

数据思维与实践

王伟

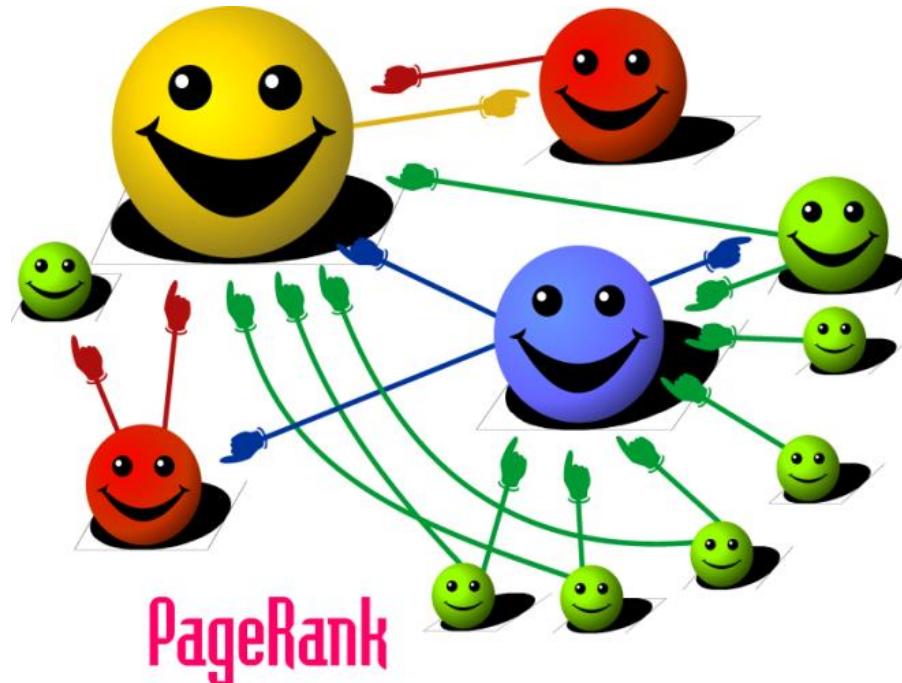
华东师范大学

数据科学与工程学院

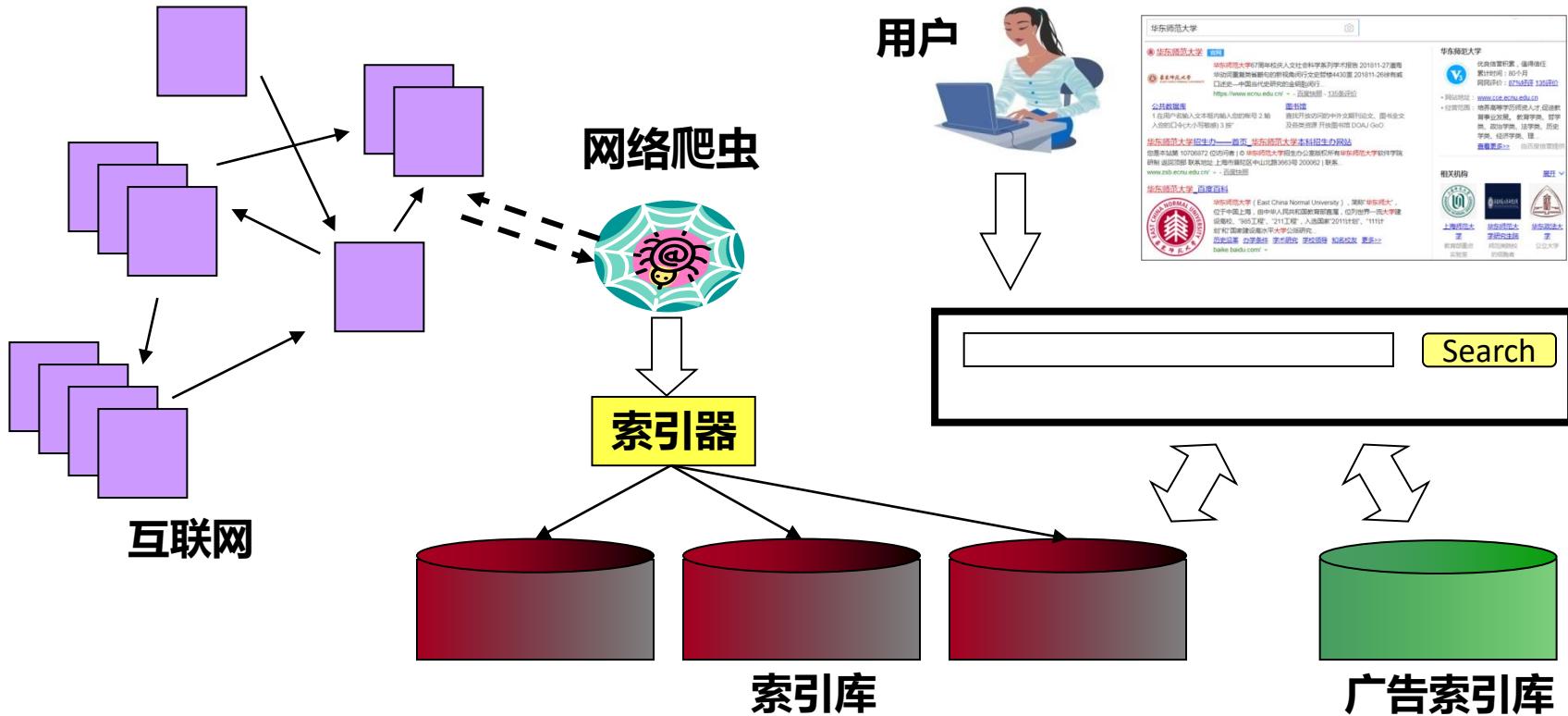
全民数字素养与技能培训基地



开篇实例：Google 的 PageRank



搜索引擎的原理



搜索示例：华东师范大学

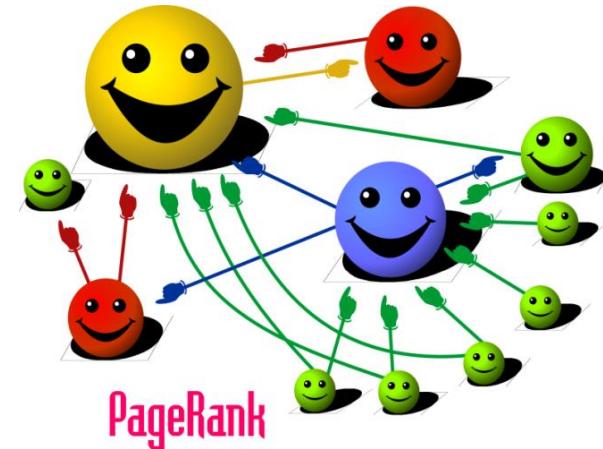
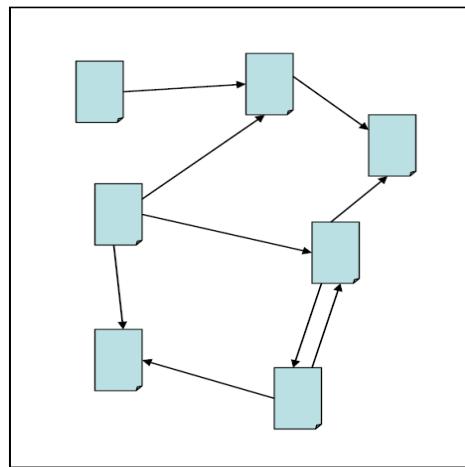
The screenshot shows a search interface with the query '华东师范大学' in the search bar. Below it, a yellow box highlights the title '网页排序结果' (Search Results). To the left of the title is a link to the official website of ECNU, which is also highlighted by a yellow arrow. To the right, there is a snippet of the search results showing a preview of a page from ECNU's website.

问题：搜索引擎怎么知道哪个网页排在前面，哪个排在后面呢？
即如何衡量网页的重要性？

This screenshot shows a search result for '华东师范大学' on Baidu. It includes the university's official website link, its entry in the Baidu Encyclopedia, and other related information. A large yellow arrow points from the text '其他相关信息' (Other Related Information) to the right side of the page, where links to affiliated institutions like Shanghai Normal University and East China Normal University Law School are listed.

大规模网页排名算法：PageRank

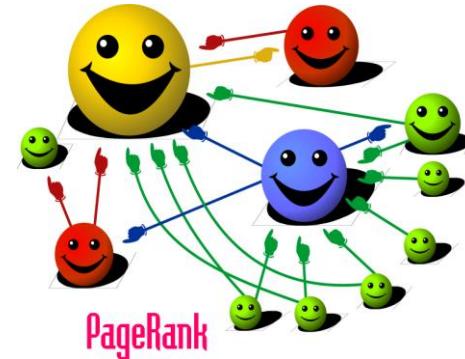
- 网页排名是网络搜索引擎的核心
- PageRank 是著名网络搜索引擎 Google 用于评测一个网页 “**重要性**” 或 “**影响力**” 的一种方法



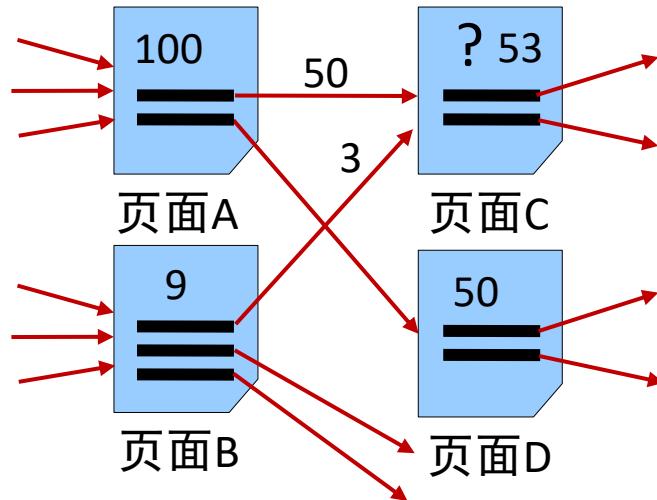
PageRank 的决定因素



- Google 的 PageRank 是基于这样一个理论：
 - 若 B 网页上有连接到 A 网页的链接，说明 B 认为 A 有链接价值，是一个“重要”的网页
 - 一个网页的重要性大致由下面两个因素决定：
 - 该网页的导入链接的数
 - 这些导入链接的重要性

