




金融科技导论

第二讲、金融大数据

金融科技导论@浙江大学 郑小林

2023年7月3日

金融大数据

- 01.金融大数据基本概念 
- 02.几类金融数据含义和来源
- 03.数据获取方式
- 04.数据预处理

- 存储技术的发展
- CPU处理能力的提高
- 网络带宽增加与云时代的开启



- 储存方法
- 检索效率
- 数据分析提取



传统的个人电脑：

GB/TB级别，硬盘通常512GB/1TB/2TB/4TB容量

大数据的级别

1 PB = 1024 TB (PB - petabyte)

1 EB = 1024 PB (EB - exabyte)



1TB：容量大约是20万张照片或20万首MP3音乐



1PB：一个人不停地听音乐，可以听1900年



1EB：21个标准篮球场

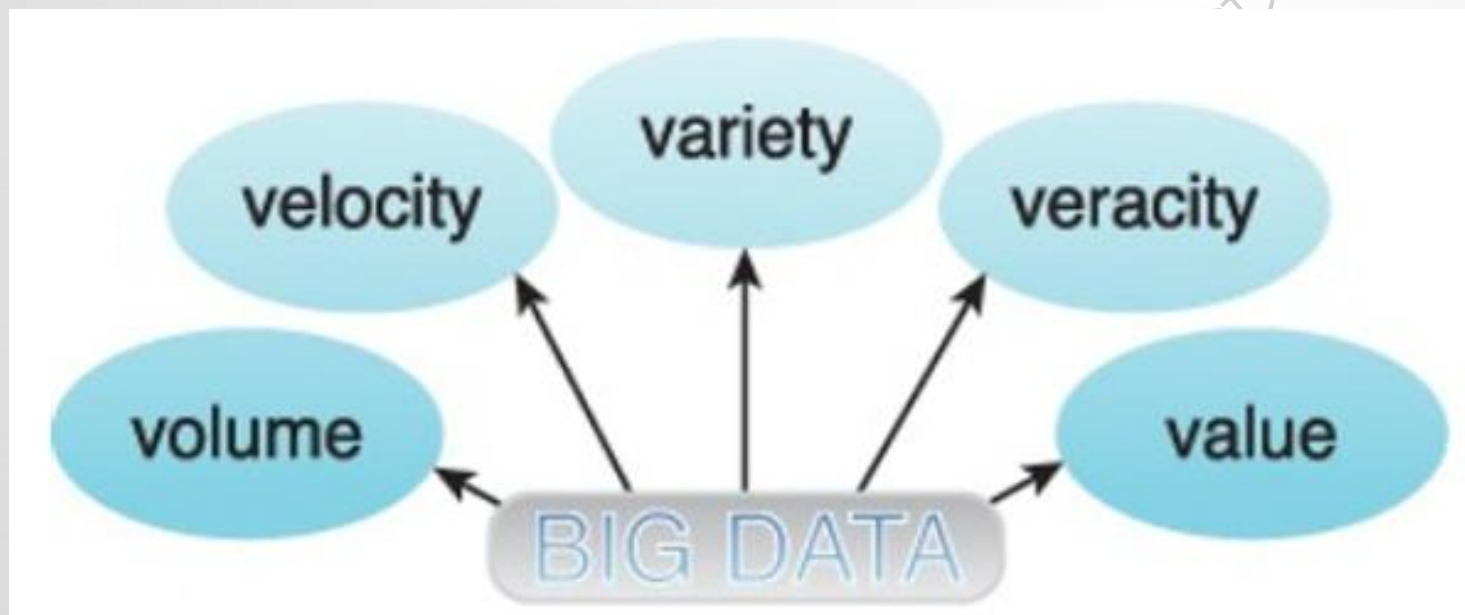


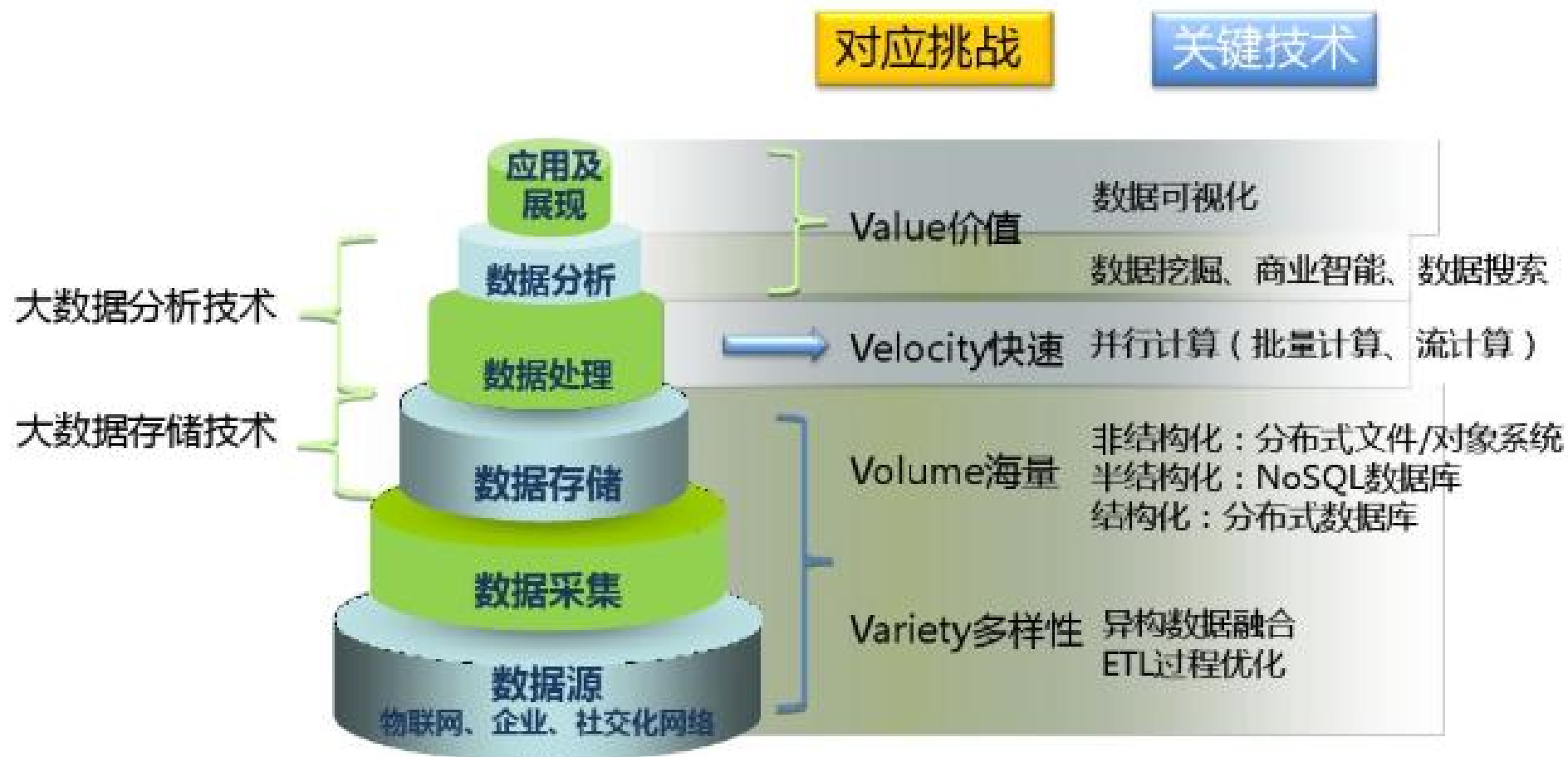
阿里巴巴数据中心



谷歌数据中心

- 数量 (Volume) —— 海量数据 (TB,PB,甚至 EB级别)
- 种类 (Variety) —— 大量不同数据源 (数字, 文本, 语音, 图像等)
- 速度 (Velocity) —— 实时处理
- 真实 (Veracity) —— 数据真实性
- 价值 (Value) —— 数据是否对业务有实际价值





