北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

Data Analytics with R
语言数据分析

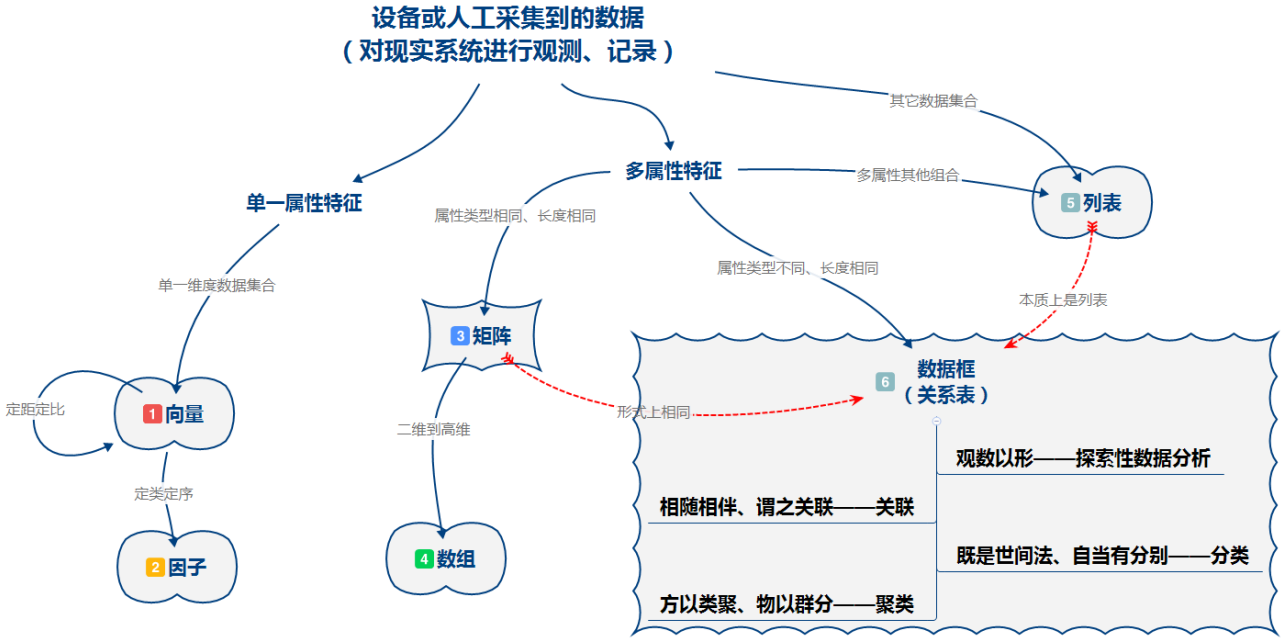
最美不过数据框

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

数据对象


设备或人工采集到的数据
（对现实系统进行观测、记录）
其它数据集合
单一属性特征
多属性特征
多属性其他组合
5 列表
属性类型相同、长度相同
单一维度数据集合
属性类型不同、长度相同
本质上是列表
3 矩阵
定距定比
1 向量
形式上相同
6 数据框
（关系表）
二维到高维
定类定序
观数以形——探索性数据分析
相随相伴、谓之关联——关联
既是世间法、自当有分别——分类
2 因子
4 数组
方以类聚、物以群分——聚类

