

第1章 絮论

李高荣

北京师范大学统计学院

E-mail: ligaorong@bnu.edu.cn



本章纲要

1 教材和相关资料

2 统计学习概述

- 统计学习的特点
- 统计学习的对象
- 统计学习的分类

微信公众号: BNULgr



- 扫二维码获取在线课件和相关教学电子资源
- 请遵守电子资源使用协议

- 2019年,中国共产党第十九届中央委员会第四次全体会议通过《中共中央关于坚持和完善中国特色社会主义制度推进国家治理体系和治理能力现代化若干重大问题的决定》,首次提出数据可作为**生产要素**按贡献参与分配.
- 目前,数据与人工智能、机器学习和大数据的先进分析技术等结合在一起,一个全新的“**人工智能大数据**”时代正在来临,特别是数字化、网络化和智能化已经成为新一轮科技革命的重要技术代表.
- 2022年11月,OpenAI推出的**ChatGPT**更是产生了深远的影响.

思考

然而,这些技术背后的方法、模型、理论和算法究竟是什么?

1 教材和相关资料

2 统计学习概述

- 统计学习的特点
- 统计学习的对象
- 统计学习的分类

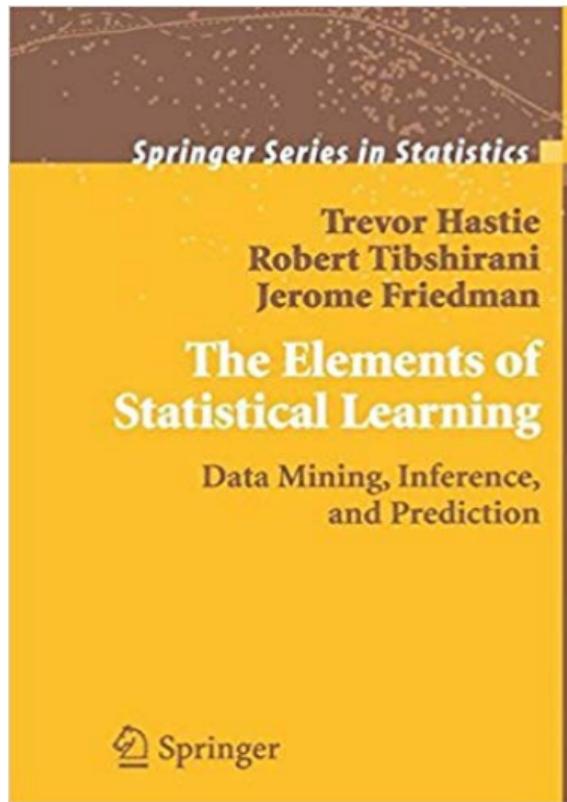
本课程教材—统计学习(R语言版)

■ 李高荣(2024). 统计学习(R语言版). 北京: 高等教育出版社.

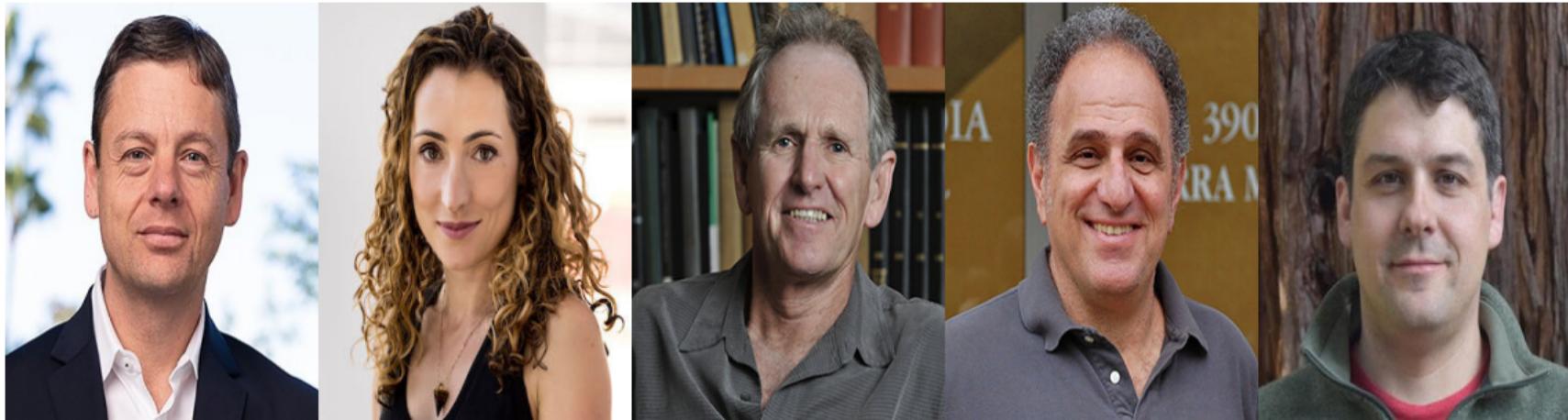


- Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd edition). New York: Springer.