- 分布式文件系统（如HDFS）
 - 高吞吐，方便扩展。
 - 有容错能力、通过相关协议保证数据一致性。
- NoSql 数据库（一些key-value数据库、列族数据库）
 - 支持规模存储、横向扩展。
 - 灵活定义存储格式。
- 云数据库
 - 购买数据库的SaaS服务。
 - 选择多样、随时调整。



金融大数据

- 01. 金融大数据基本概念
- 02. 金融数据含义和来源
- 03. 数据获取方式
- 04. 数据预处理



- 数据集由数据对象组成
- 一个数据对象代表一个实体
- 例如：
 - 销售数据库：顾客、商品、销售记录
 - 医疗数据库：患者、医生、药品
 - 大学数据库：学生、教授、课程
- 数据对象又称样本、实例、数据点、对象
- 数据库的行->数据对象；数据库的列->属性

- 属性（又称维度、特征、变量）是一个数据字段，表示数据对象的一个特征。
 - 例如，用户ID、姓名、地址
- 类型
- 标称属性 Nominal
- 二值属性 Binary
- 序数属性 Ordinal
- 数值属性 Numeric
 - 区间标度属性 Interval-scaled
 - 比率标度属性 Ratio-scaled

- 标称属性：类别、状态、编码，与名称相关
 - 头发颜色={黑色，棕色，红色，淡黄色，白色}
 - 婚姻状况，职业，身份证
- 二值属性：只有两个状态（0、1）的标称属性
 - 对称的：两种状态具有相同的价值
 - 非对称的：状态的结果是不是同样重要的
- 序数属性：值之间具有有意义的顺序，但是值之间的差是未知的
 - 例如，尺寸={大，中，小}、职位、级别

- 数值属性是定量的，整数或实数值表示
- 区间标度属性
 - 用相等的单位尺度度量
 - 属性的值有序
 - 例如，摄氏温度，日期
 - 没有真正的零点
- 比率标度属性
 - 具有固定的零点
 - 可以说一个值是另一个的倍数
 - 例如：开氏温度(K)、长度、数量

- 离散属性 Discrete
 - 具有有限或无限可数个值
 - 例如：职业、颜色、身份证
 - 离散属性有时可以用整数表示
 - 例如：二值属性取0和1，年龄属性取0到110
- 连续属性 Continuous
 - 如果属性不是离散的，则它是连续的
 - 用实数表示属性值
 - 一般用浮点数表示

- 数据质量的多维度评价指标：
 - 准确性 Accuracy: 正确或错误, 准确或不准确
 - 完整性 Completeness: 未记录的, 不可用的
 - 一致性 Consistency: 部分改变了, 但是另一些没变
 - 时效性 Timeliness: 及时的更新
 - 可信性 Believability: 数据是可信赖的吗
 - 可解释性 Interpretability: 数据是否容易理解

■ 相似性 Similarity

- 用数值评估两个数据对象间的相似程度
- 对象越相似，值越大
- 通常相似性值在范围 $[0,1]$

■ 相异性 Dissimilarity

- 用数值评估两个数据对象间的差异程度
- 对象越相似，值越小
- 通常没有上限

■ 相似性和相异性都成为邻近性 Proximity

■ 数据矩阵

- 用 $n \times p$ 的矩阵存放 n 个数据对象

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

■ 相异性矩阵

- 存放 n 个对象两两之间的临近度
- 三角矩阵

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

- 标称属性可以取两个或多个状态
- 两个对象 i 和 j 之间的相异性可以由不匹配率计算：

$$d(i, j) = \frac{p - m}{p}$$

- m 是匹配的数目， p 是刻画对象的属性总数

- 二值属性的列联表
contingency table

		Object j		
		1	0	sum
Object i	1	q	r	$q + r$
	0	s	t	$s + t$
	sum	$q + s$	$r + t$	p

- 对称的二元相异性

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

- 非对称的二元相异性

$$d(i, j) = \frac{r + s}{q + r + s}$$

- 非对称的二元相似性
Jaccard 系数

$$\text{sim}_{\text{Jaccard}}(i, j) = \frac{q}{q + r + s}$$

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- 性别是对称属性
- 其余属性是非对称二值属性

$$d(jack, mary) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$

$$d(jack, jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$$d(jim, mary) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$

- 闵可夫斯基距离 **Minkowski Distance**
$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$
- 其中, $i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ 和 $j = (x_{j1}, x_{j2}, \cdots, x_{jp})$ 是两个被 p 个数值属性描述的对象
- 数学性质:
 - 非负性: $d(i, j) > 0$ if $i \neq j$
 - 同一性: $d(i, i) = 0$
 - 对称性: $d(i, j) = d(j, i)$
 - 三角不等式: $d(i, j) \leq d(i, k) + d(k, j)$

- $h = 1$: 曼哈顿距离 Manhattan distance

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

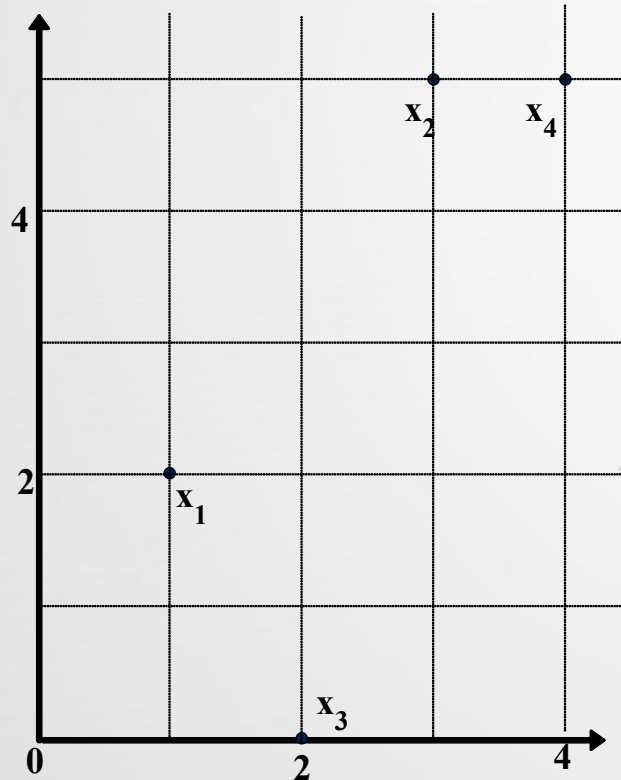
- $h = 2$: 欧几里得距离 Euclidean distance

$$d(i, j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)}$$

- $h \rightarrow \infty$: 上确界距离 “supremum” distance

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|$$

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Dissimilarity Matrices

Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

- 序数属性可以是离散的或连续的

$$r_{if} \in \{1, \dots, M_f\}$$

代替 x_{if}

- 用 x_{if} 的排位

- 由于每个序数属性可以有不同的状态数，通常需要将每个属性的值域映射到 $[0, 1]$ 上，用 z_{if} 代替第 i 个对象的 r_{if} ：

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- 之后，相异性可以用任意一种数值属性的距离度量计算

- 文档用数以千计的属性表示，每一个文档都被一个词频向量表示 **term-frequency vector**

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- 词频向量通常很长，并且是非常稀疏的，两个词频向量有很多公共的0，传统的度量会认为它们是不相似的，因此效果不好
- 我们需要一种度量，它只关注两个文档共有的词，以及这种词出现的频率，余弦相似度：

$$\cos(d_1, d_2) = (d_1 \cdot d_2) / \|d_1\| \|d_2\|$$

- Ex: Find the **similarity** between documents 1 and 2.