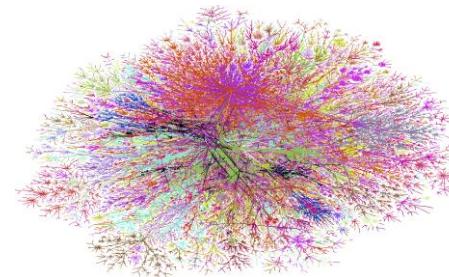


尝试计算 PageRank 值



可以用数据思维与模型
来解决这类问题

- 问题
 - 先有鸡还是先有蛋？
 - Internet的拓扑结构



有向图的知识

- ◆ 有向图
- ◆ 顶点的出度 (Out-degree)
- ◆ 顶点的入度 (In-degree)

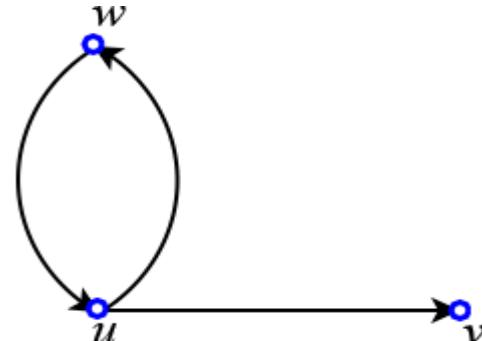
例：右图为一个有向图，记为 D

顶点组成的集合： $V(D) = \{u, v, w\}$

弧组成的集合： $A(D) = \{(u, w), (w, u), (u, v)\}$

顶点 u 的出度： $od(u) = 2$

顶点 u 的入度： $id(u) = 1$



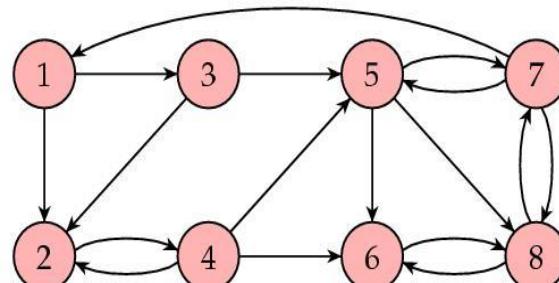
如何表示这个图，以便更好
计算 PageRank 值呢？

邻接矩阵

□ 为研究需要，我们定义邻接矩阵

$$G = (g_{ij}), \text{ 其中 } g_{ij} = \begin{cases} 1, & \text{如果存在从 } j \text{ 到 } i \text{ 的弧} \\ 0, & \text{otherwise} \end{cases}$$

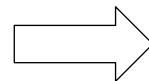
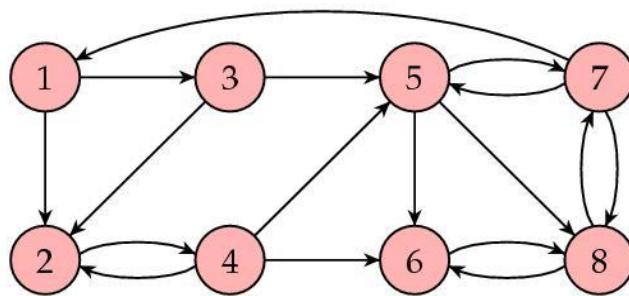
对于下例 中的有向图，其邻接矩阵为



$$G = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

超链接矩阵 (Hyperlink Matrix)

□ 进一步，如果将邻接矩阵中的元素除以对应节点的出度，可以得到该图的超链接矩阵



$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 \\ 1/2 & 0 & 1/2 & 1/3 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 1 & 1/3 & 0 \end{bmatrix} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{matrix}$$

- 超链接矩阵的特点：
 - 所有元素非负
 - 每列元素的总和为 1

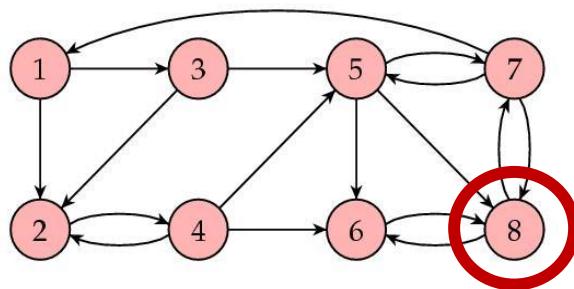
随机矩阵 (Stochastic Matrix)
马尔可夫矩阵

矩阵的特征向量和特征值

$$xA = \lambda x$$

定理：超链接矩阵 H 的最大特征向量即为该矩阵的 PageRank 值

$$I = H \cdot I \rightarrow I \text{ 是 } H \text{ 的对应于特征值 } \lambda=1 \text{ 的特征向量}$$



$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 \\ 1/2 & 0 & 1/2 & 1/3 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 1 & 1/3 & 0 \end{bmatrix} \quad I = \begin{bmatrix} 0.0600 \\ 0.0675 \\ 0.0300 \\ 0.0675 \\ 0.0975 \\ 0.2025 \\ 0.1800 \\ 0.2950 \end{bmatrix}$$

数学的奇妙：原来不知如何下手的互联网网页的排序问题，现在已经轻而易举地变成了求解矩阵 H 的特征向量问题

如何计算 PageRank 值？

幂迭代方法

$$I^{k+1} = H \cdot I^k$$

$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 \\ 1/2 & 0 & 1/2 & 1/3 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 1 & 1/3 & 0 \end{bmatrix}$$

I^0	I^1	I^2	I^3	I^4	...	I^{60}	I^{61}
1	0	0	0	0.0278	...	0.06	0.0600
0	0.5	0.25	0.1667	0.0833	...	0.0675	0.0675
0	0.5	0	0	0	...	0.03	0.0300
0	0	0.5	0.25	0.1667	...	0.0675	0.0675
0	0	0.25	0.1667	0.1111	...	0.0975	0.0975
0	0	0	0.25	0.1806	...	0.2025	0.2025
0	0	0	0.0833	0.0972	...	0.18	0.1800
0	0	0	0.0833	0.3333	...	0.295	0.2950

$$I = \begin{bmatrix} 0.0600 \\ 0.0675 \\ 0.0300 \\ 0.0675 \\ 0.0975 \\ 0.2025 \\ 0.1800 \\ 0.2950 \end{bmatrix}$$

PageRank 算法

- 第一步：将互联网作为一个有向图，并用邻接矩阵进行表示；
- 第二步：将该邻接矩阵转换为超链接矩阵；
- 第三步：求解该超链接矩阵的最大特征向量（如幂迭代法）；
- 第四步：求得的特征向量中的值即为对应网页的 PageRank 值。

PageRank算法



- 这一漂亮的想法出自于 Stanford 大学 1998 年在读博士研究生 *Larry Page* 和 *Sergey Brin*
- 第七次国际 World Wide Web 会议 (WWW'98) 上的论文 "*The PageRank citation ranking : Bringing order to the Web*"
- PageRank 算法中使用的数学知识包括：矩阵的性质、特征值和特征向量、幂迭代方法等

参考文献

- L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, *Technical Report*, Stanford University, 1998.
- K. Bryan, T. Leise, The \$25,000,000,000 eigenvector: The linear algebra behind Google, *SIAM Review*, 48 (3), 569-81, 2006.
- P. Berkin, A survey on PageRank computing, *Internet Mathematics*, 2:73–120, 2005.

数据科学的数学基础

矩阵

线性代数

微积分

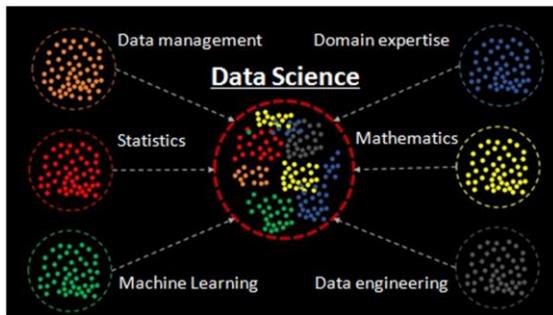
概率论

统计

图论

数据思维与实践

第02讲 数据科学的数学基础



- 矩阵和线性代数
- 概率统计基础
- 微积分与图论
- 开源数字王国的排行榜与网络分析

矩阵和线性代数

- **矩阵 (Matrix)** 是一个按照长方阵列排列的复数或实数集合。涉及到的机器学习应用有SVD、PCA、最小二乘法、共轭梯度法等。
- **线性代数** 是研究向量、向量空间、线性变换等内容的数学分支。向量是线性代数最基本的内容。中学时，数学书告诉我们向量是空间（通常是二维坐标系）中的一个箭头，它有方向和数值。在数据科学家眼中，向量是有有序的数字列表。线性代数是围绕向量加法和乘法展开的。
- 矩阵和线性代数是一体的，矩阵是描述线性代数的参数。它们构成了数据科学的庞大基石。

标量、向量和矩阵

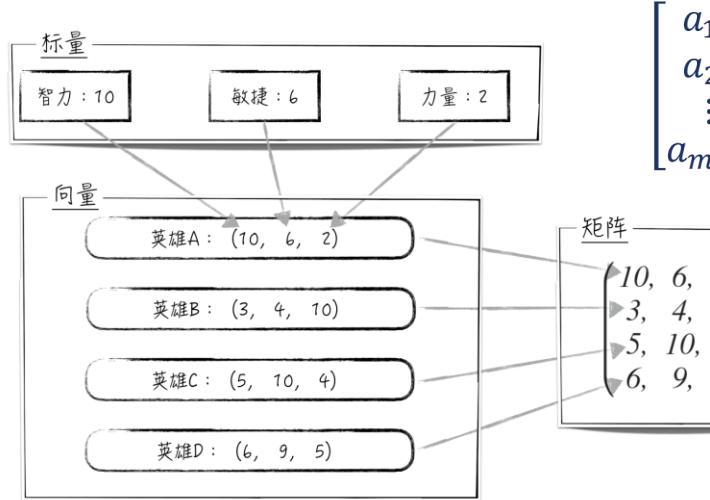


X-lab2017 / open-digger

Type to search

Code Issues 20 Pull requests 3 Discussions Actions Projects Wiki Security Insights

open-digger Public 31 8 14 Edit Pins Watch 20 Fork 78 Starred 261



向量和矩阵的数学表示:

$$\mathbf{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,m})$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} = (x_{i,j}) \in \mathbb{R}^{n \times m}$$

特殊的矩阵

- **方阵** (squared matrix) : 行数等于列数的矩阵
- **单位矩阵** (identity matrix) : 对角线元素等于 1, 其他元素等于 0

$$I_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = (1_{\{i=j\}}) \in \mathbf{R}^{n \times n} \quad E_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- **对角矩阵** (diagonal matrix) : 除对角线元素外, 其他元素都等于 0

$$diag(d_1, d_2, \dots, d_n) = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & d_n \end{pmatrix} \in \mathbf{R}^{n \times n}$$