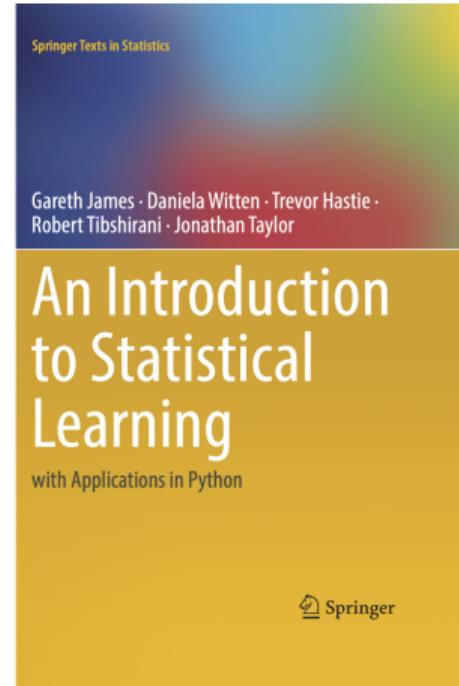
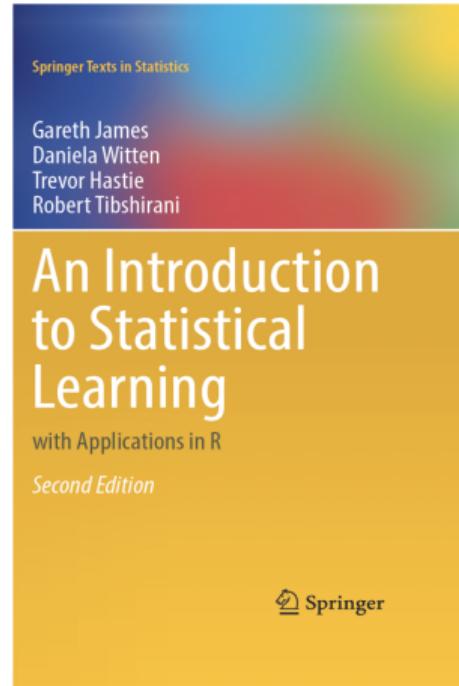
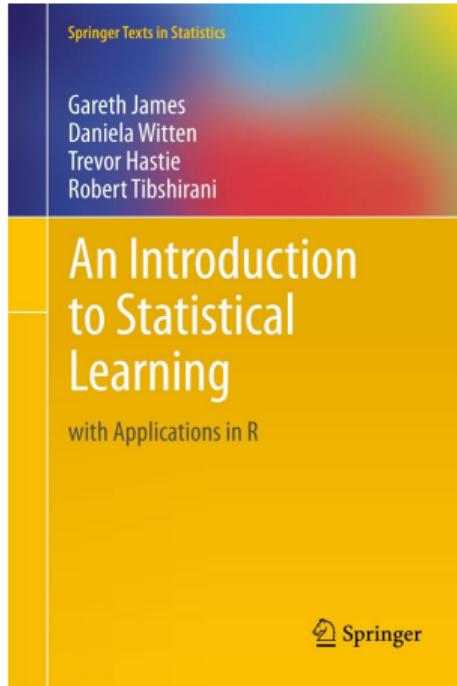




- James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R (2nd Edition)*. New York: Springer.
- James, G., Witten, D., Hastie, T., Tibshirani, R. and Taylor, J. (2023). *An Introduction to Statistical Learning with Applications in Python*. New York: Springer.



统计学习经典教材



1

教材和相关资料

2

统计学习概述

- 统计学习的特点
- 统计学习的对象
- 统计学习的分类

■ 统计学习(statistical learning, SL), 使用统计方法的一种机器学习(machine learning, ML), 可视作基于数据构建概率统计模型并运用模型对数据进行预测与分析的一门学科.

■ Jonathan Rosenberg:

- “ Data is the sword of the 21st century, those who wield it well, the Samurai.”
- “数据是21世纪的宝剑, 使用得当的人可以变成武士.”