$$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$$

$$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$$


$$d_1 \bullet d_2 = 5*3 + 0*0 + 3*2 + 0*0 + 2*1 + 0*1 + 0*1 + 2*1 + 0*0 + 0*1 = 25$$

$$\begin{aligned} \|d_1\| &= (5*5 + 0*0 + 3*3 + 0*0 + 2*2 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} \\ &= (42)^{0.5} = 6.481 \end{aligned}$$

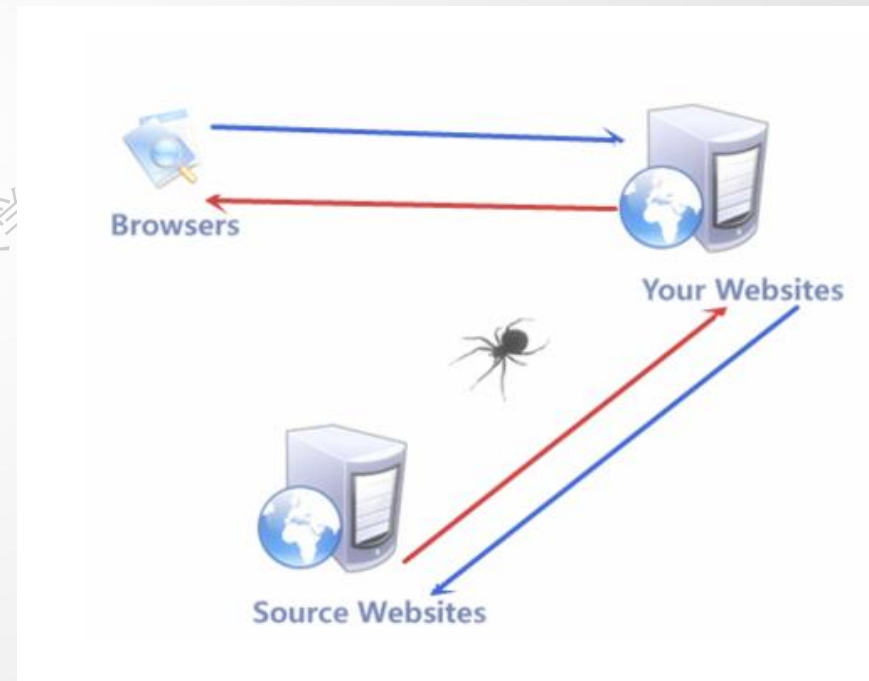
$$\begin{aligned} \|d_2\| &= (3*3 + 0*0 + 2*2 + 0*0 + 1*1 + 1*1 + 0*0 + 1*1 + 0*0 + 1*1)^{0.5} \\ &= (17)^{0.5} = 4.12 \end{aligned}$$

$$\cos(d_1, d_2) = 0.94$$

金融大数据

- 01. 金融大数据基本概念
- 02. 几类金融数据含义和来源
- 03. 数据获取方式 
- 04. 数据预处理

- 爬虫就是模拟客户端发送网络请求，接收请求响应，一种按照一定的规则，自动地抓取互联网信息的程序。
- 蓝色线条：发起请求（request）
- 红色线条：返回响应（response）



- 根据被爬网站的数量不同，我们把爬虫分为：
 - 通用爬虫：通常指搜索引擎的爬虫
 - 聚焦爬虫：针对特定网站的爬虫（主要）
- Robots 协议：网站通过 Robots 协议告诉搜索引擎哪些页面可以抓取，哪些页面不能抓取，但它仅仅是道德层面上的约束。