- **三角矩阵** (triangular matrix) : 上三角矩阵的对角线下方的元素全部为零;
下三角矩阵的对角线上方的元素全部为零

$$U = \begin{pmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,n} \\ 0 & u_{2,2} & \cdots & u_{2,n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & u_{n,n} \end{pmatrix} \in \mathbf{R}^{n \times n} \quad L = \begin{pmatrix} l_{1,1} & 0 & \cdots & 0 \\ l_{2,1} & l_{2,2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ l_{n,1} & l_{n,2} & \cdots & l_{n,n} \end{pmatrix} \in \mathbf{R}^{n \times n}$$

矩阵的运算 1

• 矩阵的加法和减法

$$X \pm Y = \begin{pmatrix} x_{1,1} \pm y_{1,1} & \dots & x_{1,m} \pm y_{1,m} \\ \vdots & & \vdots \\ x_{n,1} \pm y_{n,1} & \dots & x_{n,m} \pm y_{n,m} \end{pmatrix} = (x_{i,j} + y_{i,j}) \in \mathbf{R}^{n \times m} \quad \begin{bmatrix} 2 & 1 \\ 5 & 3 \end{bmatrix} + \begin{bmatrix} -1 & 2 \\ -5 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 0 & 3 \end{bmatrix}$$

矩阵的加法满足结合律和交换律

$$X + Y = Y + X$$

$$X + Y + Z = X + (Y + Z)$$

• 矩阵的乘法：

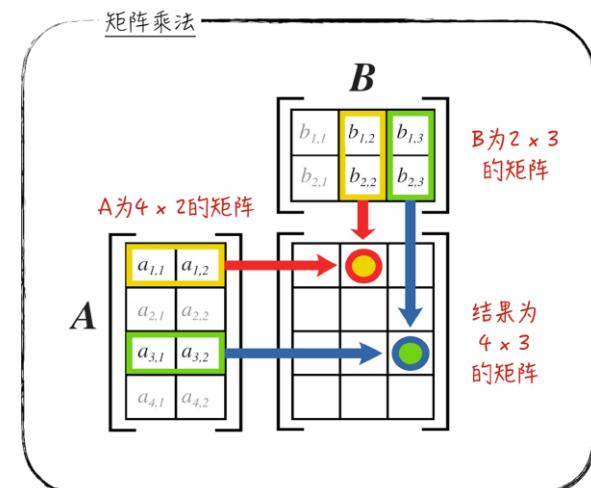
- 矩阵与数字的乘法
- 矩阵与矩阵的乘法

$$kX = \begin{pmatrix} kx_{1,1} & \dots & kx_{1,m} \\ \vdots & & \vdots \\ kx_{n,1} & \dots & kx_{n,m} \end{pmatrix} = (kx_{i,j}) \in \mathbf{R}^{n \times m}$$

$$AB = \left(\sum_{r=1}^p a_{i,r} b_{r,j} \right) \in \mathbf{R}^{n \times m}$$

$$A = \begin{bmatrix} 1 & 2 & 2 \\ 5 & 3 & 1 \\ -1 & 5 & -2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 & 3 \\ 4 & 2 & 1 \\ 5 & 0 & 3 \end{bmatrix} \quad AB = \begin{bmatrix} 19 & 5 & 11 \\ 22 & 11 & 21 \\ 9 & 9 & -4 \end{bmatrix}$$

矩阵的乘法满足结合律以及分配律，但不满足交换律



矩阵的运算 2

- 矩阵的转置： $m \times n$ 矩阵行与列互换得到 $n \times m$ 矩阵

$$A = \begin{bmatrix} 1 & 2 & 2 \\ 5 & 3 & 1 \\ -1 & 5 & -2 \end{bmatrix}, \quad A^T = \begin{bmatrix} 1 & 5 & -1 \\ 2 & 3 & 5 \\ 2 & 1 & -2 \end{bmatrix}$$

- 逆矩阵

矩阵的逆： A, B 是 n 阶方阵，若 $AB = BA = E$ ，则称 A 是可逆矩阵，同时 B 是 A 的逆矩阵，记为 A^{-1} 。

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad \text{可求得 } A^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \end{bmatrix}$$

矩阵的性质：特征值与特征向量

- 特征值与特征向量： $A\xi = \lambda\xi$ 。

【例】求矩阵 $A = \begin{bmatrix} 2 & -1 & 2 \\ 5 & -3 & 3 \\ -1 & 0 & -2 \end{bmatrix}$ 的特征值和特征向量

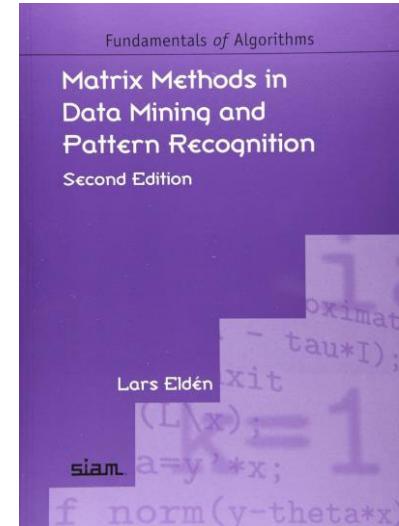
根据特征方程有：

$$0 = |\lambda E - A| = \begin{vmatrix} \lambda - 2 & 1 & -2 \\ -5 & \lambda + 3 & -3 \\ 1 & 0 & \lambda + 2 \end{vmatrix} = (\lambda + 1)^3$$

得 $\lambda_1 = \lambda_2 = \lambda_3 = -1$ ，令 $(-\mathbf{E} - \mathbf{A})\mathbf{x} = \mathbf{0}$ ，得特征向量 $k[1, 1, -1]^T$ 。

□ 其他性质：

- 矩阵的秩**: A 的线性独立的纵列的极大数目
- 矩阵范数**: 通俗的说是度量矩阵“大小”的概念



矩阵的性质：矩阵的奇异值

- 矩阵的奇异值：设 A 为 $m \times n$ 矩阵， $A^T A$ 的特征值为 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > \lambda_{r+1} = \dots = \lambda_n = 0$ ，则称 $\sigma_i = \sqrt{\lambda_i}$ ($i = 1, 2, \dots, n$) 为 A 的奇异值， A 为零矩阵时，它的奇异值均为0。

【例】求 $A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$ 的奇异值。

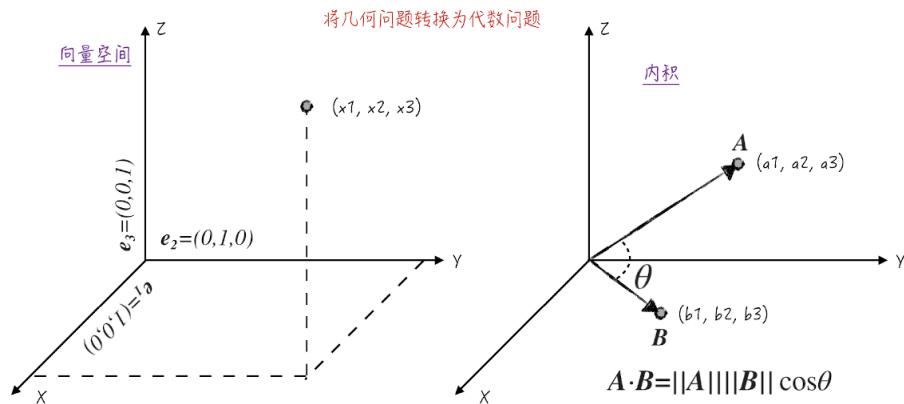
$B = A^T A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}$ ，列特征方程求得 B 的特征值 $\lambda_1 = 3, \lambda_2 = 1, \lambda_3 = 0$ ，奇异值为 $\sqrt{3}, 1, 0$ 。

向量空间

我们生活的世界在空间上是一个三维空间，在这个现实世界里建立长宽高坐标系 (x, y, z) 。这个现实中的每一个点都能被表示成一个三维行向量。从数学上来看，任意一个三维的行向量 $\mathbf{X} = (x_1, x_2, x_3)$ ，可以被写为：

$$\mathbf{X} = x_1 (1, 0, 0) + x_2 (0, 1, 0) + x_3 (0, 0, 1)$$

用学术些的话来表述就是任意一个三维行向量可以被 $e_1 = (1, 0, 0), e_2 = (0, 1, 0), e_3 = (0, 0, 1)$ 这 3 个行向量线性表示，而且反过来， e_1, e_2, e_3 这 3 个行向量的任意线性组合都对应现实空间中某一点。

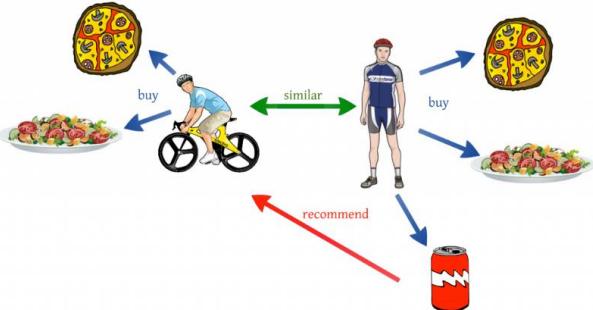


像这样的空间被称为**向量空间**，而 e_1, e_2, e_3 被称为**向量空间的基**。

实例 01：利用 SVD 进行评分预测

- 矩阵奇异值分解

- 在推荐系统中，一类重要的方法称为协同过滤，即根据相似度给相似的用户推荐相似的商品，其中最具代表性的算法就是矩阵分解——如奇异值分解（SVD矩阵分解）。



$$\boxed{A} = \boxed{X} \boxed{B} \boxed{Y}$$

矩阵奇异值分解

- 设矩阵 $A \in \mathbb{R}^{m \times n}$ 的秩 $r > 0$ ，则存在 m 阶正交矩阵 U 和 n 阶正交矩阵 V ，使得 $A = U \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} V^T$ ，其中 $\Sigma = diag(\sigma_1, \sigma_2, \dots, \sigma_r)$ ，为矩阵 A 的全部非零奇异值构成的对角矩阵。称此为矩阵 A 的奇异值分解（一般不唯一）。

实例 01：利用 SVD 进行评分预测

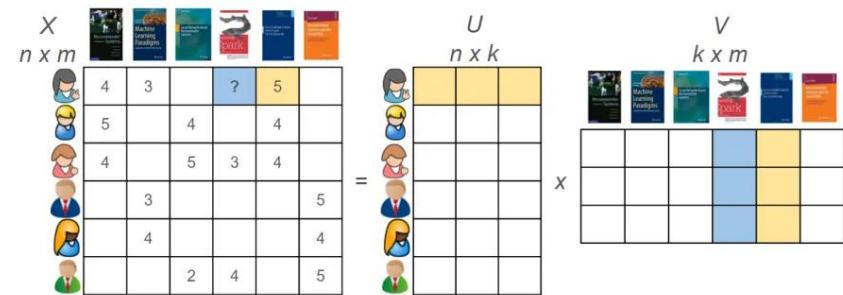
【例】求矩阵 $A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$ 的SVD分解。