1

教材和相关资料

2

统计学习概述

- 统计学习的特点
- 统计学习的对象
- 统计学习的分类

■ **统计学习**: 以理解数据为目的的庞大工具集. 李航(2019)给出了统计学习的特点:

- ① 统计学习以计算机及网络为平台, 是建立在计算机及网络上的;
- ② 统计学习以数据为研究对象, 是数据驱动的学科;
- ③ 统计学习的目的是对数据进行预测与分析;
- ④ 统计学习以方法为中心, 统计学习方法构建模型并应用模型进行预测与分析;
- ⑤ 统计学习是统计学、概率论、数学、信息论、计算理论、最优化理论及计算机科学等多个领域的交叉学科, 并且在发展中逐步形成独自的理论体系与方法论.

1

教材和相关资料

2

统计学习概述

- 统计学习的特点
- 统计学习的对象
- 统计学习的分类

统计学习的对象—数据

■ 统计学习研究的对象是**数据(data)**, 大数据时代, 数据特征:

大数据 = 高频海量数据+ 复杂类型的数据

- Variety(多样性)
- Volume(数据量大)
- Veracity(真实性)
- Velocity(速度快)
- Value(价值密度低)



■ 数据特征:

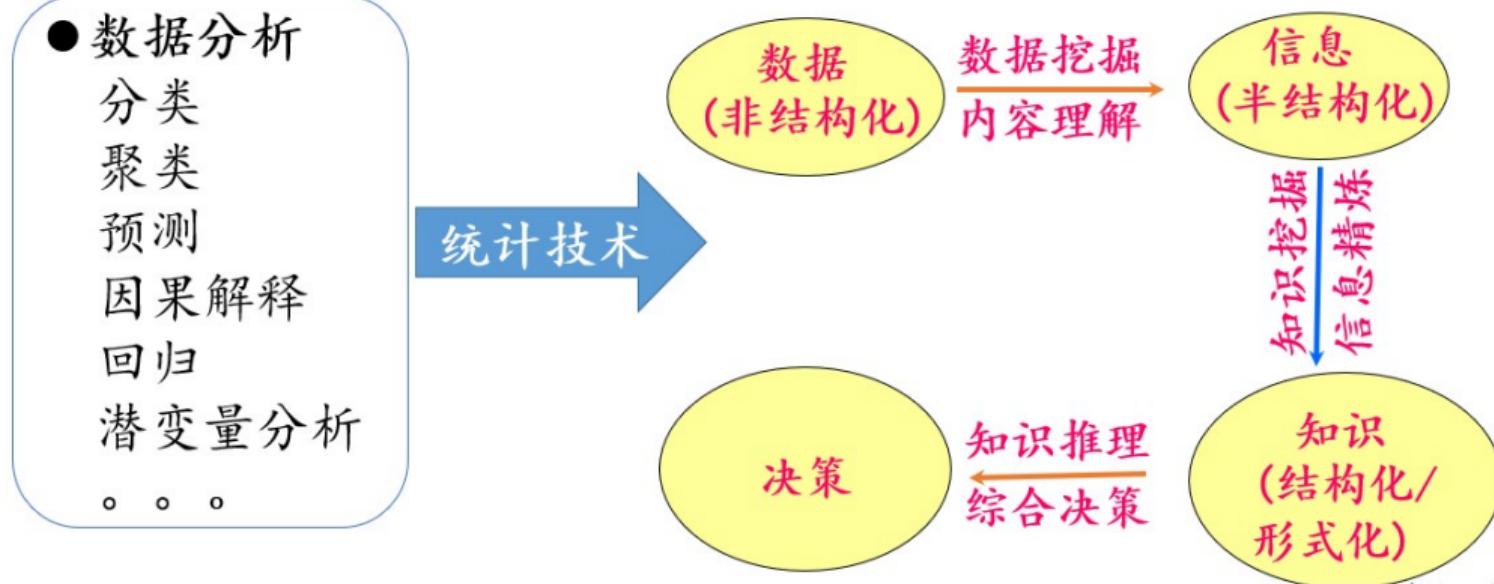
- 量: 海量, TB或PB级(volume)
- 质: 机制复杂
- 类型: 缺失、删失、纵向、相依、高维、稀疏、非线性

- 这些数据特征极大地增加了数据分析的难度;
- 从数据中学习是数据时代至关重要的挑战之一, 它给统计学家提供了难得的机遇;
- 统计学习和机器学习越来越朝着智能数据分析的方向发展, 并已成为智能数据分析技术的一个重要源泉.