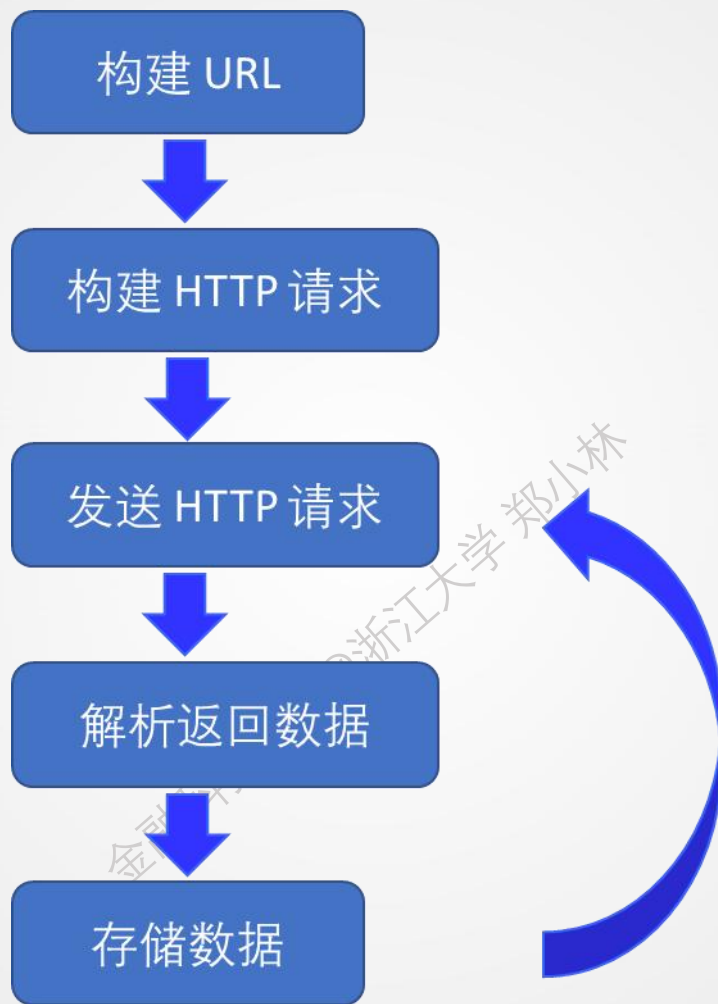
实例百度的Robots协议：

www.baidu.com/robots.txt

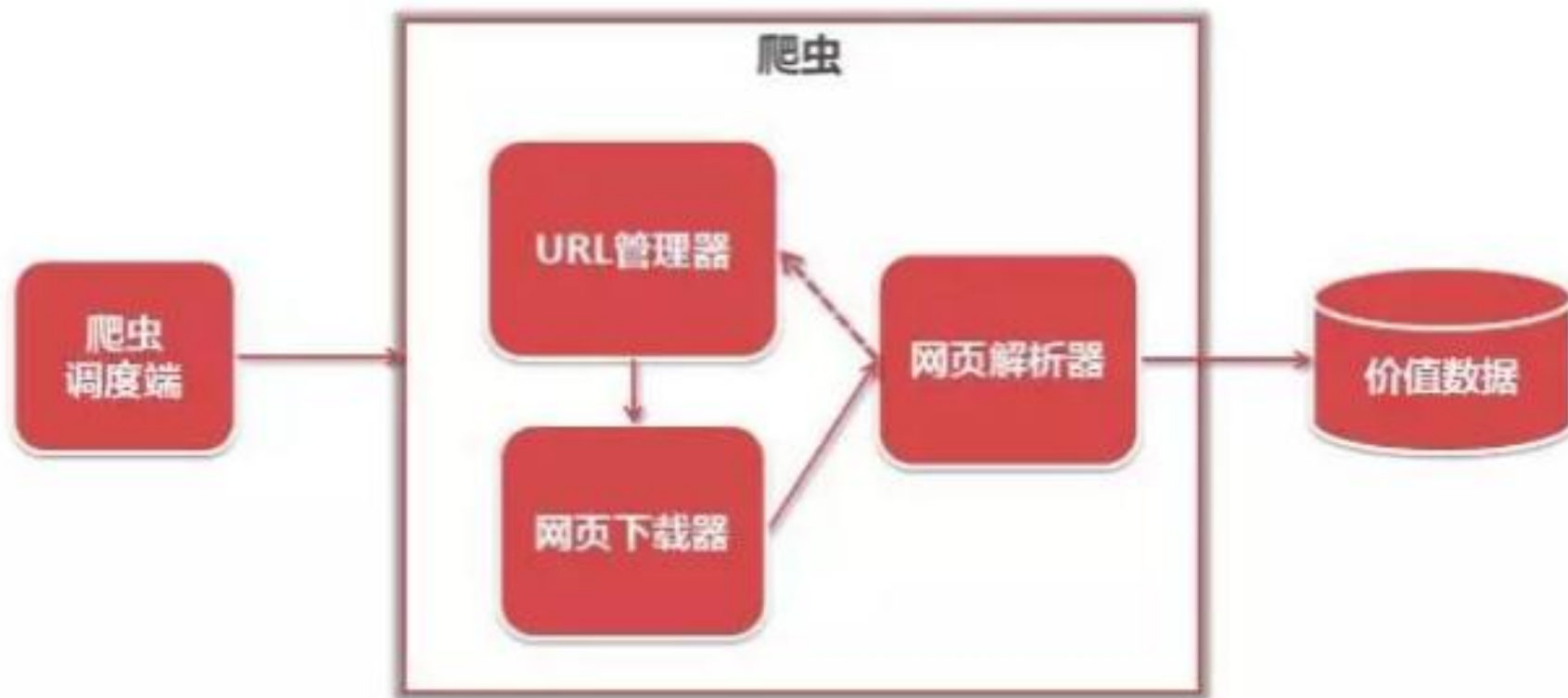
Robots协议可以通过“根域名+/robots.txt”查看。

```
User-agent: Baiduspider
Disallow: /baidu
Disallow: /s?
Disallow: /ulink?
Disallow: /link?
Disallow: /home/news/data/
Disallow: /bh
```

```
User-agent: Googlebot
Disallow: /baidu
Disallow: /s?
Disallow: /shifen/
Disallow: /homepage/
Disallow: /cpro
Disallow: /ulink?
Disallow: /link?
Disallow: /home/news/data/
Disallow: /bh
```



简单爬虫架构



- 根据《中华人民共和国[网络安全法](#)》的最新规定，以下行为可能构成违法犯罪：
 1. 爬虫程序规避网站经营者设置的反爬虫措施或者破解服务器防抓取措施，非法获取相关信息，情节严重的，有可能构成“非法获取计算机信息系统数据罪”。
 2. 爬虫程序干扰被访问的网站或系统正常运营，后果严重的，触犯刑法，构成“破坏计算机信息系统罪”。
 3. 爬虫采集的信息属于公民个人信息的，有可能构成非法获取公民个人信息的违法行为，情节严重的，有可能构成“侵犯公民个人信息罪”。

因此，在使用爬虫时，我们尽量坚持以下原则：

1. 遵守 Robots 协议。
2. 不能造成对方服务器瘫痪。
3. 不能非法获利。



金融大数据

- 01. 金融大数据基本概念
- 02. 几类金融数据含义和来源
- 03. 数据获取方式
- 04. 数据预处理



- 数据清理
 - 填充缺失值
 - 光滑噪声数据
 - 识别并删除离群点
 - 解决不一致性
- 数据集成
 - 集成多个数据库、数据立方体、数据文件等
- 数据规约
 - 维规约
 - 数值规约
 - 数据压缩
- 数据变换
 - 规范化
 - 离散化

数据预处理

● 数据预处理流程

- 数据清理
- 数据集成
- 数据规约
- 数据变化



金融科技导论@浙江大学 郑小林

- 现实世界的数据一般是不完整的、有噪声的、不一致的
 - 不完整：缺失属性值、缺少感兴趣的属性
 - 例如：职业= “ ”（缺失值）
 - 噪声：包含噪声、错误、离群点
 - 例如：薪水= “-10”（错误数据）
 - 不一致：两个属性的含义出现矛盾
 - 例如：年龄= “30”，生日= “01/01/2005”
 - 故意错误：被掩盖的缺失数据
 - 例如：为生日选择默认值（1月1日）

- 忽略元组：类标号 label 缺失时通常丢弃这条数据
- 人工填充：数据集很大或缺失值很多时，费时费力
- 自动填充：
 - 全局常量：将缺失值用同一个常量替换，例如：-1
 - 属性均值
 - 同一个类别的属性均值
 - 计算最可能的值：使用线性回归等方法推断属性值

- 分箱
 - 排序后的数据划分到若干个箱中（等深或等频的）
 - 箱中的每个值用箱中的中位数、均值、或边界值替换
- 回归
 - 用一个函数拟合数据来光滑数据
- 聚类
 - 聚类可以用来检测并去除离群点
- 结合人工
 - 算法检测可疑数据，人工检查这些数据并做删除

数据预处理

● 数据预处理流程

- 数据清理
- 数据集成
- 数据规约
- 数据变化



金融科技导论@浙江大学郑小

- 数据集成：合并来自多个数据存储的数据
- 模式集成：集成多个数据源的元数据（metadata）
- 实体识别问题：如何确定多个数据源中相同的实体
- 属性值冲突：
 - 同一对象在不同数据源中的某些属性不匹配
 - 属性的度量不同

- 集成多个数据源时经常会有数据的冗余
 - 同一个属性在不同数据源的命名不同
 - 一个属性可以由另一个推导出来
- 冗余属性可以用相关性分析检测
- 正确地集成多个数据源的数据，有助于减少结果数据集的冗余和不一致，可以提高后续过程的准确性和速度。