解：可知 $B = A^T A$ 的特征值有 $\lambda_1 = 3, \lambda_2 = 1, \lambda_3 = 0$ ，对应的特征向量分别为

$$\xi_1 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \quad \xi_2 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \quad \xi_3 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$$

于是可得 $r(A) = 2$, $\Sigma = \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \end{bmatrix}$, 令 V 等于归一化后的特征向量拼接构成的矩阵，即

$$V = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{2}{\sqrt{6}} & 0 & -\frac{1}{\sqrt{3}} \end{bmatrix}$$



概率统计基础

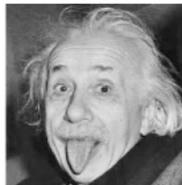
Statistics
Probability
Median
Mode
Deviation
Mean
Standard
Distribution
Population
Event
Experiment
Union
Intersection
Independence
Randomness
Skewed
Normal
Box-plot
Observation
Dotplot
Pie-chart
Histogram
Bargraph
Conditional
Random
Chance
Stem-leaf
Z-score
Standard deviation
Median
Mode
Deviation
Mean
Standard
Distribution
Population
Event
Experiment
Union
Intersection
Independence
Randomness
Skewed
Normal
Box-plot
Observation
Dotplot
Pie-chart
Histogram
Bargraph
Conditional
Random
Chance

什么是概率？

在现实世界里，充满了各种随机事件

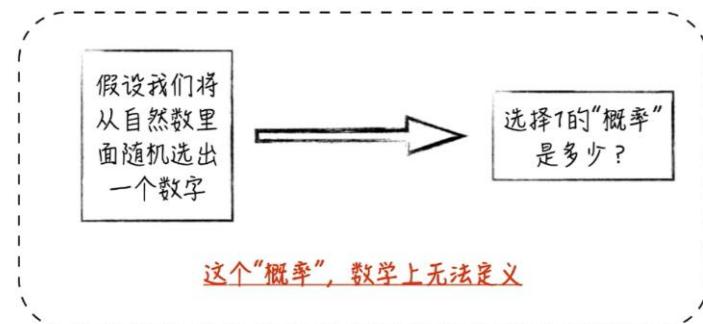
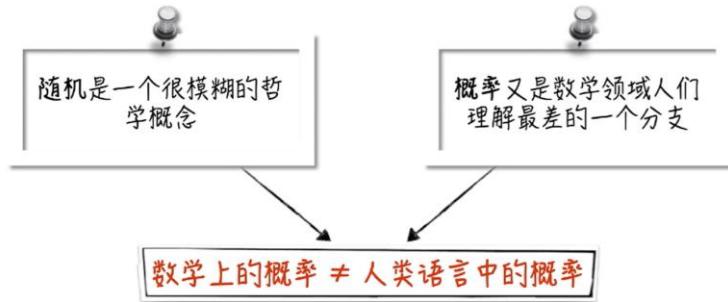
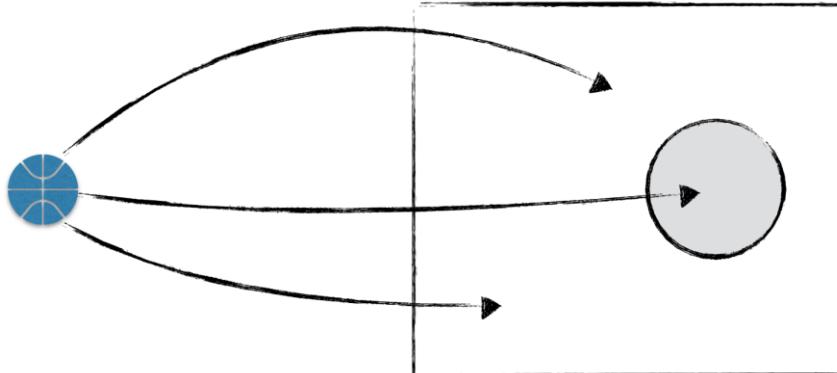


- 彩票的中奖、掷骰子的点数
- 微观世界里，粒子出现的位置

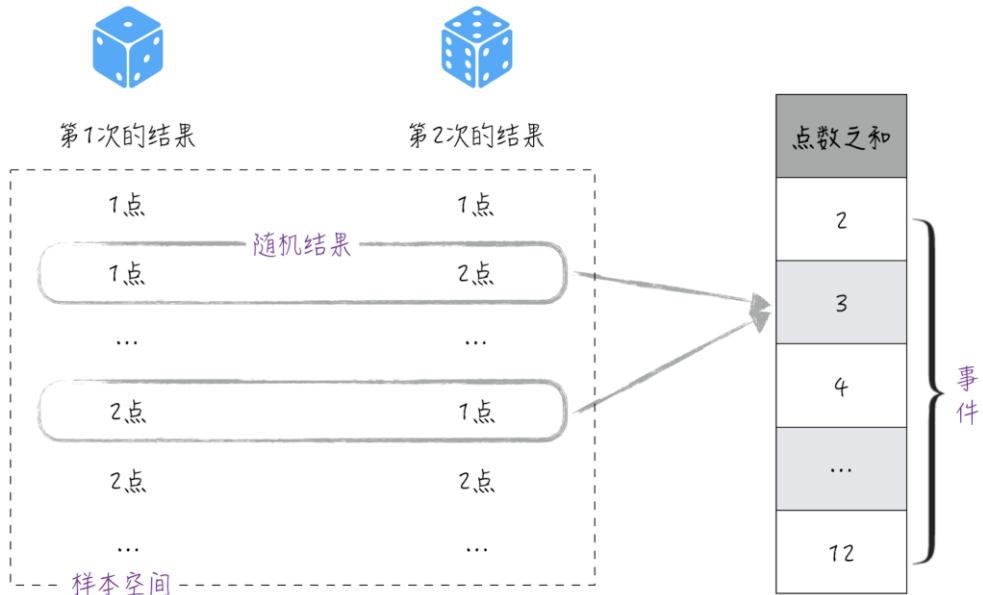


概率是用来刻画随机的一种数学工具

球落入圆圈的概率 = 圆圈面积 / 方框面积



概率的定义



- 在随机结果有限的情况下：
 - 定义样本空间 S : 所有随机结果 ω 组成的集合
 - 定义概率: 满足如下三个条件的, 从样本空间到实数的函数
 - $P(\omega) \geq 0$
 - $\sum_{\omega \in S} P(\omega) = 1$
 - $P(E) = \sum_{\omega \in E} P(\omega)$ 其中 E 为任意一个事件

条件概率

假设在一个大学班级里：

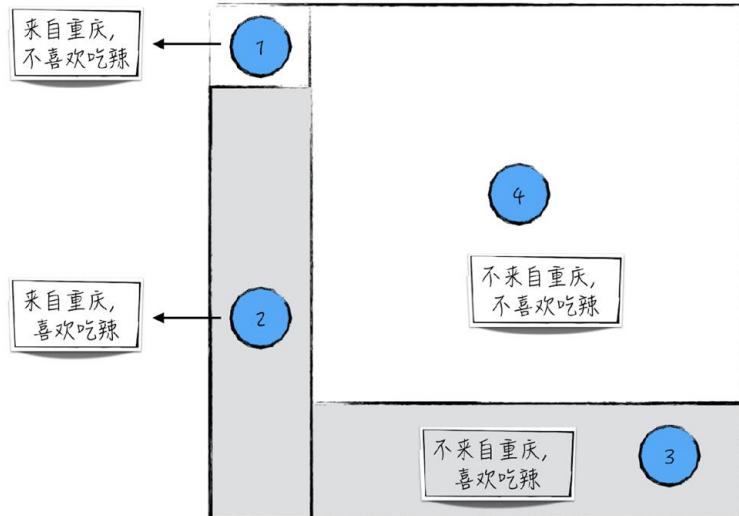
- 来自重庆的学生比例为10%，而这批学生中喜欢吃辣的比例为90%
- 来自其他省份的学生比例为90%，而这些学生中喜欢吃辣的比例为10%

用事件A表示学生
来自重庆

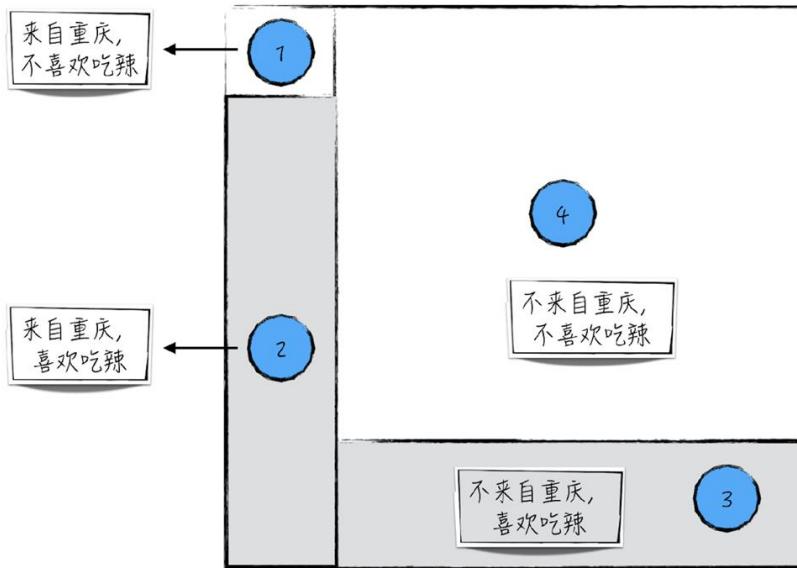
用事件B表示学生
喜欢吃辣

想知道：已知学生喜欢吃辣的情况下，学生来自重庆的概率是多少？

这就是条件概率想要解决的问题，记为： $P(A | B)$



条件概率



来自重庆的概率： $P(A) = ((1) + (2)) \div ((3) + (4) + (1) + (2))$

喜欢吃辣的条件下，
来自重庆的概率： $P(A|B) = (2) \div ((2) + (3))$

对于两个事件A、B

- 两个事件同时发生记为 $A \cap B$
- 定义已知B发生的情况下，A发生的条件概率为

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

同理可以得到：

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

结合上面两个条件概率的定义，可以得到如下的贝叶斯公式

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

量化信息的价值

再次回顾问题的描述

假设在一个大学班级里：

- 来自重庆的学生比例为10%，而这批学生中喜欢吃辣的比例为90%
- 来自其他省份的学生比例为90%，而这些学生中喜欢吃辣的比例为10%

记号描述

用事件A表示学生来自重庆

用事件B表示学生喜欢吃辣

翻译成数学语言

已知：

$$P(A) = 0.1, P(B|A) = 0.9$$

$$P(A^c) = 0.9, P(B|A^c) = 0.1$$

求解

$$P(A|B) = ?$$

已知一个学生喜欢吃辣这一条信息

他是重庆的人概率从10%上升到50%

$$P(A) = 0.1$$

$$P(A|B) = 0.5$$

量化信息的价值

独立事件



若 $P(A) = P(A|B)$ 则称A、B为相互独立事件

$$P(A \cap B) = P(A)P(B)$$

同理可以定义任意多个相互独立的事件

事件 A_1, \dots, A_n 相互独立，当且仅当，对于任意子集， A_{i1}, \dots, A_{ik} ，都成立

$$P(A_{i1} \cap \dots \cap A_{ik}) = P(A_{i1}) \dots P(A_{ik})$$

条件概率与事件独立性

【例】两个口袋各10个球，第一个有10个白球，第二个有7个白球，3个黑球，所有球除颜色外无其他区别，从一个口袋中取出一个球，已知为白球，放回原口袋。再从该口袋中取一个球，求该球是黑球的概率。

