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

分类与回归：根据目前所拥有的信息（数据）来建立人们所关心的变量和其他有关变量的关系

假如用y表示感兴趣的变量，用x表示其他可能与Y有关的变量（也可能是由若干变量组成的向量），分类与回归就是建立以下函数关系

$$y = f(x)$$

y被称为因变量或响应变量（类别/标签）

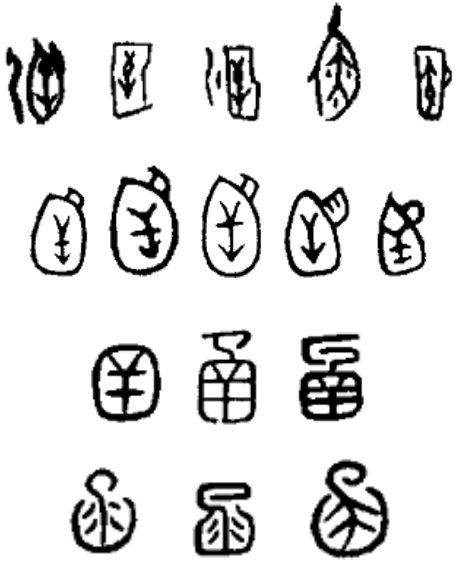
x被称为自变量（属性/特征），也称为解释变量或协变量

温故而知新：什么是函数

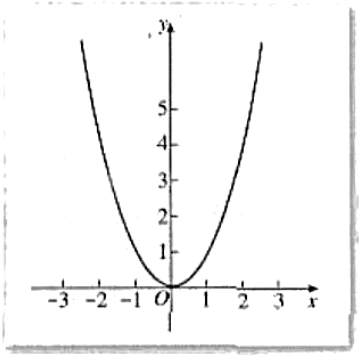
在同一个自然现象或技术过程中，往往同时有几个变量在变化着。这几个变量并不是孤立地在变，而是相互联系并遵循着**一定的变化规律**。……抽去上面几个例子中所考虑的量的实际意义，它们都表达了两个变量之间的**相依关系**，这种相依关系给出了**一种对应法则**，根据这一对应法则，当其中一个变量在变化范围内任意取定一个数值时，另一个变量就有确定的值与之对应。两个变量之间的这种对应关系就是函数概念的实质。

定义 设x和y是两个变量，D是一个给定的数集。如果对于每个数$x \in D$，变量y按照**一定的法则**总有确定的数值和它对应，则称y是x的函数，记作$y = f(x)$

温故而知新：什么是函数



温故而知新：什么是函数



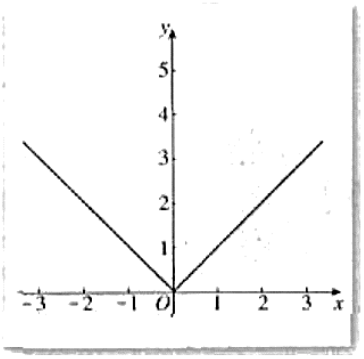


图 1.3.7

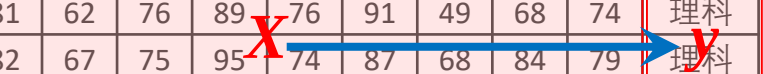
x	-3	-2	-1	0	1	2	3
$f(x)=x^2$	9	4	1	0	1	4	9

x	-3	-2	-1	0	1	2	3
$f(x)=\lvert x\rvert$	3	2	1	0	1	2	3

温故而知新：什么是函数

x	-3	-2	-1	0	1	2	3
$f(x)=x^2$	9	4	1	0	1	4	9

xm	xb	yw	sx	wy	zz	ls	dl	wl	hx	sw	wlfk
祝友香	女	88	88	95	96	84	95	98	98	88	文科
班维	男	87	94	86	84	85	94	81	88	90	理科
崔辉	男	87	72	88	92	87	86	49	84	80	文科
贲惊姣	女	83	83	75	86	85	94	43	88	78	理科
昌肖峰	男	81	62	76	89	76	91	49	68	74	理科
储承香	男	82	67	75	95	74	87	68	84	79	理科
房果平	女	92	93	90	94	94	94	99	98	97	理科
苍旺金	男	86	75	81	89	91	90	87	84	96	理科
锺志浩	男	88	95	87	93	96	92	77	92	90	理科
柯婷	女	87	82	92	91	95	100	75	86	85	理科
浦丹华	女	88	79	80	95	93	96	58	94	77	文科