■ 卡方检验

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

■ χ^2 值越大，说明变量间相关性越强

■ 例子：

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

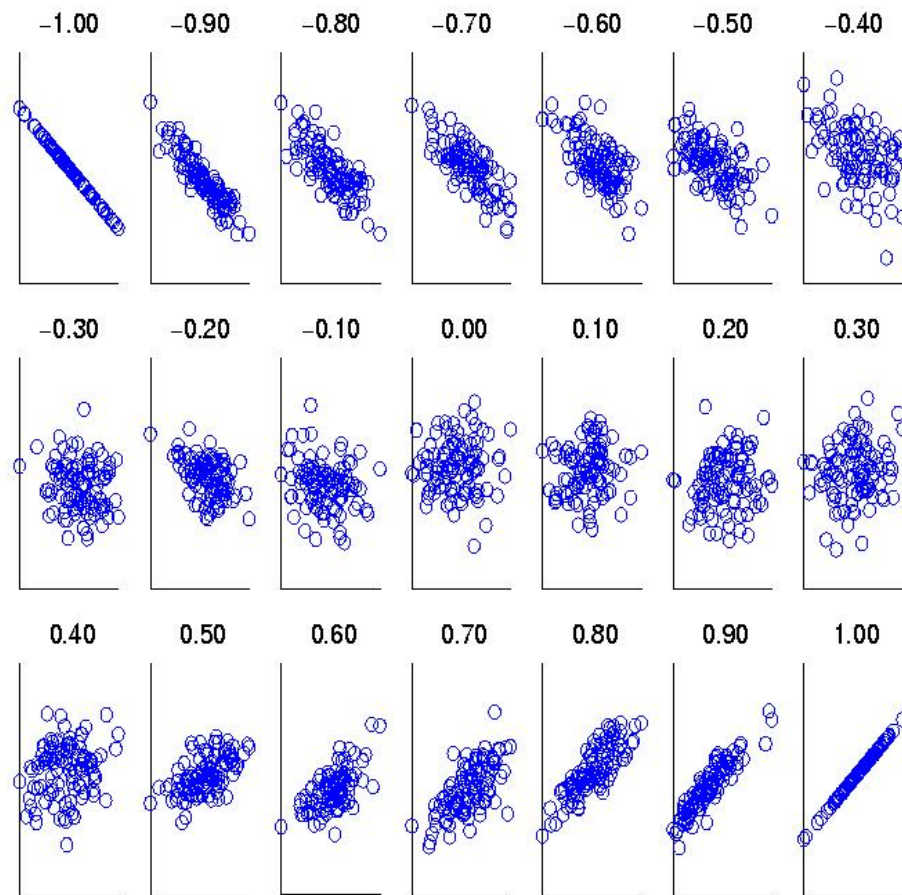
$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- 相关系数（Correlation coefficient）

$$r_{A,B} = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum_{i=1}^n (a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

- 其中， n 是元组的个数， \bar{A} 和 \bar{B} 是 A 和 B 的均值， σ_A 和 σ_B 分别是 A 和 B 的标准差
- 如果 $r_{A,B} > 0$ ，则 A 和 B 是正相关的，意味着 A 值随 B 值增加而增加，该值越大，相关性越强
- 如果 $r_{A,B} = 0$ ：不相关； $r_{A,B} < 0$ ：负相关

- 用散点图表示相关系数从 -1 到 $+1$



- 协方差（Covariance）与相关系数是类似的：

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n}$$

- 相关系数：

$$r_{A,B} = \frac{Cov(A, B)}{\sigma_A \sigma_B}$$

- 如果 $Cov_{A,B} > 0$ ，那么 A 和 B 倾向于同时大于它们的期望
- 如果 $Cov_{A,B} < 0$ ，那么当 A 大于期望时，B 很可能小于它的期望
- 如果 A 和 B 是独立的，那么 $Cov_{A,B} = 0$

反之不成立

- 协方差计算可以简化为

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

- 给定5个时间点两支股票的价格 (6, 20), (5, 10), (4, 14), (3, 5), (2, 5).

- $E(A) = (6 + 5 + 4 + 3 + 2) / 5 = 20/5 = 4$
- $E(B) = (20 + 10 + 14 + 5 + 5) / 5 = 54 / 5 = 10.8$
- $Cov(A, B) = (6 \times 20 + 5 \times 10 + 4 \times 14 + 3 \times 5 + 2 \times 5) / 5 - 4 \times 10.8 = 7$

- 由于协方差为正，因此可以说两个股票同时上涨

数据预处理

● 数据预处理流程

- 数据清理
- 数据集成
- 数据规约
- 数据变化



金融科技导论@浙江大学 郑小林

- 数据规约：得到数据集的更小的表示形式，仍保持数据的完整性，在规约数据上的分析仍可以得到相同的结果
- 数据规约策略
 - 维规约：减少随机变量和属性的个数
 - 小波变换
 - 主成分分析 PCA
 - 属性子集选择
 - 数值规约
 - 参数化数据规约
 - 直方图
 - 数据压缩

- 维度灾难
 - 随着特征维度的增加，数据在特征空间的密度是呈指数型下降
 - 数据变得越来越稀疏，数据间的距离和密度（分类、聚类）变得没有意义
- 维度规约
 - 避免维度灾难
 - 消除不相关属性和噪声
 - 减小时间和空间需求
 - 更容易进行可视化

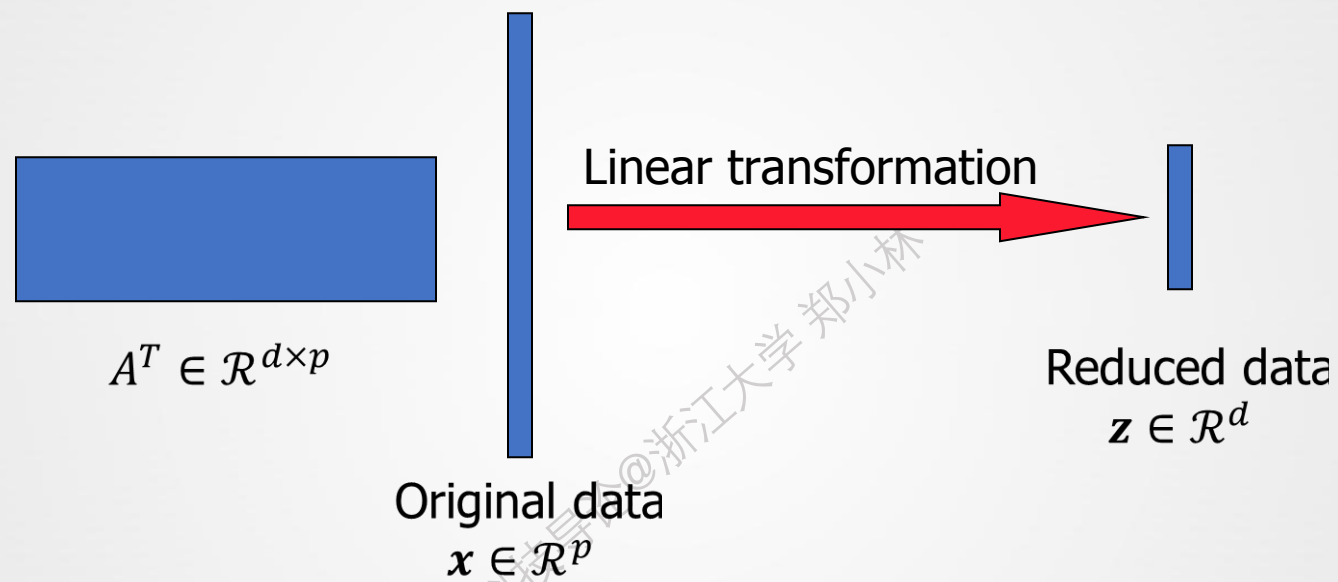
- 无监督
 - Latent Semantic Indexing (LSI): truncated SVD
 - **Principal Component Analysis (PCA)**
 - Independent Component Analysis (ICA)
 - Canonical Correlation Analysis (CCA)
- 有监督
 - **Linear Discriminant Analysis (LDA)**
- 半监督
 - Semi-supervised Discriminant Analysis (SDA)

■ 线性

- Latent Semantic Indexing (LSI): truncated SVD
- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Canonical Correlation Analysis (CCA)