统计学习的对象—数据

- 基因组数据、芯片数据
- 生物医学数据
- 金融交易数据
- 卫星遥感数据
- 医学成像数据
- 传感器数据、视频、监控数据
- 互联网文本数据
-

统计学习的对象—数据



例1.1: 前列腺癌数据

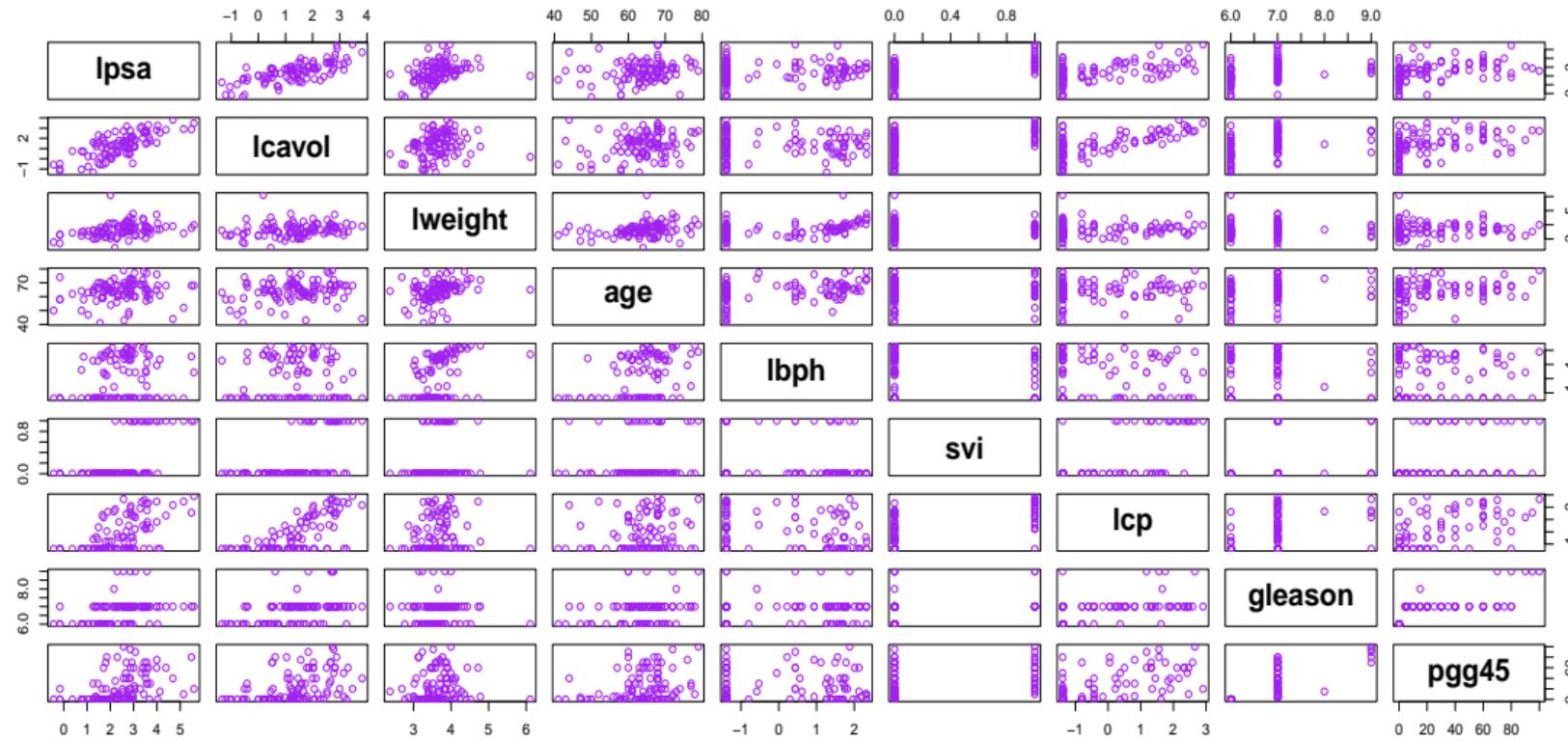
例1.1: 前列腺癌数据

程序包[faraway](#)中的前列腺癌数据集prostate包含97个样本和9个变量: `lpsa` (PSA的对数)、`lcavol` (癌体积的对数)、`lweight` (前列腺重量的对数)、`age` (患者年龄)、`lbph` (良性前列腺增生量的对数)、`svi` (精囊浸润)、`lcp` (包膜穿透的对数)、`gleason` (格里森分数)和`pgg45` (格里森分为4或5的比例).

研究目的

通过8个临床指标预测`lpsa` (PSA的对数).

