北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

Data Analytics with R
语言数据分析

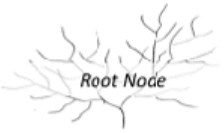
既是世间法、自当有分别

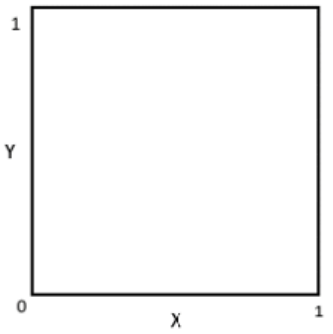
艾新波 / 2018 • 北京

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

不纯度与不确定性的减少

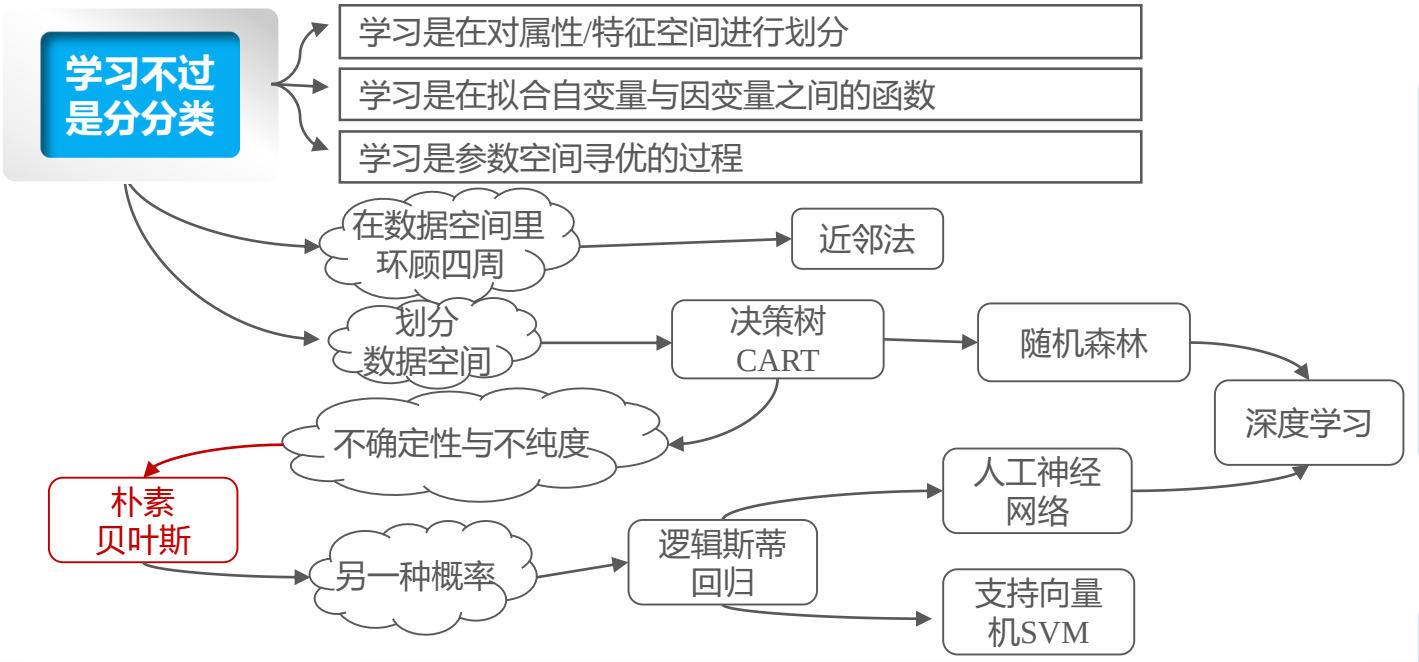




另一种不确定性的减少：有了更多的证据

$$p(y) \rightarrow p(y|X)$$

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

贝叶斯公式用于分类

贝叶斯公式：

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

换一种形式：

$$P(\text{类别}\mathbf{1}|\text{特征组合}) = \frac{P(\text{特征组合}|\text{类别}\mathbf{1})P(\text{类别}\mathbf{1})}{P(\text{特征组合})}$$

$$P(\text{类别}\mathbf{2}|\text{特征组合}) = \frac{P(\text{特征组合}|\text{类别}\mathbf{2})P(\text{类别}\mathbf{2})}{P(\text{特征组合})}$$