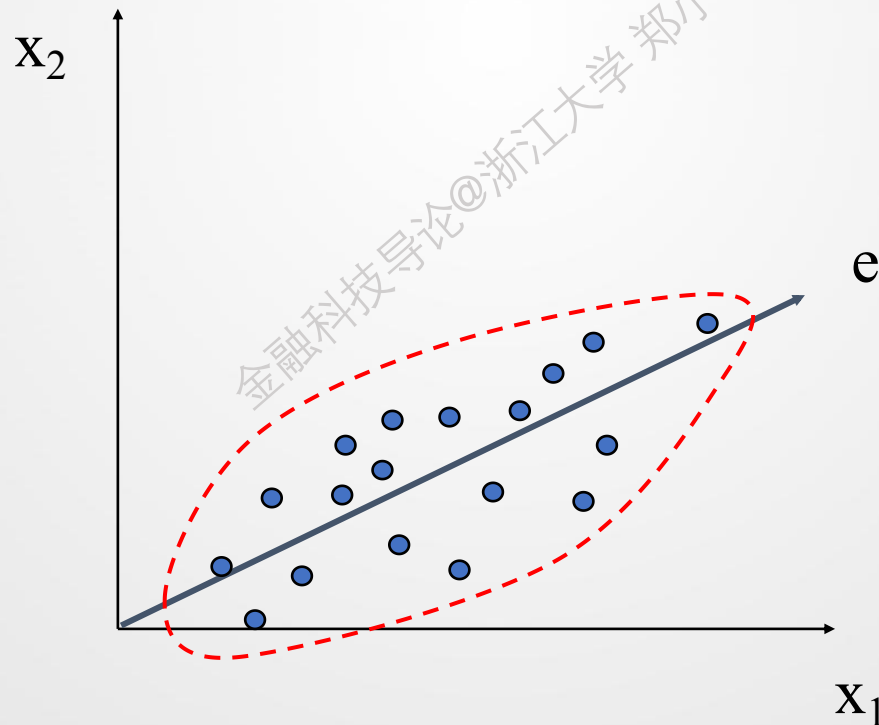
■ 非线性

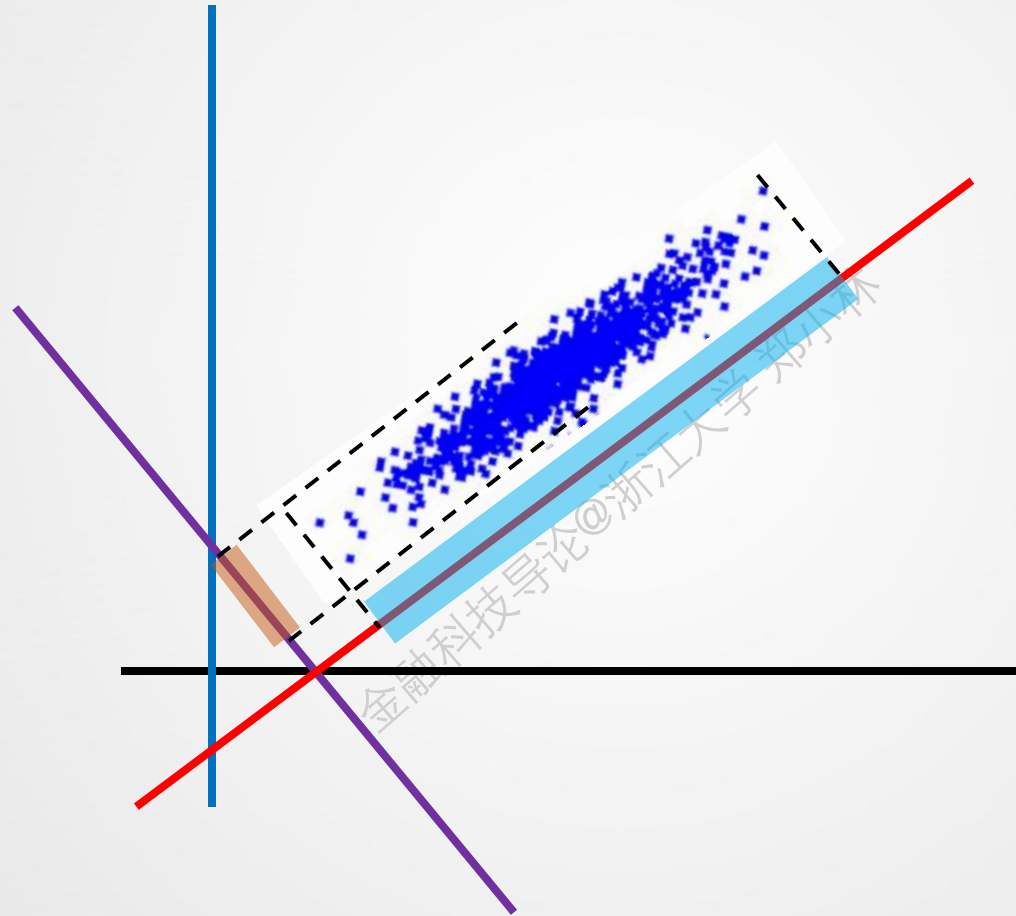
- Nonlinear feature reduction using kernels: 利用核函数进行非线性特征约简
- Manifold learning: 流形学习



$$A \in \mathcal{R}^{p \times d} : x \in \mathcal{R}^p \rightarrow z = A^T x \in \mathcal{R}^d$$

- 搜索 k 个最能代表数据的 n 维正交向量, $k \ll n$
- 找到使得投影后的数据方差最大的映射
- 原数据投影到一个小的多的空间上, 形成维规约
- 协方差矩阵的特征向量定义了新的空间





- 给定 n 个 p 维的样本观察值

$$\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \in \mathbb{R}^p$$

- 定义第一个主成分

$$z_i^{(1)} = \mathbf{a}_1^T \mathbf{x}_i, \quad i = 1, \dots, n$$

- 目标：使得，在这个空间的投影的方差 $\text{var}(z^{(1)})$ 最大

$$\begin{aligned} \text{var}(z^{(1)}) &= E((z^{(1)} - \bar{z}^{(1)})^2) \\ &= \frac{1}{n} \sum_{i=1}^n (\mathbf{a}_1^T \mathbf{x}_i - \mathbf{a}_1^T \bar{\mathbf{x}})^2 \\ &= \frac{1}{n} \sum_{i=1}^n \mathbf{a}_1^T (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{a}_1 = \mathbf{a}_1^T S \mathbf{a}_1 \end{aligned}$$

其中, $S = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T$ 是协方差矩阵

$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ 是平均数

- 通常来说:

$$\text{var}(z^{(k)}) = \mathbf{a}_k^T \mathbf{S} \mathbf{a}_k = \lambda_k$$

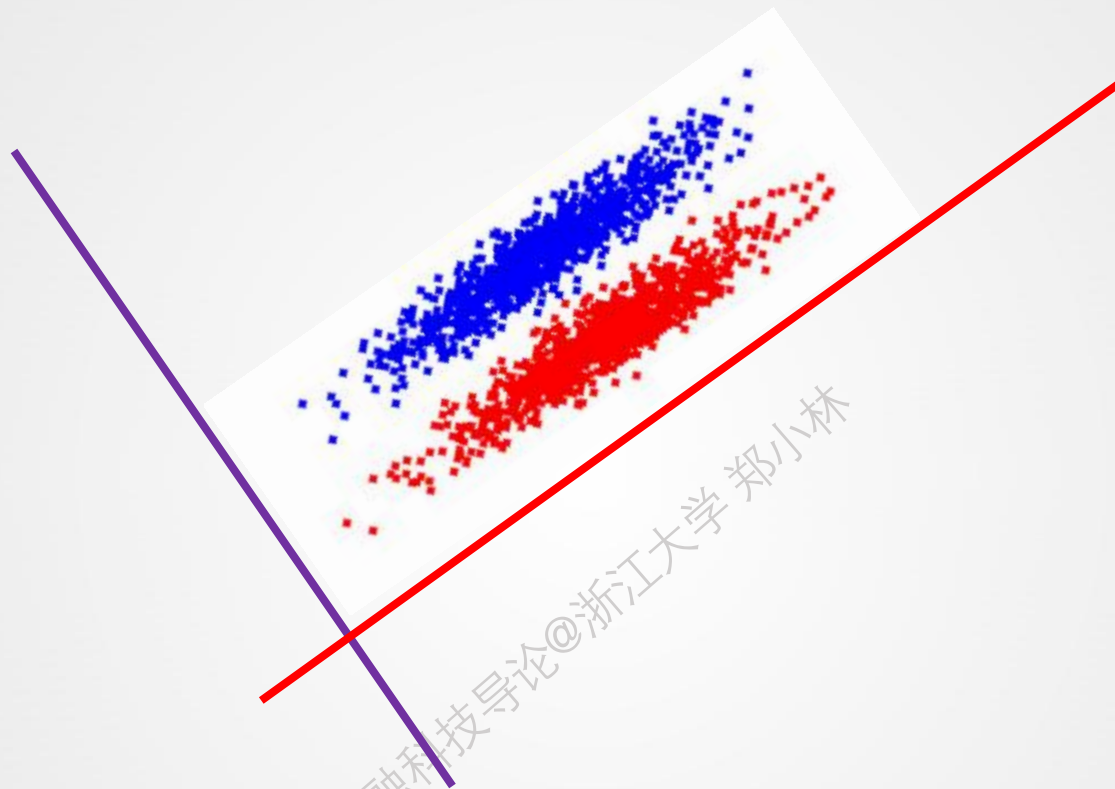
- S 的第 k 大的特征值对应于第 k 个主成分的方差
- 第 k 个主成分 $z^{(k)}$ 保留了样本中方差的第 k 大的影响

- 计算主成分的步骤：
 - 构造数据的协方差矩阵 S
 - 计算协方差矩阵的特征根: $\{\mathbf{a}_i\}_{i=1}^p$
 - 用前 d 个特征向量 $\{\mathbf{a}_i\}_{i=1}^d$ 组成 d 个主成分
 - 转换矩阵 A 定义为

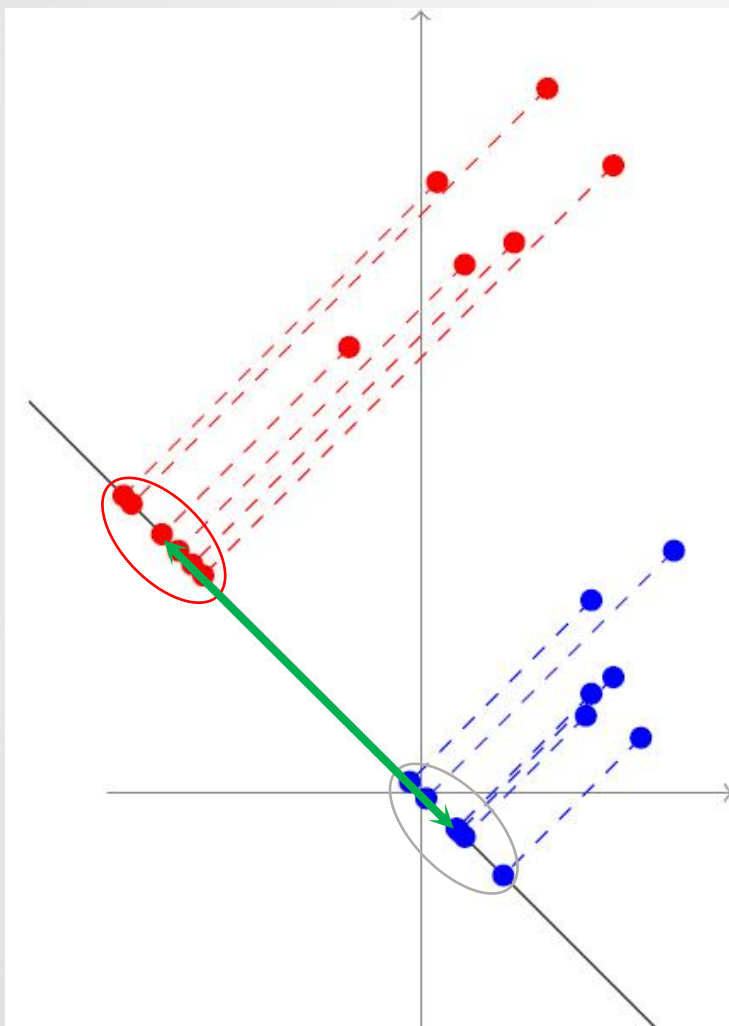
$$A = [\mathbf{a}_1, \dots, \mathbf{a}_d]$$

- 对于一个样本点:

$$\mathbf{x} \in \mathbb{R}^p \rightarrow A^T \mathbf{x} \in \mathbb{R}^d$$



- 尽可能保持类之前的差异性，寻找使得类别分离度最大的方向，类间距离最大，类内距离最小



■ 类内距离公式

$$S_w = \sum_{i=1}^C \sum_{j=1}^{M_i} (x_{ij} - \mu_i)(x_{ij} - \mu_i)^T$$

■ 类间距离公式

$$S_b = \sum_{i=1}^C (\mu_i - \mu)(\mu_i - \mu)^T$$

$$\mu = \frac{1}{C} \sum_{i=1}^C \mu_i$$

■ 判别式

$$J = \frac{|S_b|}{|S_w|}$$

- LDA 主要步骤:

- 构造 scatter 矩阵 S_B 和 S_W

- 计算广义特征值问题的**非零特征根对应的特征向量**

$$S_B \mathbf{a} = \lambda S_W \mathbf{a} \quad \text{or} \quad S_B \mathbf{a} = \lambda S_T \mathbf{a}$$

- 对特征向量进行排列, 最终选择一定数量的特征向量, 得到转移矩阵 A

$$A = [\mathbf{a}_1, \cdots \mathbf{a}_{c-1}]$$

- 对于一个样本点:

$$\mathbf{x} \in \mathbb{R}^p \rightarrow A^T \mathbf{x} \in \mathbb{R}^{(c-1)}$$

- 冗余属性
 - 重复包含其它特征中已经包含的信息
 - 例如，一个商品的价钱和购物发票总额
- 无关属性
 - 不包含对后续任务有用的信息
 - 例如，学号信息对预测学生的成绩没有意义

- 对于 n 个属性，有 2^n 种可能的属性组合，无法通过穷举找出最优属性子集
- 通常使用压缩搜索空间的启发式算法：
 - **逐步向前选择**：由空属性集开始，每次确定剩余属性集中最好的属性加入到该集合中。
 - **逐步向后删除**：由整个属性集开始，每次删除尚在属性集中的最差属性。
 - **逐步向前和向后的组合**：每次选择一个最好的属性，并在剩余属性中删除一个最差的属性。
 - **决策树**：决策树的每个节点选择最好的属性，最后不出现在树中的属性是不相关的