例1.1: 前列腺癌症数据



例1.2: Credit数据

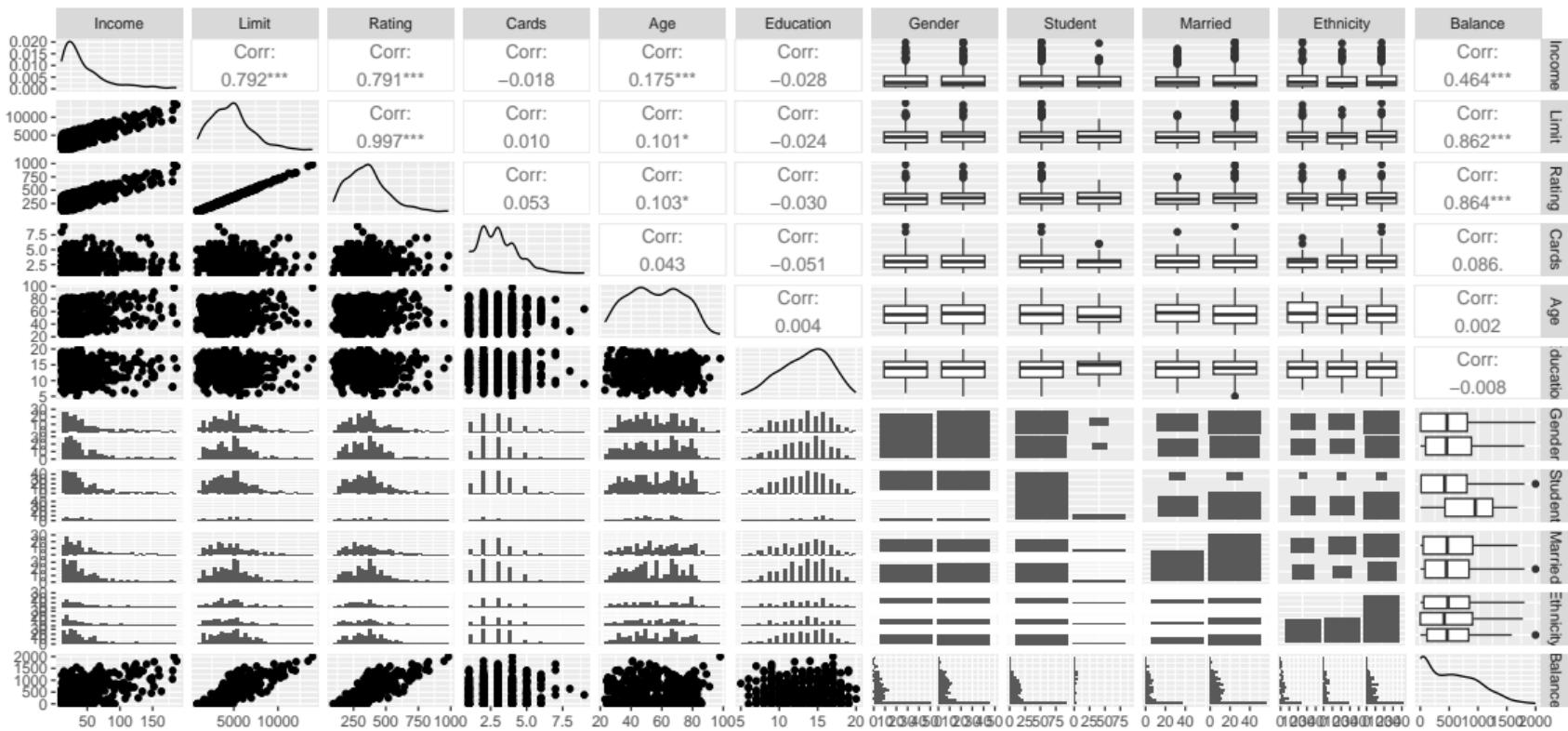
例1.2: Credit数据

程序包ISLR中的Credit数据集包含400个样本和11个变量: ① Balance, 表示个体客户的平均信用卡债务(单位: 美元); ② Income, 表示客户的收入(单位: 千美元); ③ Limit, 表示信用额度; ④ Rating, 表示信用评级; ⑤ Cards, 表示信用卡数量; ⑥ Age, 表示客户的年龄; ⑦ Education, 表示客户的受教育年限; ⑧ Gender, 表示客户的性别(取Male或Female); ⑨ Student(取Yes或No); ⑩ Married, 表示客户是否结婚; ⑪ Ethnicity, 表示客户的种族, 取“Caucasian”表示白种人, 取“American”表示非裔美国人, 取“Asian”表示亚洲人.

研究目的

通过其他10个变量预测个体客户的平均信用卡债务Balance.

