
  train_set <- cjb[-curr_fold,] #训练集
  test_set <- cjb[curr_fold,] #测试集
  imodel_kfold <- randomForest(wlfk~., ntree = 220,
                               data = train_set)
  predicted_train <-  predict(imodel_kfold,
                              train_set, type = "response")
```

R语言实现

```
  imetrics("randomForest", "Train",
           predicted_train, train_set$wlfk)
  predicted_test <- predict(imodel_kfold,
                            test_set, type = "response")
  imetrics("randomForest", "Test",
           predicted_test, test_set$wlfk)
}
ep <- Sys.time()
cat("\tFinised at:", as.character(ep), "]\n")
cat("[Time Ellapsed:\t",
    difftime(ep, sp, units = "secs"), " seconds]\n")
```

R语言实现

```
#> 41 randomForest Train 1.0000000  0.0000000
#> 42 randomForest  Test 0.7820513  0.2179487
#> 45 randomForest Train 1.0000000  0.0000000
#> 46 randomForest  Test 0.8461538  0.1538462
#> 49 randomForest Train 1.0000000  0.0000000
#> 50 randomForest  Test 0.8051948  0.1948052
#> 51 randomForest Train 1.0000000  0.0000000
#> 52 randomForest  Test 0.8181818  0.1818182
.....
#> 59 randomForest Train 1.0000000  0.0000000
#> 60 randomForest  Test 0.7662338  0.2337662
```

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