发现数据框背后的规律

先看一个简单的逻辑：

数据框（关系表）是最常见的数据对象

发现数据背后的规律

很大程度上就是

发现**数据框**背后的规律

发现数据背后的规律

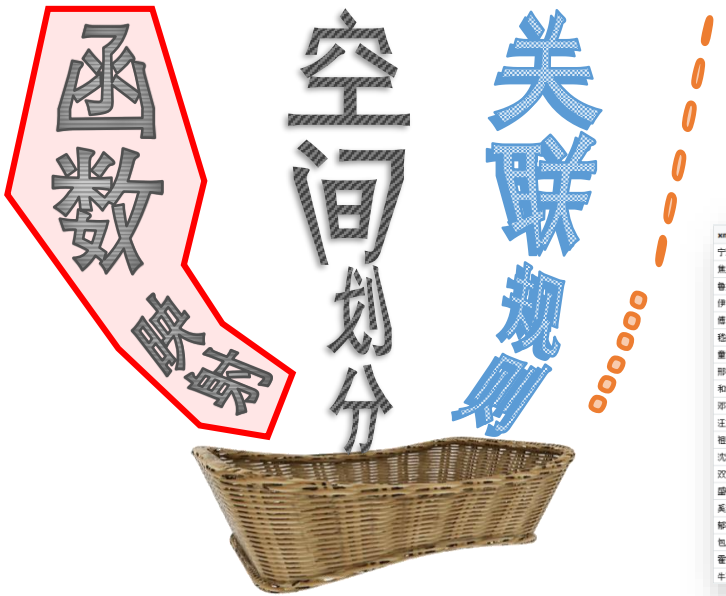
以机器学习为内核

数据分析 ≈ 机器学习 / 数据挖掘

≈ 认识数据 + **关联 + 分类 + 聚类**

≈ 寻找关系结构（核心是归归类）

数据框里有乾坤

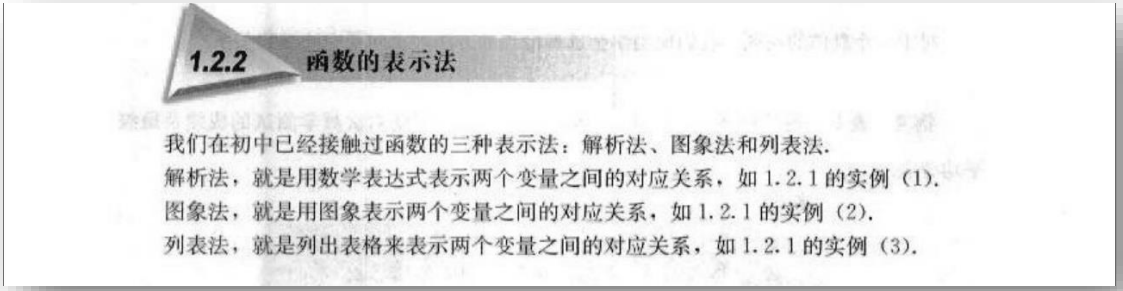


xm	bj	xb	yw	sx	wy	zz	ls	dl	wl	hx	sw	wlfk
宁境	1110	男	94	97	97	97	100	100	100	100	100	理科
焦金鼎	1108	女	91	98	95	99	99	100	100	100	97	理科
鲁孟秋	1109	女	93	98	98	97	96	98	100	100	98	理科
伊礼贤	1108	男	92	97	99	95	94	100	100	100	99	理科
傅世鸿	1110	男	88	98	99	95	100	100	97	100	95	理科
嵇婉玉	1107	女	94	100	96	98	100	98	88	100	97	文科
童秋艳	1109	女	92	98	95	95	95	100	100	98	98	理科
那任林	1106	女	93	98	92	99	100	98	97	100	93	文科
和五筠	1108	女	96	98	94	97	94	98	98	98	97	理科
邓丽凤	1106	女	94	95	91	97	98	100	95	100	99	文科
汪文昊	1108	男	93	97	93	97	94	98	100	98	99	理科
祖雅秋	1109	男	91	93	92	95	100	100	100	100	98	理科
沈秋艳	1109	女	94	92	94	96	96	100	100	100	97	理科
况文华	1110	男	92	97	93	98	95	96	100	100	98	理科
盛叶倩	1110	女	88	96	97	96	97	100	100	100	95	理科
高梦欢	1106	女	93	97	98	96	97	100	91	100	96	文科
郸巧月	1108	女	93	99	92	95	100	97	95	100	97	理科
包威	1110	男	93	99	98	99	93	98	95	100	92	理科
霍学超	1108	男	94	96	97	97	95	96	95	100	96	理科
牛飞雨	1109	男	90	98	94	97	95	100	94	100	98	理科

温故而知新：什么是函数

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$y = f(x)$

令人费解：什么叫一定的法则?



1.2.2　函数的表示法

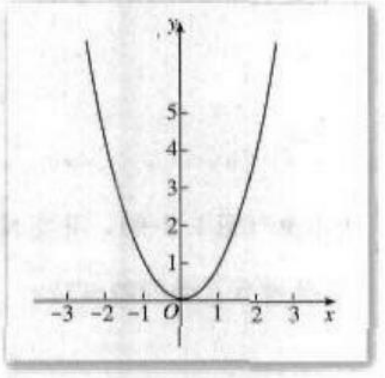
我们在初中已经接触过函数的三种表示法：解析法、图象法和列表法.

解析法，就是用数学表达式表示两个变量之间的对应关系，如 1.2.1 的实例 (1).

图象法，就是用图象表示两个变量之间的对应关系，如 1.2.1 的实例 (2).

列表法，就是列出表格来表示两个变量之间的对应关系，如 1.2.1 的实例 (3).