数据预处理

● 数据预处理流程

- 数据清理
- 数据集成
- 数据规约
- 数据变化



金融科技导论@浙江大学 郑小林

- 使用一个函数将一个给定的特征的所有值映射为一个新的值的集合，使得每一个旧值可以对应应用一个新值表示。
- 变换方法：
 - 光滑：去掉数据中的噪声。
 - 特征构造：由给定的特征构造新的特征并添加到属性集中。
 - 聚集：对数据进行汇总，例如汇集日销量数据计算月销量。
 - 规范化：把属性按比例缩放，使之落入一个特定的小区间。
 - 最小-最大规范化
 - Z-score 规范化
 - 小数定标规范化
 - 离散化：数值属性的原始值用区间标签或概念标签替换。

- 最小-最大规范化：映射到区间 $[new_min_A, new_max_A]$

$$v' = \frac{v - min_A}{max_A - min_A} (new_max_A - new_min_A) + new_min_A$$

- Z-score 规范化 (μ : 均值, σ : 标准差):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

- 小数定标规范化:

$$v' = \frac{v}{10^j}$$

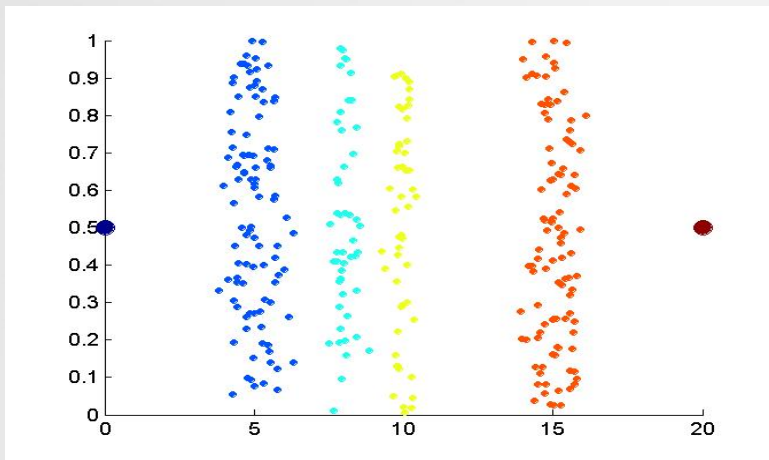
- j 是使得 $\text{Max}(|v'|) < 1$ 的最小整数
- 比如属性A的取值范围是-999到88，那么最大绝对值为999，小数点就会移动3位，即新数值=原数值/1000，那么A的取值范围就被规范为-0.999到0.088。

- 离散化：将连续属性的范围划分为区间
 - 方式：区间标签可以用于替换真实数据值
 - 优势：离散化可以减小数据大小
 - 类型：
 - ◆ 有监督 vs. 无监督
 - ◆ 分裂（自顶向下） vs. 合并（自底向上）
 - ◆ 可以使用递归的方法进行离散化

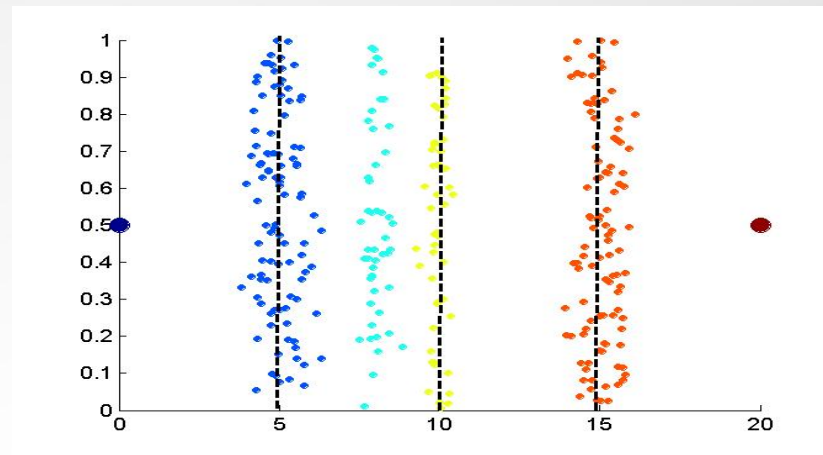
- 分箱 Binning: 自顶向下, 无监督
- 直方图 Histogram: 自顶向下, 无监督
- 聚类 Clustering: 自顶向下或**自底向上**, 无监督
- 决策树 Decision-tree: 自顶向下, **有监督**
- 相关系数 Correlation: **自底向上**, 无监督

- 等宽（等距）分箱
 - 将原始范围划分为相等大小的 N 个区间
 - 如果 A 和 B 是范围的最小最大值，那么划分区间的宽度为： $W = (B - A) / N$
 - 优势：方法直接
 - 缺陷：对离群点造成影响，倾斜的数据无法处理
- 等深（等频）分箱
 - 将范围划分为 N 个区间，使得每个区间的样本量近似相同，
 - 类别属性不容易划分

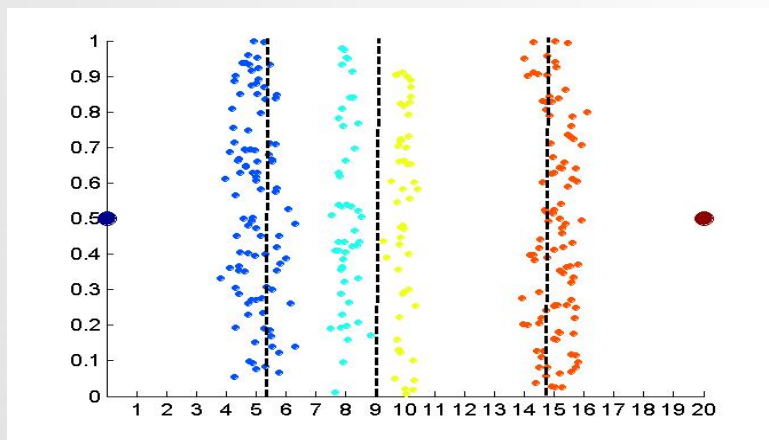
- 排序好的价格数据: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- 等深分箱:
 - Bin 1: 4, 8, 9, 15
 - Bin 2: 21, 21, 24, 25
 - Bin 3: 26, 28, 29, 34
- 用箱均值光滑:
 - Bin 1: 9, 9, 9, 9
 - Bin 2: 23, 23, 23, 23
 - Bin 3: 29, 29, 29, 29
- 用箱边界光滑:
 - Bin 1: 4, 4, 4, 15
 - Bin 2: 21, 21, 25, 25
 - Bin 3: 26, 26, 26, 34



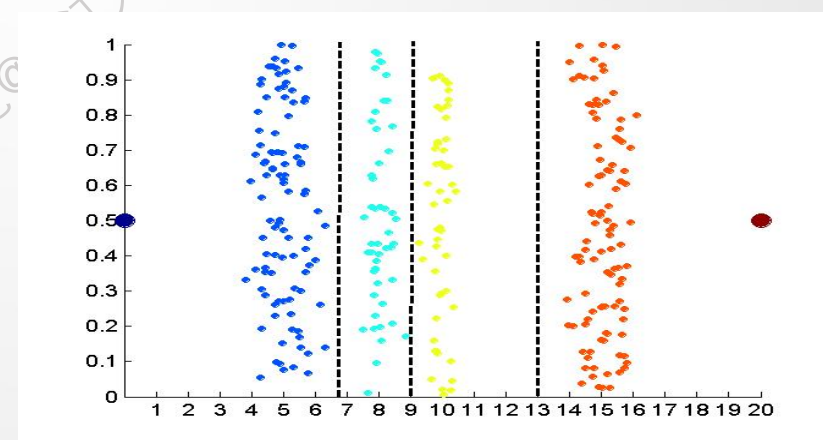
Data



Equal interval width (binning)



Equal frequency (binning)



K-means clustering leads to better results