解：假设 $H_i(i = 1,2)$ 为“从第*i*个口袋中取球”， A 为“取白球”，则：

$$P(H_1) = P(H_2) = \frac{1}{2}, \quad P(A|H_1) = 1, \quad P(A|H_2) = \frac{7}{10}$$

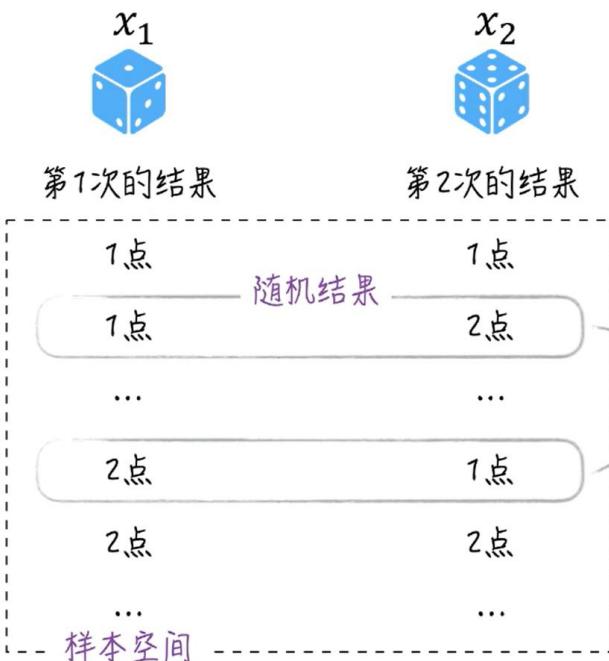
$$P(A) = P(A|H_1)P(H_1) + P(A|H_2)P(H_2) = \frac{17}{20}$$

$$P(H_1|A) = \frac{P(H_1)P(A|H_1)}{P(A)} = \frac{10}{17}, \quad P(H_2|A) = \frac{P(H_2)P(A|H_2)}{P(A)} = \frac{7}{17}$$

已知第一次取的是白球，而且从同一袋中去取，则有：

$$P(\bar{A}|A) = P(\bar{A}|H_1)P(H_1|A) + P(\bar{A}|H_2)P(H_2|A) = 0 \times \frac{10}{17} + \frac{3}{10} \times \frac{7}{17} = \frac{21}{170}$$

随机变量



刻画随机变量的方法

离散型随机变量



对于离散型的随机变量，使用概率分布
函数来刻画它

$$P(x_1 = i) = 1/6; i = 1, \dots, 6$$

连续型随机变量

对于连续型的随机变量，使用概率密度
函数来刻画它

$$P(a \leq x \leq b) = \int_a^b f_x(t) dt$$

累积分布函数

期望

方差

协方差

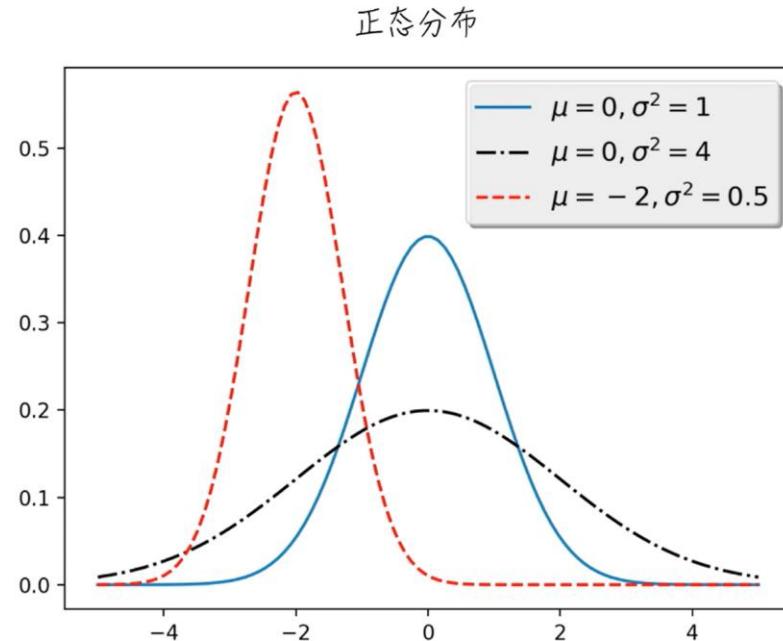
$$\begin{aligned} C_x(a) &= P(x \leq a) & E[x] &= \sum_i P(x = x_i)x_i & Var(x) &= E[(x - E[x])^2] & Cov(x, y) &= E[(x - E[x])(y - E[y])] \\ & & & & & = E[x^2] - (E[x])^2 & \uparrow & = E[xy] - E[x]E[y] \\ E[x] &= \int xf(x)dx & & & & & & \end{aligned}$$

正态分布

正态分布也称高斯分布，是最为重要的
一种概率分布

- 如果一个随机变量服从正态分布，则
它可能的取值是任意实数
- 相应的概率密度函数如下

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-(x-\mu)^2/2\sigma^2}$$



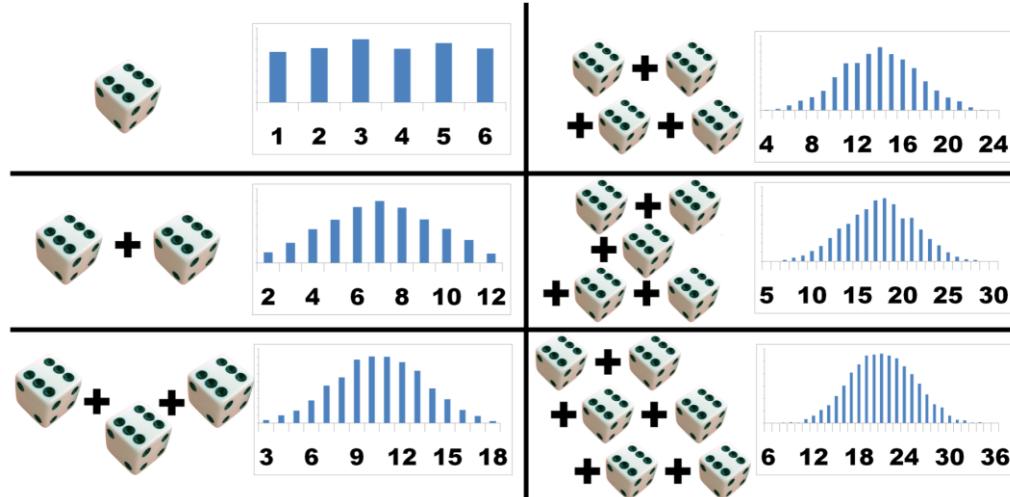
中心极限定理

中心极限定理

n个独立同分布的随机变量相叠加，得到的和将越来越近似于一个正态分布

在实际中，一个随机现象往往
是多个随机因素的叠加

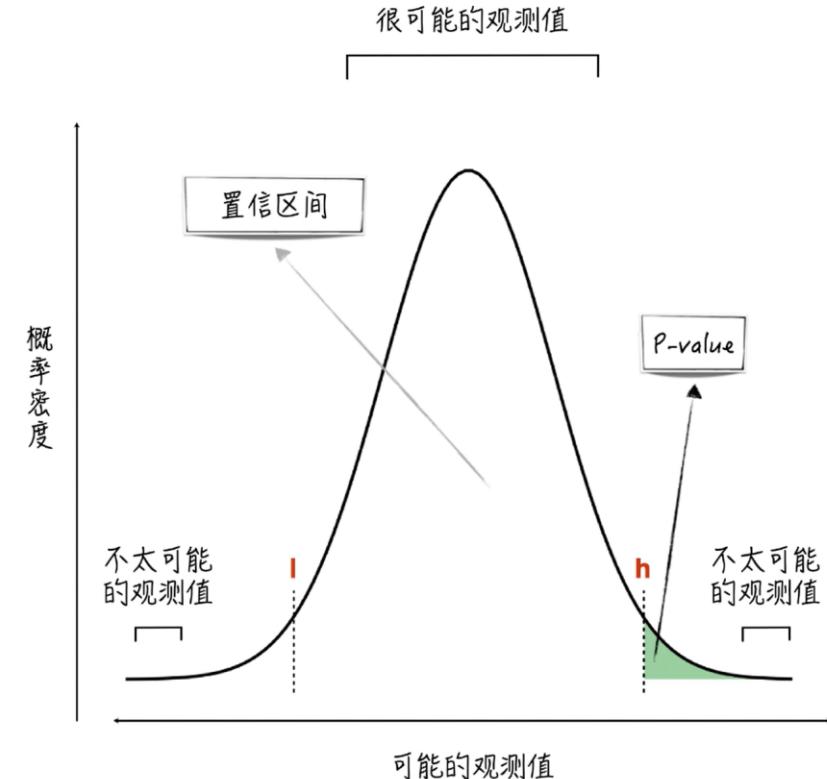
很多随机现象的分布都可以用
正态分布来描述



置信区间

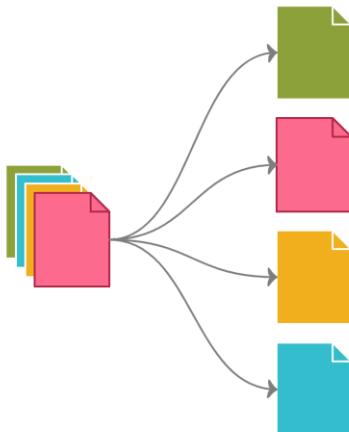
置信区间：概率值等于 a ，且以期望为中心的对称区域（在实际中 a 常常等于0.95或者0.99）

对于置信区间的两个边界值 l, h ，它们的P-value为 $(1 - a)/2$



实例 02：利用朴素贝叶斯算法进行文本分类

- 利用朴素贝叶斯算法进行文本分类
 - N-gram是一种基于概率的判别的语言模型，基于这样一种假设：第N个词的出现只与前面N-1个词相关，而与其它任何词都不相关。因此词序列 w_1, w_2, \dots, w_m 出现的概率：
$$p(w_1, w_2, \dots, w_m) = p(w_1) * p(w_2 | w_1) \dots p(w_m | w_1, \dots, w_{m-1})$$



- 当N=1时，称为一元语言模型，此时模型简化为
$$P(w_1, w_2, \dots, w_m) = \prod_{i=1}^m P(w_i)$$
, 其中 $P(w_i) = \frac{C(w_i)}{M}$, $C(\cdot)$ 表示词在训练语料中出现的次数，M是语料库中的总字数。

实例 02：利用朴素贝叶斯算法进行文本分类

【例】现在假设有语料，左三句为侮辱性句子，右三句非侮辱性句子：

Maybe not take him to dog park, stupid.