温故而知新：什么是函数



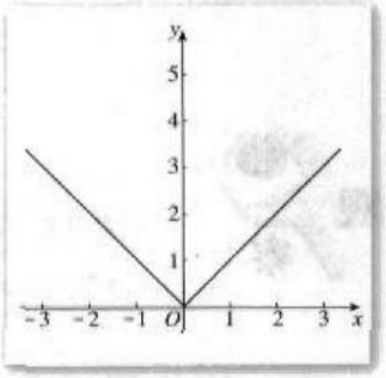


图 1.3-7

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

数据框里的函数

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

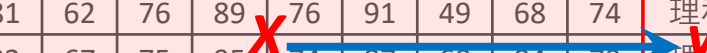
xm	xb	yw	sx	wy	zz	ls	dl	wl	hx	sw	wlfk
祝友香	女	88	88	95	96	84	95	98	98	88	文科
班维	男	87	94	86	84	85	94	81	88	90	理科
崔辉	男	87	72	88	92	87	86	49	84	80	文科
贲惊姣	女	83	83	75	86	85	94	43	88	78	理科
昌肖峰	男	81	62	76	89	76	91	49	68	74	理科
储承香	男	82	67	75	95	74	87	68	84	79	理科
房果平	女	92	93	90	94	94	94	99	98	97	理科
苍旺金	男	86	75	81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

数据框：函数的三种表现形式之一——列表法

数据框里的函数

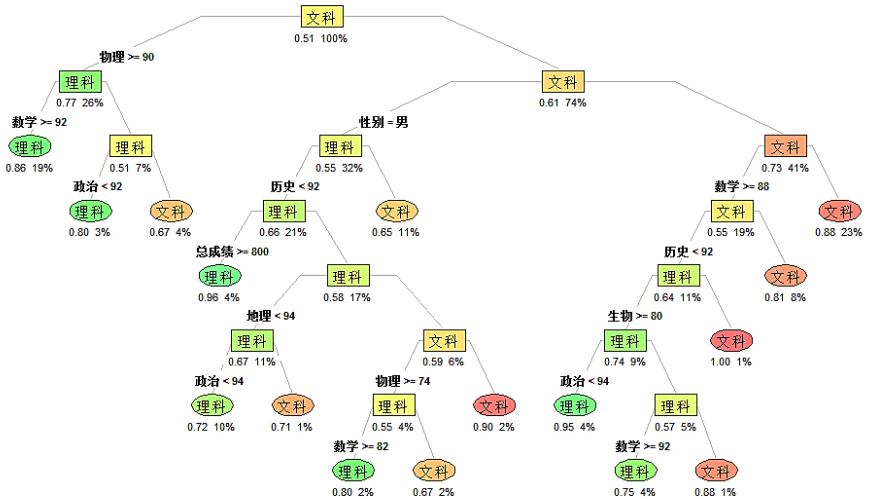
x	$f(x)=x^2$
−3	9
−2	4
−1	1
0	0
1	1
2	4
3	9