分类与回归的任务：**通过学习获得一个目标函数f，将每个属性集x映射到y**

分类与回归

建立这种映射关系的过程就叫做回归或者分类

当因变量为数量变量时，叫做回归，而当因变量为类别变量（也称名义变量或分类变量）时叫做分类

分类：构造一个分类器classifier来预测类标号

回归：构造一个预测器predictor来预测一个数值

根据模型可以通过自变量对因变量进行预测，这种预测只是估计，只是一种函数关系，并非**决定性**因素或**因果**关系

分类的两个步骤：训练

训练数据

性别	语文	……	文理分科
女	94	…	文科
男	87	…	理科
男	92	…	理科
女	91	…	文科
…	…	…	…

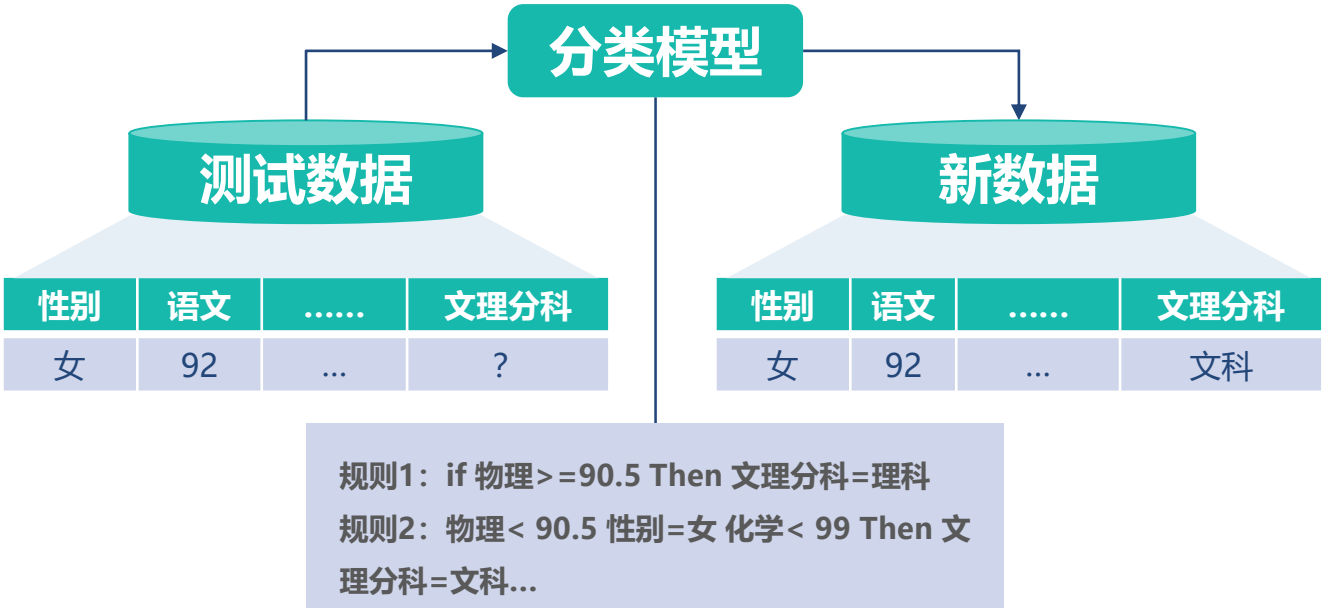
分类算法

分类模型

规则1：if 物理>=90.5 Then 文理分科=理科

规则2：物理< 90.5 性别=女 化学< 99 Then 文理分科=文科…

分类的两个步骤：测试


分类模型
测试数据
新数据
性别 语文 …… 文理分科
女 92 … ?
性别 语文 …… 文理分科
女 92 … 文科
规则1：if 物理>=90.5 Then 文理分科=理科
规则2：物理< 90.5 性别=女 化学< 99 Then 文理分科=文科…