朴素贝叶斯

贝叶斯公式：

$$p(C|F_1,\cdots,F_n)=\frac{p(C)p(F_1,\cdots,F_n|C)}{p(F_1,\cdots,F_n)}$$

分母不发挥作用：

$$p(C|F_1,\cdots,F_n)\propto p(C)p(F_1|C)p(F_2|C,F_1)\cdots p(F_n|C,F_1,\cdots,F_{n-1})$$

朴素一点——特征之间相互独立：

$$p(F_i|C,F_j)=p(F_i|C)$$

朴素贝叶斯

于是有：

$$p(C|F_1,\cdots,F_n) \propto p(C)p(F_1|C)p(F_2|C)\cdots p(F_n|C) = p(C)\prod_{i=1}^{n} p(F_i|C)$$

用于分类：

$$classify(f_1,f_2,\cdots,f_n) = \underset{c}{argmax}\, p(C=c)\prod_{i=1}^{n} p(F_i=f_i|C=c)$$

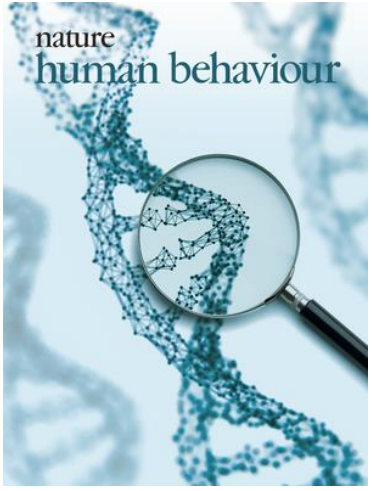
朴素贝叶斯

Article

Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth

Marcel Adam Just, Lisa Pan, Vladimir L. Cherkassky, Dana L. McMakin, Christine Cha, Matthew K. Nock & David Brent

The clinical assessment of suicidal risk would be substantially complemented by a biologically based measure that assesses alterations in the neural representations of concepts related to death and life in people who engage in suicidal ideation. This study used machine-learning algorithms (Gaussian Naive Bayes) to identify such individuals (17 suicidal ideators versus 17 controls) with high (91%) accuracy, based on their altered functional magnetic resonance imaging neural signatures of death-related and life-related concepts. The most discriminating concepts were 'death', 'cruelty', 'trouble', 'carefree', 'good' and 'praise'. A similar classification accurately (94%) discriminated nine suicidal ideators who had made a suicide attempt from eight who had not. Moreover, a major facet of the concept alterations was the evoked emotion, whose neural signature served as an alternative basis for



Marcel Adam Just, et al. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. Nature Human Behaviour (2017). doi:10.1038/s41562-017-0234-y

R语言实现

```
library(e1071)
imodel <- naiveBayes(wlfk~.,
                     data = cjb[train_set_idx, ])
predicted_train <- predict(imodel,
                           newdata = cjb[train_set_idx,],
                           type = "class")
Metrics::ce(cjb$wlfk[train_set_idx], predicted_train)
#> [1] 0.2920518
```

R语言实现

```
predicted_test <- predict(imodel,
                          newdata = cjb[-train_set_idx,],
                          type = "class")
Metrics::ce(cjb$wlfk[-train_set_idx], predicted_test)
#> [1] 0.27897
```

R语言实现

```
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 <- naiveBayes(wlfk~.,
                             data = train_set)
  predicted_train <-  predict(imodel_kfold,
                              train_set, type = "class")
```

R语言实现

```
  imetrics("naiveBayes", "Train",
           predicted_train, train_set$wlfk)
  predicted_test <- predict(imodel_kfold,
                            test_set, type = "class")
  imetrics("naiveBayes", "Test",
           predicted_test, test_set$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 naiveBayes Train 0.7040230  0.2959770
#> 42 naiveBayes  Test 0.6794872  0.3205128
#> 43 naiveBayes Train 0.7155172  0.2844828
#> 44 naiveBayes  Test 0.7179487  0.2820513
#> 45 naiveBayes Train 0.7068966  0.2931034
#> 46 naiveBayes  Test 0.7307692  0.2692308
#> 47 naiveBayes Train 0.7183908  0.2816092
#> 48 naiveBayes  Test 0.7051282  0.2948718
.....
#> 59 naiveBayes Train 0.6944046  0.3055954
#> 60 naiveBayes  Test 0.7142857  0.2857143
```

究竟是一种什么关系

既然所有规律都是关系

那么，请问：

朴素贝叶斯

究竟是什么关系

得到规律的表现形式是什么

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