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



一定的法则在数据框中的体现：X和y出现在同一行之中

数据框里的函数



$$wlfk = f\left(xb, yw, sx, \cdots, sw\right)$$

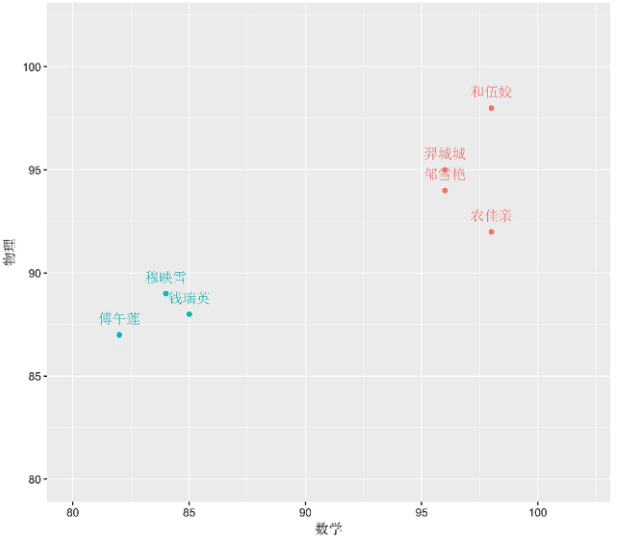
数据框里有乾坤

xm	bj	xb	yw	sx	wy	zz	ls	dl	wl	hx	sw	wlfk
宁境	1110	男	94	97	97	97	100	100	100	100	100	理科
焦金鼎	1108	女	91	98	95	99	99	100	100	100	97	理科
鲁孟秋	1109	女	93	98	98	97	96	98	100	100	98	理科
伊礼贤	1108	男	92	97	99	95	94	100	100	100	99	理科
傅世鸿	1110	男	88	98	99	95	100	100	97	100	95	理科
稽振玉	1107	女	94	100	96	98	100	98	88	100	97	文科
童依艳	1109	女	92	98	95	95	95	100	100	98	98	理科
那任林	1106	女	93	98	92	99	100	98	97	100	93	文科
和五筠	1108	女	96	98	94	97	94	98	98	98	97	理科
邓丽凤	1106	女	94	95	91	97	98	100	95	100	99	文科
汪文昊	1108	男	93	97	93	97	94	98	100	98	99	理科
祖雅秋	1109	男	91	93	92	95	100	100	100	100	98	理科
沈依艳	1109	女	94	92	94	96	96	100	100	100	97	理科
双文华	1110	男	92	97	93	98	95	96	100	100	98	理科
盛叶倩	1110	女	88	96	97	96	97	100	100	100	95	理科
高梦欢	1106	女	93	97	98	96	97	100	91	100	96	文科
郸巧月	1108	女	93	99	92	95	100	97	95	100	97	理科
包威	1110	男	93	99	98	99	93	98	95	100	92	理科
霍学超	1108	男	94	96	97	97	95	96	95	100	96	理科
牛飞雨	1109	男	90	98	94	97	95	100	94	100	98	理科

数据框与数据空间

数据集包含n个属性/特征，张成一个n维数据空间

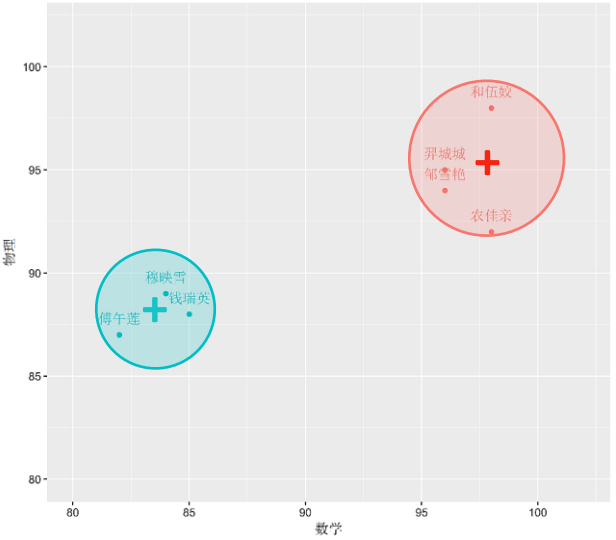
xm	sx	wl
傅午莲	82	87
钱瑞英	85	88
穆映雪	84	89
和伍姣	98	98
邹雪艳	96	94
农佳亲	98	92
羿城城	96	95



数据框与数据空间

距离关系的远近，在数据空间中形成了自然的结构：簇

xm	sx	wl
傅午莲	82	87
钱瑞英	85	88
穆映雪	84	89
和伍姣	98	98
邹雪艳	96	94
农佳亲	98	92
羿城城	96	95



数据框里有乾坤

xm	bj	xb	yw	sx	wy	zz	ls	dl	wl	hx	sw	wlfk
宁境	1110	男	94	97	97	97	100	100	100	100	100	理科
焦金鼎	1108	女	91	98	95	99	99	100	100	100	97	理科
鲁孟秋	1109	女	93	98	98	97	96	98	100	100	98	理科
伊礼贤	1108	男	92	97	99	95	94	100	100	100	99	理科
傅世鸿	1110	男	88	98	99	95	100	100	97	100	95	理科
稂振玉	1107	女	94	100	96	98	100	98	88	100	97	文科
童依驰	1109	女	92	98	95	95	95	100	100	98	98	理科
那任林	1106	女	93	98	92	99	100	98	97	100	93	文科
和五俊	1108	女	96	98	94	97	94	98	98	98	97	理科
邓丽凤	1106	女	94	95	91	97	98	100	95	100	99	文科
汪文昊	1108	男	93	97	93	97	94	98	100	98	99	理科
祖雅秋	1109	男	91	93	92	95	100	100	100	100	98	理科
沈依驰	1109	女	94	92	94	96	96	100	100	100	97	理科
双文华	1110	男	92	97	93	98	95	96	100	100	98	理科
盛叶倩	1110	女	88	96	97	96	97	100	100	100	95	理科
高梦欢	1106	女	93	97	98	96	97	100	91	100	96	文科
郸巧月	1108	女	93	99	92	95	100	97	95	100	97	理科
包威	1110	男	93	99	98	99	93	98	95	100	92	理科
霍学超	1108	男	94	96	97	97	95	96	95	100	96	理科
牛飞雨	1109	男	90	98	94	97	95	100	94	100	98	理科