Stop posting stupid worthless garbage.

My dog has flea problems, help please.

My dalmatian is so cute. I love him.

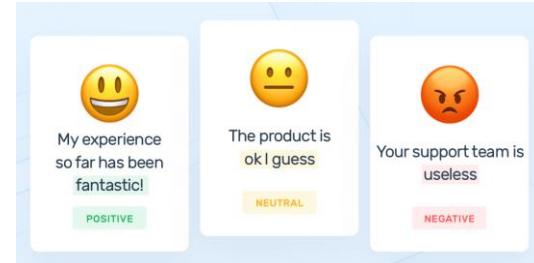
Quit buying worthless dog food, stupid.

Mr. Licks ate my steak. How to stop him?

为了避免测试中出现某些未在训练语料中出现过的词，从而造成概率值归零的情形出现，将每个词的出现次数都自增1，以词“stupid”为例，计算其在侮辱性句子和非侮辱性句子中出现的概率。

$$P(\text{stupid}|\text{侮辱性}) = \frac{3 + 1}{19} \approx 0.2105$$

$$P(\text{stupid}|\text{非侮辱}) = \frac{0 + 1}{24} \approx 0.0417$$



可以看出，两种类别下的词分布有着明显的差异，从而使得分类器可以生效。

实例 02：利用朴素贝叶斯算法进行文本分类

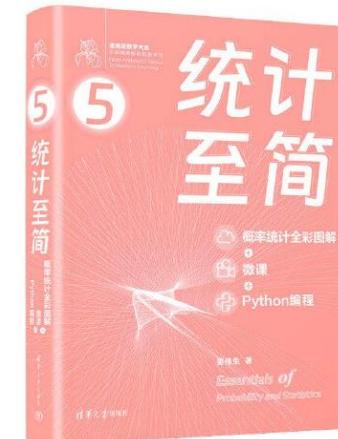
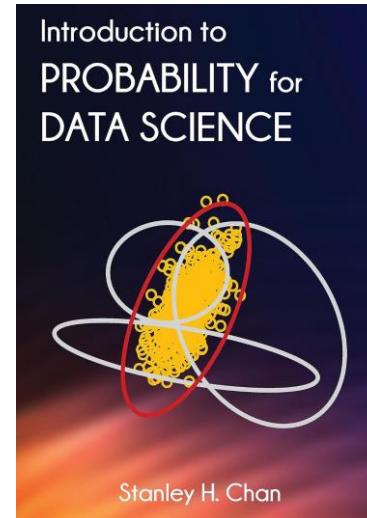
考虑测试用例：Stupid garbage，我们应用朴素贝叶斯公式计算此句是侮辱性句子的概率。

$$\begin{aligned} P(\text{侮辱性}|\text{Stupid garbage}) &= \frac{P(\text{侮辱性})P(\text{stupid}|\text{侮辱性})P(\text{garbage}|\text{侮辱性})}{\sum_{Y=\{\text{侮辱性, 非侮辱}\}} P(\text{stupid}|Y)P(\text{garbage}|Y)} \\ &= \frac{\frac{3}{6} \times \frac{4}{19} \times \frac{2}{19}}{\frac{4}{19} \times \frac{2}{19} + \frac{1}{24} \times \frac{1}{24}} \\ &= 0.9274 \end{aligned}$$

同理 $P(\text{非侮辱}|\text{Stupid garbage}) = 0.0726$ ，从而将其正确分类为“侮辱性句子”。

面向数据科学的概率统计

- Random Variables (随机变量)
- Bayes Theorem (贝叶斯定理)
- Cumulative Distribution Function (累计分布函数)
- Continues Distributions (连续分布)
- Probability Density Function (概率密度函数)
- ANOVA (方差分析)
- Central Limit Theorem (中心极限定理)
- Monte Carlo Method (蒙特卡罗方法)
- Hypothesis Testing (假设检验)
- p-Value (P值)
- Estimation (估计)
- Confidence interval (置信区间)
- Maximum Likelihood Estimate (极大似然估计)
- Regression (回归)
- Covariance (协方差)
- Correlation (相关性)
- Pearson correlation coefficient (Pearson 相关系数)
- Least Squares Fitting (最小二乘法)
- Euclidean Distance (欧氏距离)



统计学

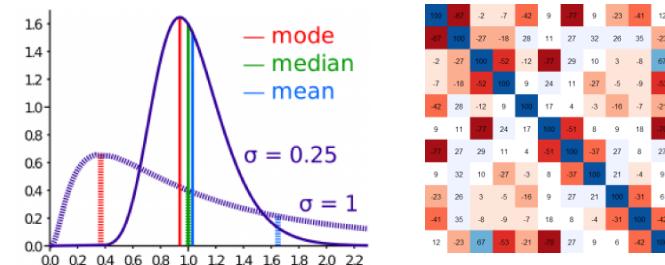


- **统计学 (Statistics)** 是指用于理解数据的**数学和技术**。
- 通过搜索、整理、分析、描述数据等手段，以达到推断所测对象的本质，甚至预测对象未来的一门综合性科学。
- 事物的发展充满了**不确定性**，而统计学，既研究如何从数据中把信息和规律提取出来，找出最优化的方案；也研究如何把数据当中的不确定性量化出来。

统计的起源

- 用单个数或者数的小集合捕获可能很大值集的各种特征，如：
 - 频率度量：**众数
 - 位置度量：**均值和中位数
 - 散度度量：**极差和方差
 - 数据分布：**频率表、直方图
 - 多元汇总统计：**相关矩阵、协方差矩阵

汇总数据的初衷如公司的组织结构，高层期望看到工作概要，而不是细节



统计量的设计

汇总数据指标的设计，源于非常朴素的思想。

Standard Deviation Formula

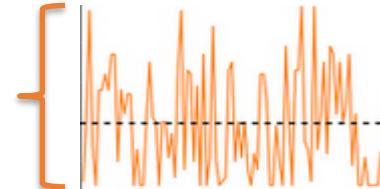


$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{N}}$$

- **(标准差)** 设计一个指标，可以用来衡量数据集合的发散性，经过如下思考：
 - 每个样本的偏差累加就可以衡量 $\sum (x_i - \hat{x}_i)$
 - 偏差较大的值应该具有更大的权重 $\sum (x_i - \hat{x}_i)^2$
 - 集合中数字越多，方差越大，应该与集合大小无关 $\frac{\sum (x_i - \hat{x}_i)^2}{n}$
 - 量纲与原始数据不同，无法比 $\sqrt{\frac{\sum (x_i - \hat{x}_i)^2}{n}}$
 - 最终结果，RMSE（均方根误差）



貌似这个宽度就
可以体现数据的
波动性大小



回顾：数据分析 (AD) 的四个层次

- **描述性分析** (Descriptive analysis) : 发生了什么 (过去与现在)
- **诊断性分析** (Diagnostic analysis) : 发生的原因 (动因与洞察)
- **预测性分析** (Predictive analysis) : 将要发生什么 (趋势与可能)
- **指导性分析** (Prescriptive analysis) : 应该做什么 (决策与优化)

描述性统计
(Descriptive Statistics)

探索性统计
(Exploratory Statistics)

推断性统计
(Inferential Statistics)

统计决策
(Statistical decision)

统计分析方法

– 描述性统计 (Descriptive Statistics) : 解释数据的一些特征

描述性统计分析包括数据的频数分析、数据的集中趋势分析（如均值、中位数、众数等）、数据的离散程度分析（如标准差、极差等）、数据的分布（如偏度值、峰度值等）以及一些基本的统计图形（如饼图、直方图、箱线图等）

– 探索性统计 (Exploratory Statistics) : 关注数据内在规律

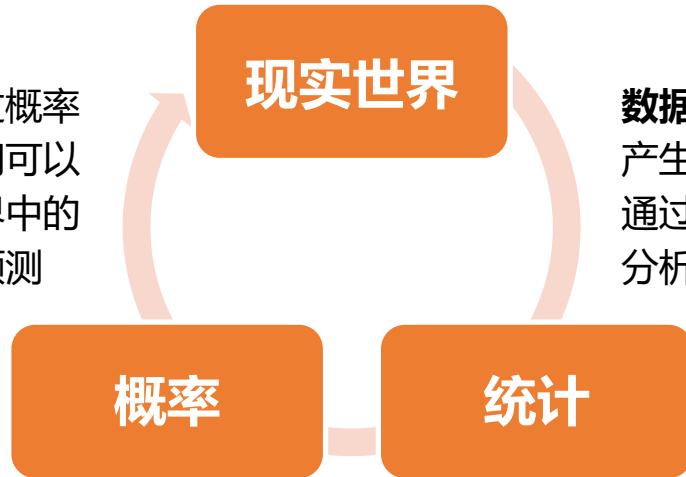
探索性统计分析主要用于数据分析过程中的探索，通过探索可以发现数据背后隐藏的内在规律和联系，通常探索性统计分析还可以挖掘出数据中出现异常的原因。例如，变量之间是否存在一定的相关性、两组样本之间是否存在显著的差异、探索企业内某指标（如曝光量、广告点击率、支付成功率、某支付渠道占比等）没有达标的原因，探索企业内某指标在接下来的一段时间内将会有怎样的变化趋势等。

– 推断性统计 (Inferential Statistics) : 怎样用已知数据来进行预测和判断

统计学实质上就是根据样本的特征来推断总体的情况。例如，借助于随机抽样的方法，从总体中抽出部分样本，并根据样本推断出总体的平均水平（解决问题的方法是统计推断中的均值检验）；根据样本的两个属性（即两个变量），判断属性间是否存在相关性（需利用统计推断中的相关系数检验或卡方检验）；根据样本的分布，判断其总体是否服从正态分布（该问题的解决可以使用数据的正态性检验技术）。