例1.2: Credit数据



例1.3: 信用卡违约数据

信用卡违约数据

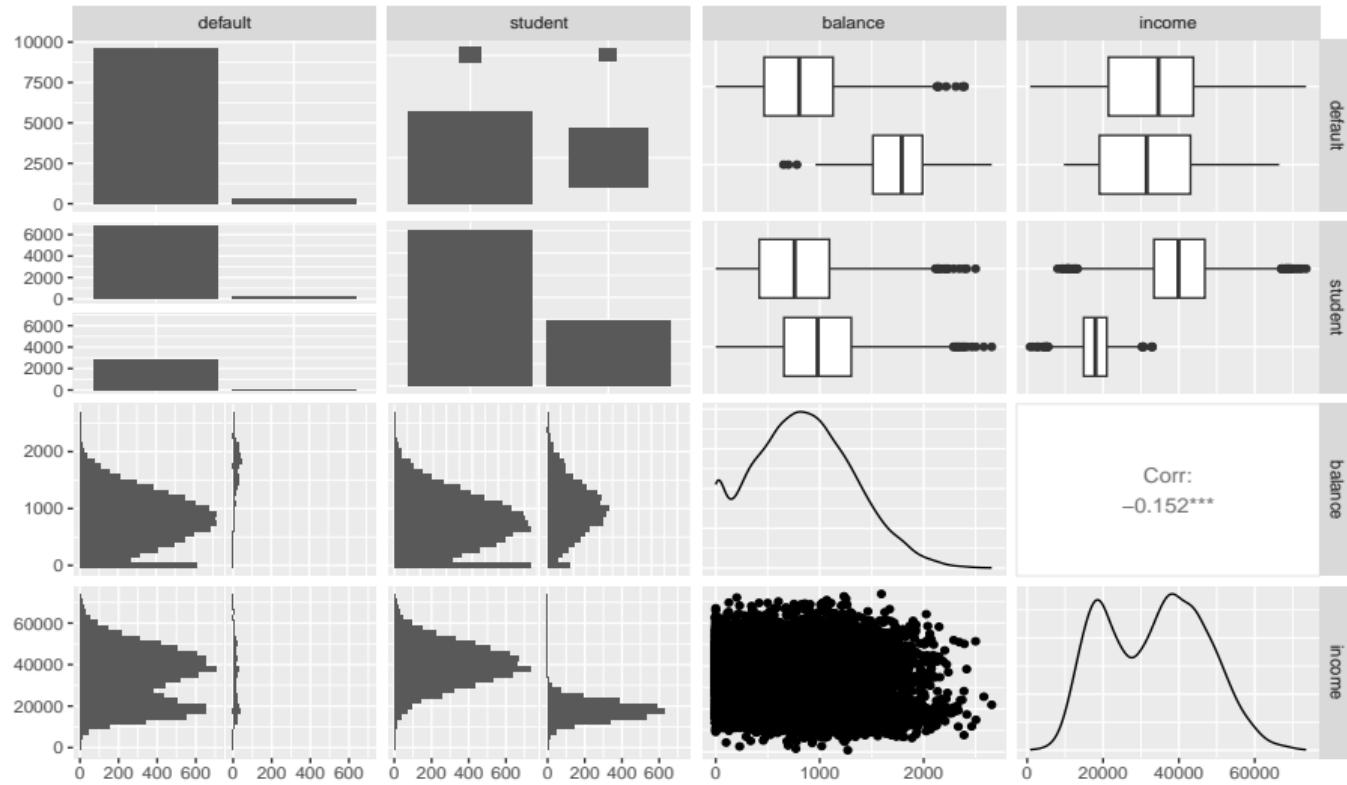
程序包**ISLR2**中的信用卡违约数据Default包含10000个样本和4个变量：

- ① **default**, 表示客户是否违约, 为定性变量或因子变量, 如果为"Yes"表示客户违约, 如果为"No"表示客户不违约;
- ② **student**, 表示客户是否为学生, 如果取"Yes"表示客户为学生, 如果取"No"表示客户不是学生;
- ③ **balance**, 表示客户每月信用卡的平均余额;
- ④ **income**, 表示客户的年收入.

研究目的

通过**student**、**balance**和**income**三个变量预测客户是否具有违约行为.