模型评估的方法

模型的评估和模型的建立同等重要！

为了评估、选择、调试不同的分类器，我们需要通过实际的数据来测试不同分类器的泛化误差；采用测试误差作为泛化误差的近似

对于一个包含m个样例的数据集D

$$D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$$

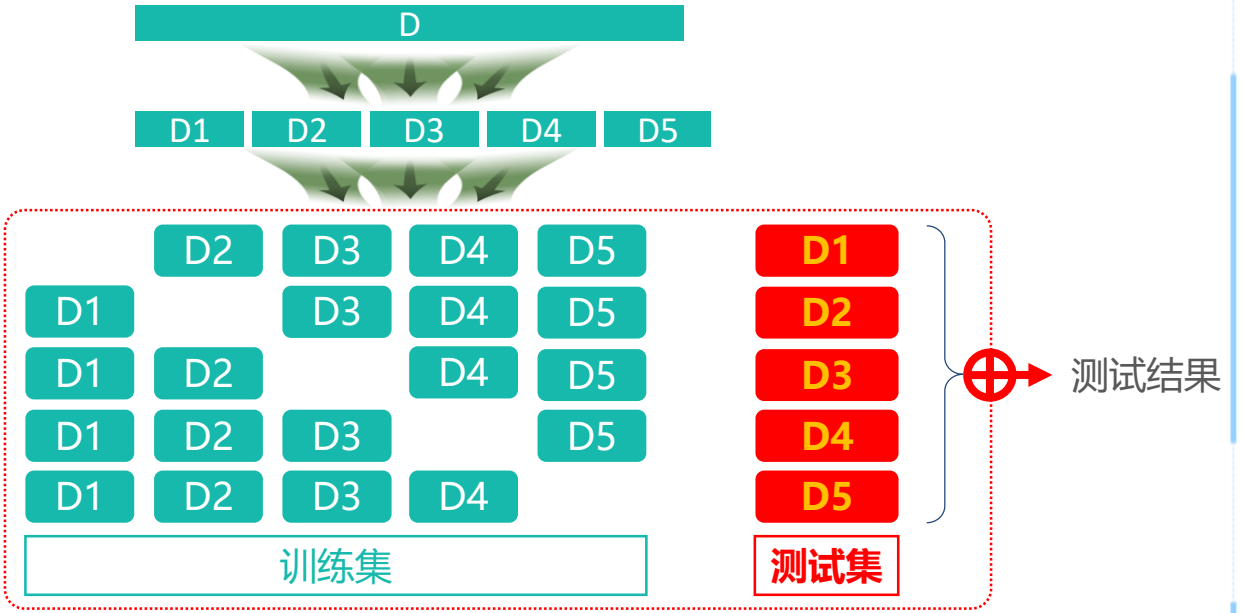
在数据集D的基础上，产生训练集S和测试集T

在同一轮训练、测试中，一般要求：$S \cap T = \Theta$

模型评估的方法

评估方法	基本原理
留出法 Hold-out	$D = S \cup T, S \cap T = \Phi$ 随机抽样，而不是原始数据的先后顺序 分层抽样，保持比例
交叉验证法 cross validation	将数据分为大小大致相同的 k 份 每一次将其中一份作为测试集，剩余的 k-1份作为训练集 以 k 次测试结果的平均值作为最终的测试误差
自助法 out-of-bag	采用有放回抽样，产生训练集 有36%左右的样本不会被抽到，作为测试集 在随机森林等组合学习算法中使用较多

k折交叉检验示意图


D
D1
D2
D3
D4
D5
D2
D3
D4
D5
D1
D1
D3
D4
D5
D2
D1
D2
D4
D5
D3
D1
D2
D3
D5
D4
D1
D2
D3
D4
D5
训练集
测试集
测试结果

模型评估：k折交叉检验

```
cv_kfold <- function(data, k = 10, seed = 2012) {
  n_row <- nrow(data)#计算数据的行数
  n_foldmarkers <- rep(1:k, ceiling(n_row / k))[1:n_row]
  set.seed(seed)
  n_foldmarkers <- sample(n_foldmarkers)#打乱顺序
  kfold <- lapply(1:k, function(i){
    (1:n_row)[n_foldmarkers == i]
  })
  return(kfold)
}
```

模型评估：留出法

```
#留出法hold-out
set.seed(2012)
train_set_idx <- sample(nrow(cjb), nrow(cjb)*0.7)
str(train_set_idx)
#> int [1:541] 169 576 218 722 575 673 411 700 687 696 ...
length(train_set_idx) / nrow(cjb)
#> [1] 0.6989664
train_set <- cjb[train_set_idx, ]
# test_set <- ?
```

模型评估：k折交叉检验

```
cv_kfold(cjb)
#> [[1]]
#> [1]   7  14  15  25  35  48  56  59  60  61  65  91  92
#> [14] 102 109 114 128 130 135 141 156 169 178 180 181 185
#> [27] 189 190 191 196 208 217 244 245 247 263 280 282 291
#> [40] 293 301 309 319 324 327 328 329 330 332 356 361 362
#> [53] 376 384 412 413 446 456 485 489 499 500 519 525 531
#> [66] 534 550 559 578 585 586 598 607 619 620 675 685 719
sapply(cv_kfold(cjb), length)
#>  [1] 78 78 78 78 77 77 77 77 77 77
```

模型评估：评估指标

```
global_performance <- NULL
imetrics <- function(method, type, predicted, actual) {
  con_table <- table(predicted, actual)
  cur_one <- data.frame(
    method = method, #算法模型的名称
    type = type, #取值为train或是test
    accuracy = sum(diag(con_table)) / sum(con_table),
    error_rate = 1 - accuracy
  )
  assign("global_performance",
         rbind(get("global_performance", envir = .GlobalEnv),
               cur_one),
         envir = .GlobalEnv)
}
```

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