关于概率论和统计学的区别

预测：通过概率模型，我们可以对现实世界中的事件做出预测

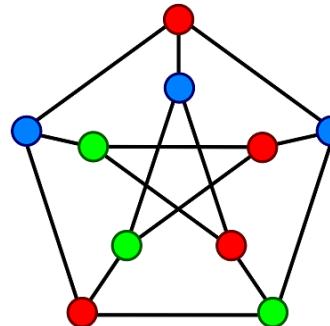
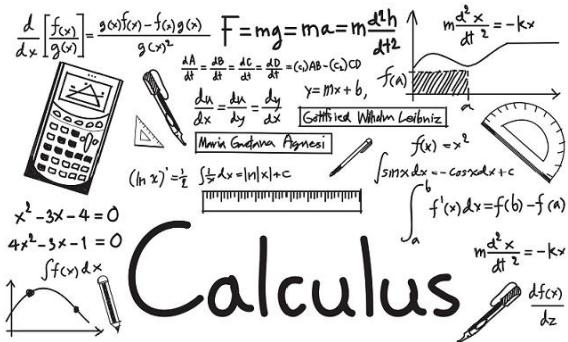


模型：利用统计和数据，我们可以生成概率模型

数据：现实世界产生的数据可以通过统计学进行分析



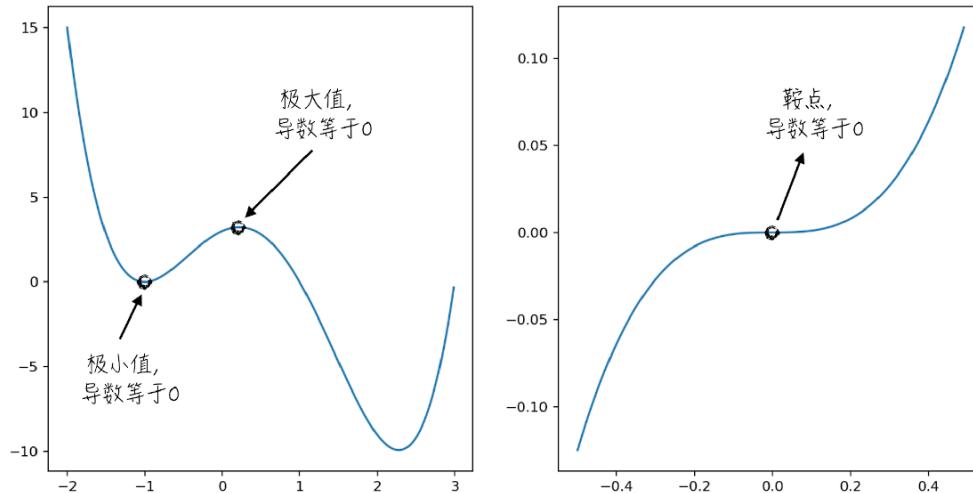
微积分与图论



微积分

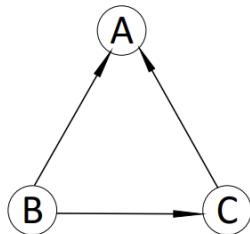
- 导数和积分
- 极限：变化的终点
- 复合函数
- 多元函数与偏导数
- 极值与最值

数据科学中，常常遇到寻求曲线最值点的问题

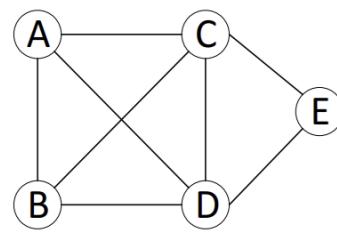


图论

- 图 G 由一个有序二元组组成，记为 $G = (V, E)$, $V = \{v_1, v_2 \dots v_n\}$ 是图 G 中顶点有限非空集合, $E = \{(u, v) | u \in V, v \in V\}$ 是图 G 中两个不同顶点的边的集合。若 E 是有向边（也称为弧）的有限集合时，称图 G 为有向图，若是无向边的有限集合时，称图 G 为无向图。



有向图



无向图

一个有向图和一个无向图

图的顶点

- 图中每个顶点的度为以该顶点为另一个端点的边的数目。
在有向图中可细分入度和出度。

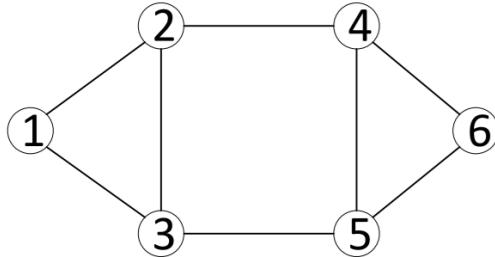
【例】无向图中有17条边，其中度为4的顶点有4个，度为2的顶点有3个，其余顶点度为3，求总节点个数。

$$\sum_{i=1}^n \deg(v_i) = 2e = 34$$

度为4和2的顶点总度数为 $4 \times 4 + 2 \times 3 = 22$ ，于是度为3的顶点个数为 $\frac{34-22}{3} = 4$ ，得总节点个数为11。

图的矩阵表示

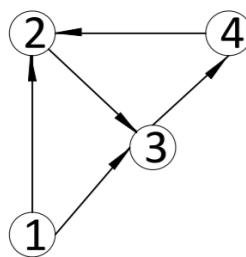
- 邻接矩阵



图与它的邻接矩阵

$$A_1 = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \quad A_2 = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