例1.3: 信用卡违约数据



例1.4: CIFAR-10数据集

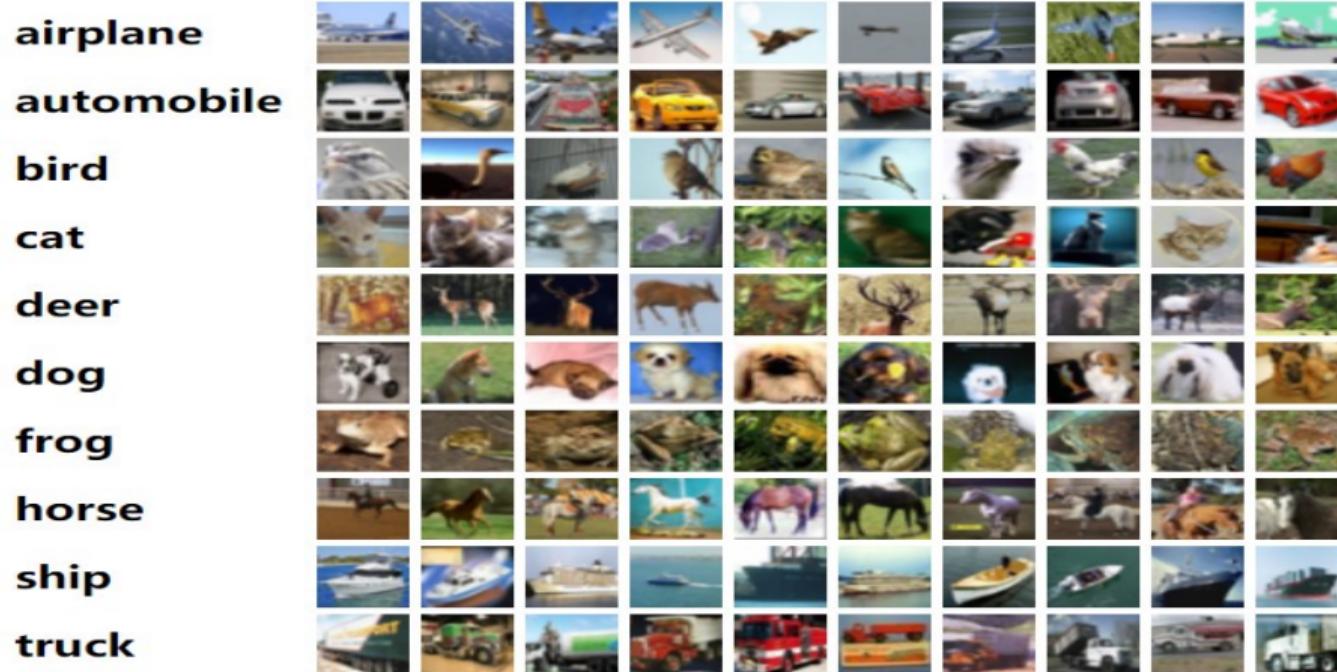
例1.4: CIFAR-10数据集

CIFAR-10是一个用于识别普适物体的数据集，由10个类别60000张 32×32 的RGB彩色图片组成，每类6000张图片，其中50000张为训练图片，10000张为测试图片。10个类别分别是“airplane”、“automobile”、“bird”、“cat”、“deer”、“dog”、“frog”、“horse”、“ship”和“truck”。