数据框里的关联规则

关联规则A⇒B ： A发生时伴随着B发生

啤酒和尿不湿的销量次第增长

数据框里的关联规则

关联规则A⇒B ： A发生时伴随着B发生

支持度：support(A⇒B) = P(A∪B)=53/774=0.068

置信度：confidence(A⇒B) = P(B|A) = P(A∪B)/P(A)=53/63=0.841

xb	yw	sx	wy	zz	ls	dl	wl	hx	sw	wlfk
女	优	优			优	优	优	优	优	理科
男	优	优	中	优	优	优	良	优	优	文科
男	良	优	良			优	中	优	优	理科
男	良	不及格	良	良	良	良	及格	及格	中	理科
男	良	优	优	良	良	优	良	良	中	理科
男	及格	及格	及格	良	良	优	不及格	优		科
男	良	良	中	优	优	优	及格	良	中	文科

wlfx=文科：53

ls=优, hx=良：63

全部记录数：774

{ls=优, hx=良} ⇒ {wlfk=文科}

数据框里的关联规则

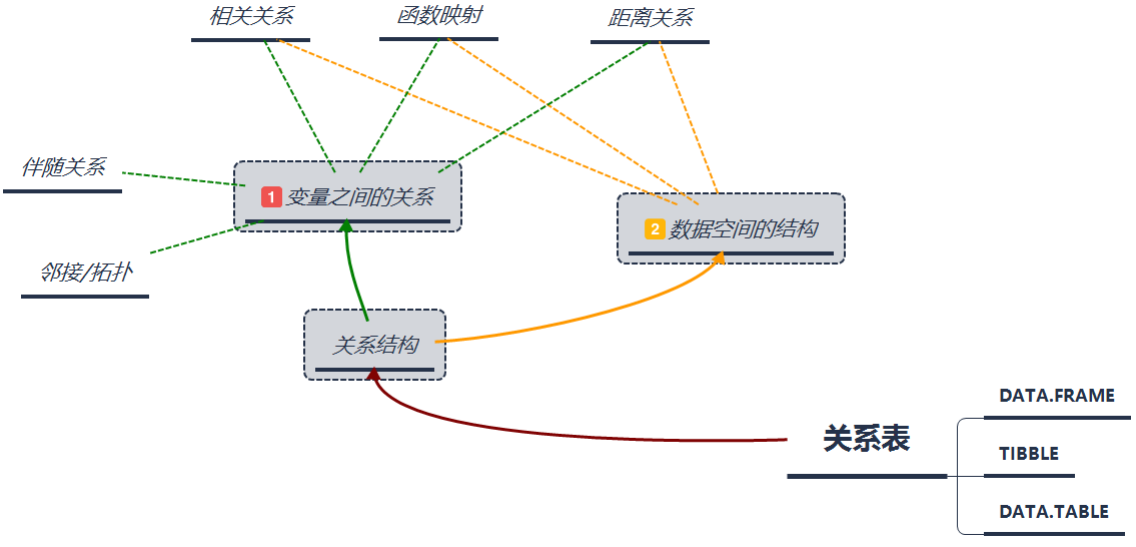
关联规则A⇒B： A发生时伴随着B发生

支持度：support(A⇒B) = P(A∪B)

置信度：confidence(A⇒B) = P(B|A) = P(A∪B)/P(A)

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[60]	{yw=优,ls=优,sw=良}	{wlfk=文科}	0.068	0.869	1.707	53.000
[63]	{xb=男,ls=优,wl=优}	{wlfk=理科}	0.099	0.837	1.705	77.000
[73]	{xb=男,yw=良,ls=良,hx=优}	{wlfk=理科}	0.075	0.829	1.688	58.000
[80]	{xb=女,ls=优,sw=良}	{wlfk=文科}	0.083	0.853	1.676	64.000
[81]	{sx=优,ls=良,dl=优,hx=优}	{wlfk=理科}	0.066	0.823	1.675	51.000
[83]	{sx=优,ls=良,hx=优}	{wlfk=理科}	0.081	0.818	1.667	63.000
[89]	{ls=优,hx=良}	{wlfk=文科}	0.068	0.841	1.653	53.000
[90]	{zz=优,ls=优,dl=优,sw=良}	{wlfk=文科}	0.116	0.841	1.652	90.000

最美不过数据框



一切都是关系结构：关系表几乎可以上升为一个数学概念

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