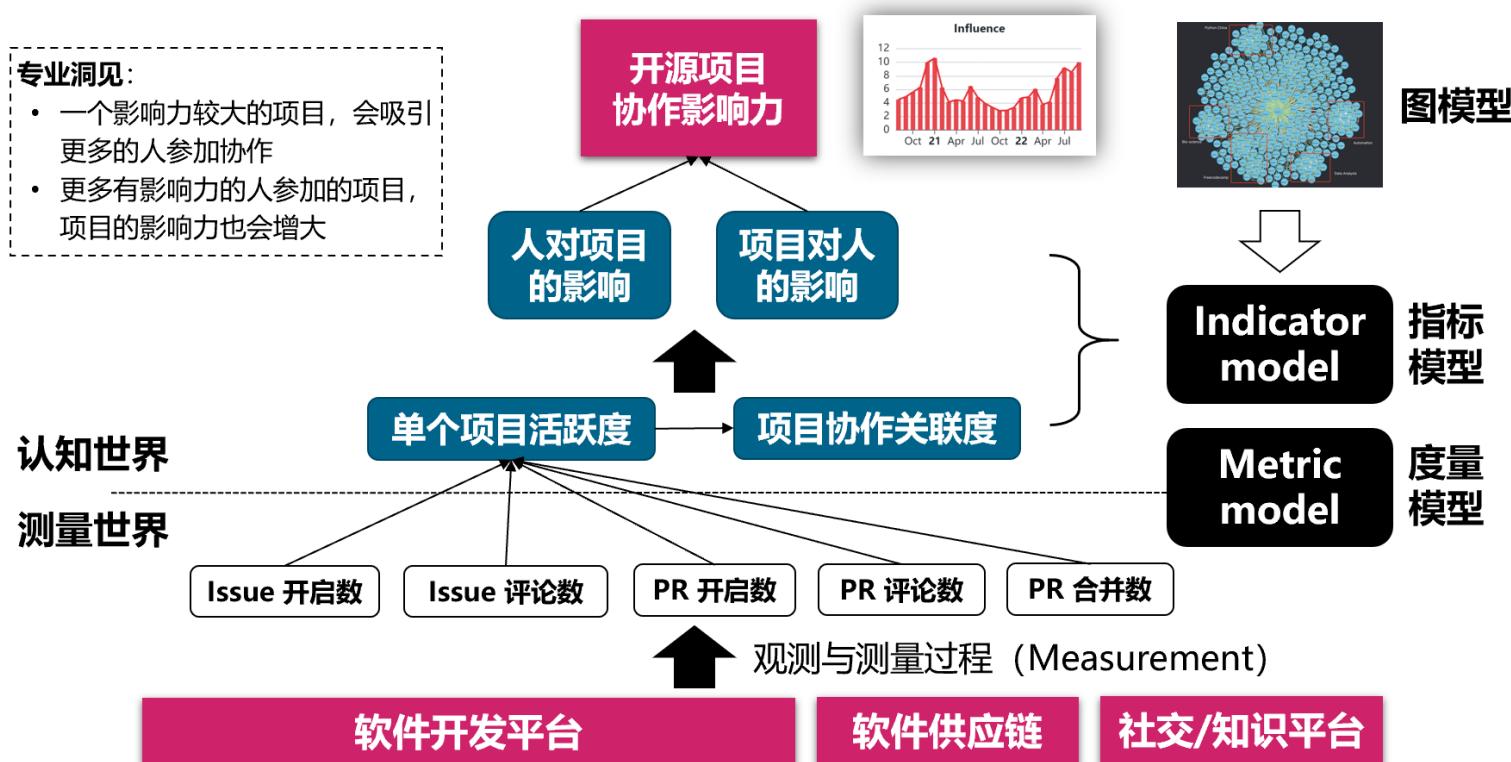
- 关联矩阵



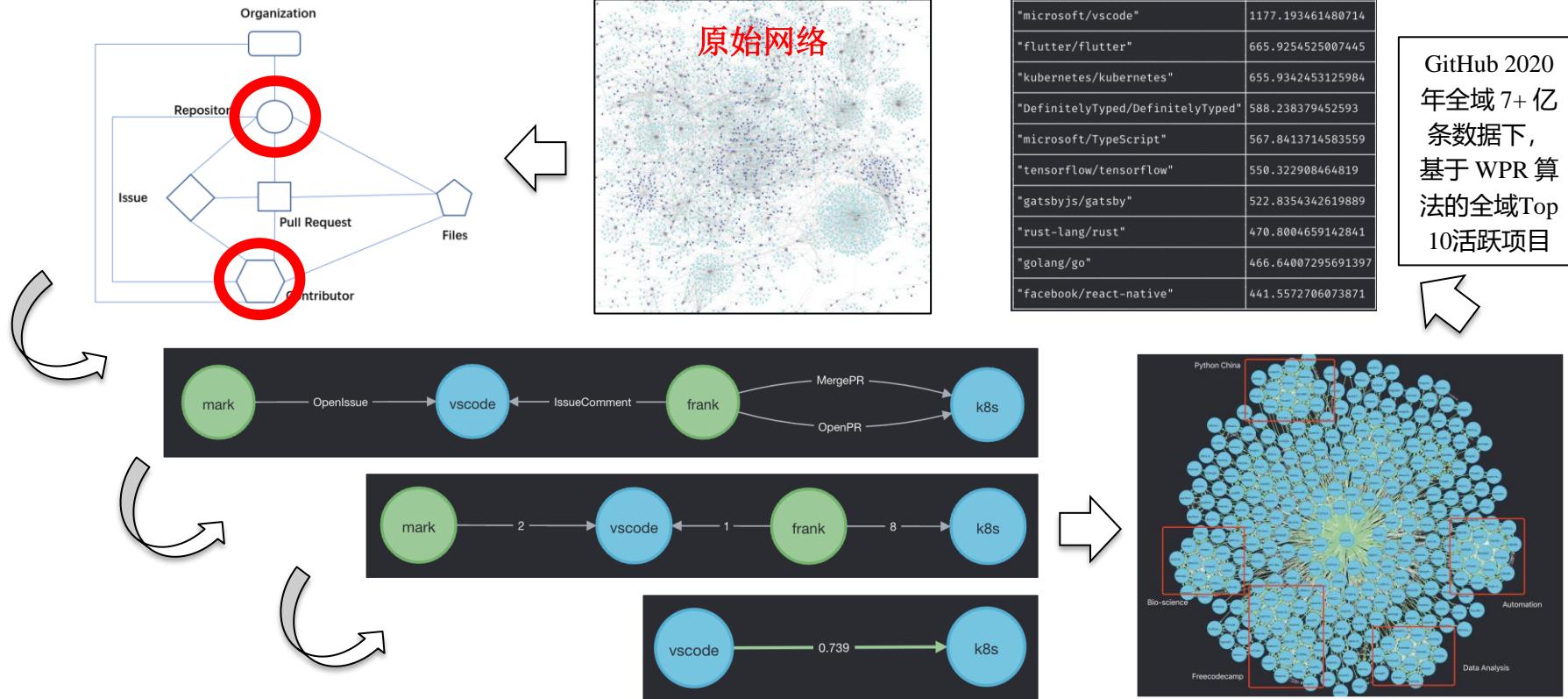
图与它的关联矩阵

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

实例 03：初识 OpenRank 算法

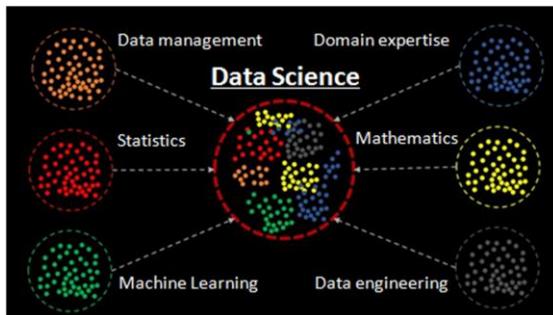


建模过程：异质信息网络降维



数据思维与实践

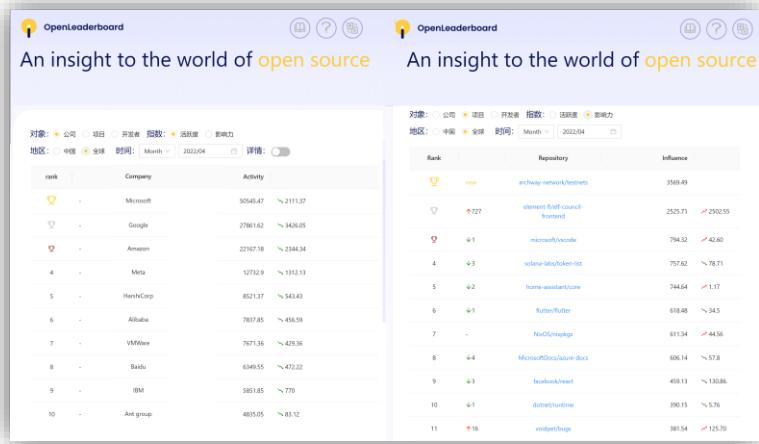
第02讲 数据科学的数学基础



- 矩阵和线性代数
- 概率统计基础
- 微积分与图论
- **开源数字王国的排行榜与网络分析**

第二周的任务

开源数字王国的开源爱好者们，根据 GitHub 上面的开源活动数据，构建了一个开源排行榜 OpenLeaderboard (<https://open-leaderboard.x-lab.info>)，用来展示开源世界中的那些优秀的项目、企业、与开发者，排行榜里面的各种数字引起了李纳斯浓厚的兴趣。



以 2023 年 6 月份的数据为例：

- 用矩阵表示当月排名前 50 的项目活跃度详情。
- 全球排名前 100 项目的 OpenRank 值，最大的是多少、最小的是多少、均值多少、中位数多少？

以 2022 年一整年的数据为例：

- 全球排名前十的项目活跃度与影响力平均增长率多少？
- 中美排名前十的企业，总体活跃度和影响力差距是多大？

第二周的任务

根据 OpenLeaderboard 上对前 10000 个活跃的项目统计，工具组件型项目占比 50%，系统应用型占比 25%，而内容资源型（非软件类）项目占比 25%，成三分天下的态势。

非软件类项目中，带有 HTML/Markdown 标签的项目占 85%，而软件类项目中带 HTML/Markdown 标签的项目占比则为 10%（注：HTML/Markdown 一般可用来书写文档内容）

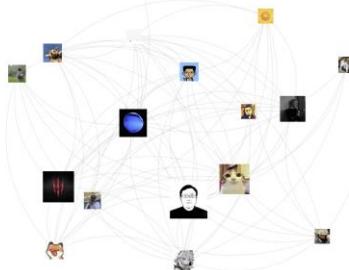
工具组件型项目中，JavaScript 语言的项目占比 35%，而非工具组件型项目中，JavaScript 语言的项目占比则为 10%（注：JavaScript 是一种脚本编程语言，可以在网页上实现复杂的功能）

根据 OpenDigger 仓库中的标注数据：

- 系统软件类型：225 个 (2.25%)
- 应用软件类型：2434 个 (24.34%)
- 组件框架类型：3136 个 (31.36%)
- 软件工具类型：1888 个 (18.88%)
- 内容资源类型：2317 个 (23.17%)

1. 已知一个项目带有 HTML/Markdown 标签，那么该项目是非软件型项目的概率是多少？
2. 已知一个项目是由 JavaScript 语言编写的，那么它是工具组件型项目的概率是多少？

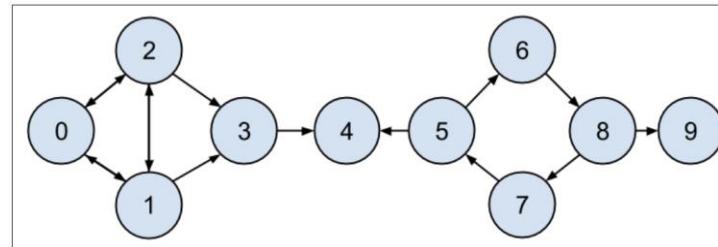
第二周的任务



开源数字王国的市民在 GitHub 上面的开源活动形成了多种多样的社交关系，例如协作关系、Follow 关系、代码评审关系等。

- 谁是这个开源数字王国中最具影响力的开发者？

```
users = [  
    { "id": 0, "name": "Hero" },  
    { "id": 1, "name": "Dunn" },  
    { "id": 2, "name": "Sue" },  
    { "id": 3, "name": "Chi" },  
    { "id": 4, "name": "Thor" },  
    { "id": 5, "name": "Clive" },  
    { "id": 6, "name": "Hicks" },  
    { "id": 7, "name": "Devin" },  
    { "id": 8, "name": "Kate" },  
    { "id": 9, "name": "Klein" }  
]
```



GitHub Social Network (follow)

```
endorsements = [(0, 1), (1, 0), (0, 2), (2, 0), (1, 2),  
    (2, 1), (1, 3), (2, 3), (3, 4), (5, 4),  
    (5, 6), (7, 5), (6, 8), (8, 7), (8, 9)]
```

Will Wang
will-ww

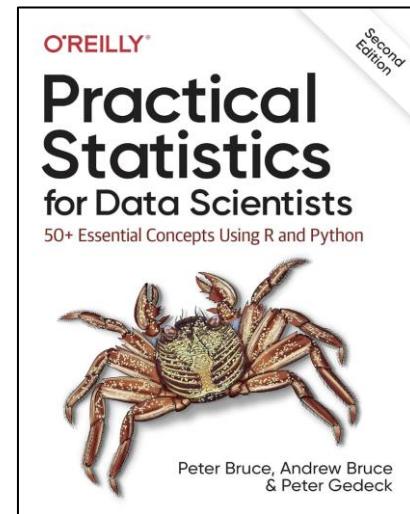
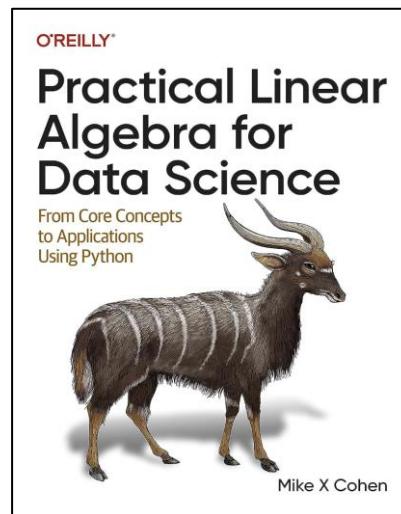
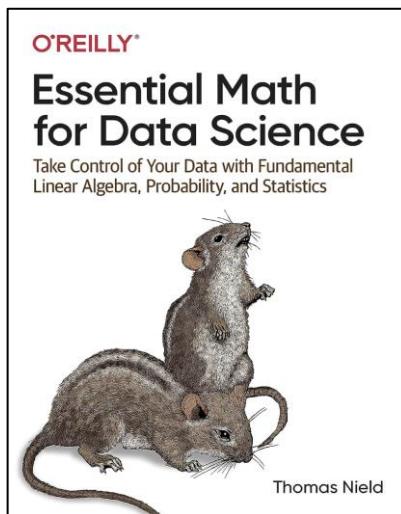
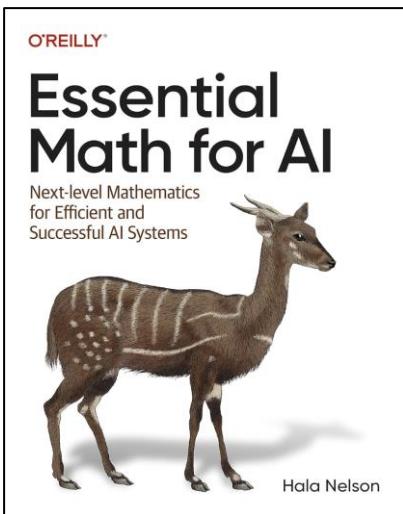
Will Wang/ECNU research professor/director of department of computer science education/founder of X-lab

Edit profile

153 followers - 113 following

ECNU/X-lab

扩展阅读



THANK YOU