研究目的

通过训练集建立统计模型，然后通过测试集验证分类效果，最后把所建立的统计模型用于普适物体的识别和分类。

例1.4: CIFAR-10数据集



例1.5: 手写数字识别

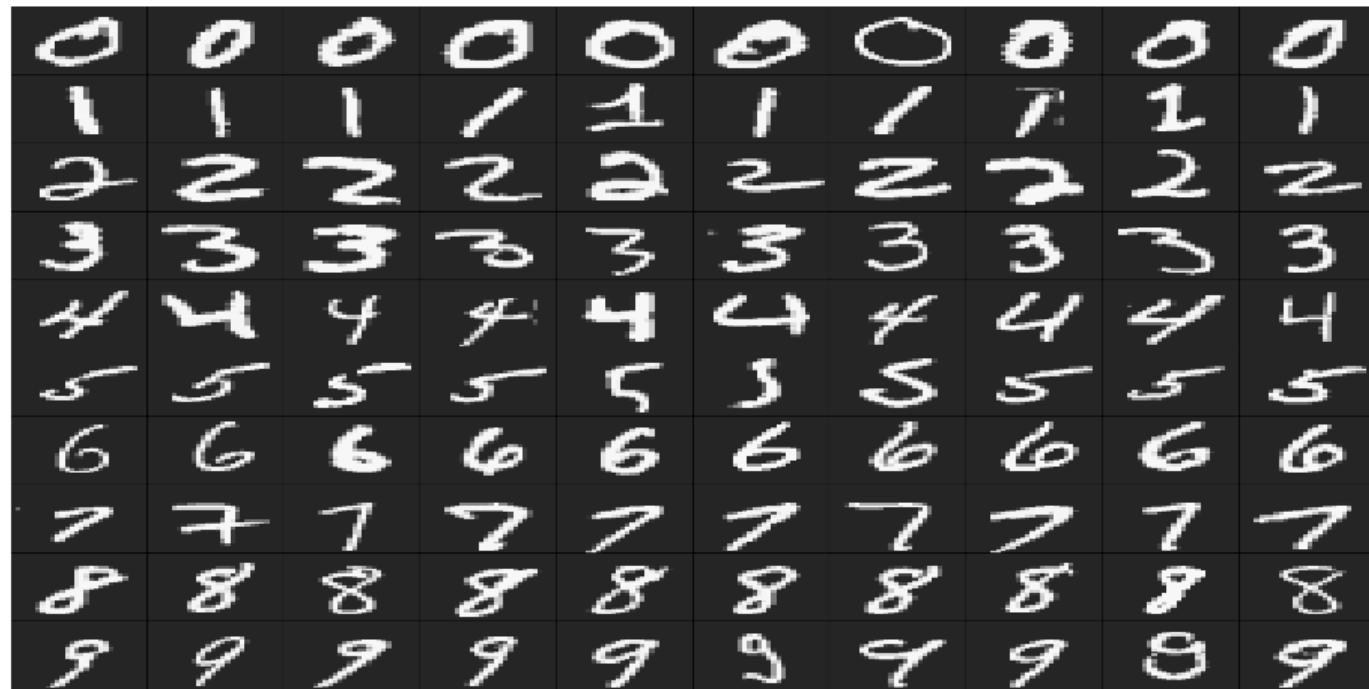
例1.5: 手写数字识别

MNIST是一个包含数字0 ~ 9的手写数字图片数据集，该数据集由60000个训练集和10000个测试集组成，其中每个图片由 28×28 个像素点组成，每个像素点的取值区间为 $[0, 255]$ ，0表示白色，255表示黑色。手写数字识别是一个多分类问题，共有10个类，每个手写数字图像的类别标签是0 ~ 9中的其中一个数字。

研究目的

从 28×28 的灰度值矩阵中快速又准确地判断每张图片上的数字，使得分类错误率尽量低。

例1.5: 手写数字识别



■ 根据变量取值的不同，可把变量分为两大类：定量变量和定性(或属性)变量。

① **定量变量**：通常就是指连续性变量，例如时间、长度、重量、产量、温度和速度等，它们是由测量或计数、统计所得到的具有数值特征的量，称为定量变量。

② **定性变量**：又称为分类或属性变量。

- **有序变量**，它没有数量关系，只有次序关系，如某种产品分为一等品、二等品、三等品等；矿石的质量分为贫矿和富矿。
- **名义变量**，这种变量既无等级关系，也无数量关系，如天气(阴、晴)，性别(男、女)，职业(工人、农民、教师、干部、医生等)等。

1

教材和相关资料

2

统计学习概述

- 统计学习的特点
- 统计学习的对象
- 统计学习的分类

■ 统计学习的分类:

① 有监督的学习(supervised learning), 有两种用途:

- (1) 建立面向预测的统计模型, 如回归模型或分类模型;
- (2) 对一个或多个给定的输入(input)估计某个输出(output).

② 无监督的学习(unsupervised learning): 对无标记的训练数据集 $D = \{x_1, \dots, x_n\}$ 进行建模, 寻找数据的模型和规律, 并作出推断结论, 其中 $x_i = (x_{i1}, \dots, x_{ip})^T$ 为 p 维的观测样本数据, 且 $i = 1, \dots, n$.

■ 无监督学习方法有聚类分析、主成分分析和因子分析等.

有监督统计学习的数据结构

输出变量		输入变量			
序号	Y	X_1	X_2	...	X_p
1	y_1	x_{11}	x_{12}	...	x_{1p}
2	y_2	x_{21}	x_{22}	...	x_{2p}
3	y_3	x_{31}	x_{32}	...	x_{3p}
:	:	:	:	:	:
n	y_n	x_{n1}	x_{n2}	...	x_{np}

■ \mathbf{X} 表示一个 $n \times p$ 矩阵, 表示为

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix};$$

- $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ 表示长度为 p 的列向量;
- $\mathbf{x}_j = (x_{1j}, x_{2j}, \dots, x_{nj})^T$ 表示长度为 n 的列向量;
- $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$.



谢谢，请多提宝贵意见！