



北京交通大学  
BEIJING JIAOTONG UNIVERSITY



# 《大数据概论》

## 大数据分析挖掘

鲍鹏  
软件学院





# 目录

---

- 数据理解与特征工程
  - 数据类型
  - 数据规范
  - 度量方法
  - 特征工程
- 常用数据挖掘算法
- 高级数据建模技术
- 数据可视化技术



# 数据类型

- 数据
  - 狭义概念：数字。
  - 广义概念：数据对象及其属性的集合，其表现形式可以为数字、符号、文字、图像或计算机代码等。
- 属性
  - 别称：特征、维或字段。
  - 通常指一个对象的某方面性质或特性。一个对象通过若干属性来刻画。
- 数据集
  - 数据对象的集合(同分布、同特征)。



# 数据类型

## 包含电信客户信息的样本数据集

属性

对象

客户编号	客户类别	行业大类	通话级别	通话总费用	...
N2201100 2518	大客户	采矿业和一般制造业	市话	16352	...
C1400483 9358	商业客户	批发和零售业	市话+国内长途(含国内IP)	27891	...
N2200489 5555	商业客户	批发和零售业	市话+国际长途(含国际IP)	63124	...
32210261 96	大客户	科学教育和文化卫生	市话+国际长途(含国际IP)	53057	...
D1400473 7444	大客户	房地产和建筑业	市话+国际长途(含国际IP)	80827	...
...	...	...	...	...	...



# 属性分类

属性类型		描述	例子	操作
分类的 (定性的)	标称	其属性值只提供足够的信息以区分对象。这种属性值没有实际意义。	颜色、性别、产品编号	众数、熵、列联相关
	序数	其属性值提供足够的信息以区分对象的序。	成绩等级(优、良、中、及格、不及格)、年级(一年级、二年级、三年级、四年级)	中值、百分位、秩相关、符号检验
数值的 (定量的)	区间	其属性值之间的差是有意义的。	日历日期、摄氏温度	均值、标准差、皮尔逊相关
	比率	其属性值之间的差和比率都是有意义的。	长度、时间和速度	几何平均、调和平均、百分比变差



# 数据集的特性

- 维度(Dimensionality)
  - 指数据集中的对象具有的**属性个数总和**。
- 稀疏性(Sparsity)
  - 指在某些数据集中，有意义的数据**非常少**，对象在大部分属性上的**取值为0**；非零项不到1%。
- 分辨率(Resolution)
  - 不同分辨率下数据的**性质不同**。
  - 举例：数米的分辨率下，地球表面看上去很不平坦，但在数十公里的分辨率下却相对平坦。



# 数据集的类型

- 记录数据
  - 每个记录包含固定的数据字段集（即属性集）。
  - 事务数据或购物篮数据、文本数据。
- 基于图的数据
  - 图捕获数据对象之间的联系。
  - 数据对象映射到图上的节点，对象之间的联系用链的方向和权值等链性质表示。
  - 万维网、化合物结构。
- 有序数据
  - 时序数据、序列数据、空间数据。



# 数据智能分析

- 数据智能分析是大数据应用中的一个重要环节，其目标是在对大数据进行预处理的基础上进行有效建模，并为具体的应用目标提供服务支撑。
- 当前大数据智能分析必须有效响应来自数据层的若干挑战：
  - 异构的数据格式
  - 异构的数据组织方式
  - 数据的时序性
  - 数据的交互性





# 数据规范

- 数据规范化
  - 定义：使不同规格的数据转换到同一规格。
- 数据归一化（最大-最小规范化）
  - 将数据映射到[0, 1]区间。
  - 目的：将特征数据进行伸缩变化，使得各特征对目标变量的影响一致，数据归一化会改变特征数据的分布。
  - 归一化计算公式：

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$



# 数据规范

- 数据标准化（z-score变换）

- 将特征数据变换为均值为0，方差为1的正态分布。
- 目的：使得不同特征之间具备可比性，经过标准化变换之后的特征数据分布没有发生改变。
- 应用场景：数据集的各个特征取值范围存在较大差异或各特征取值单位差异较大。
- 标准化的计算公式：

$$x^* = \frac{x - \mu}{\sigma}$$

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu)^2$$
$$\mu = \frac{1}{m} \sum_{i=1}^m x^{(i)}$$

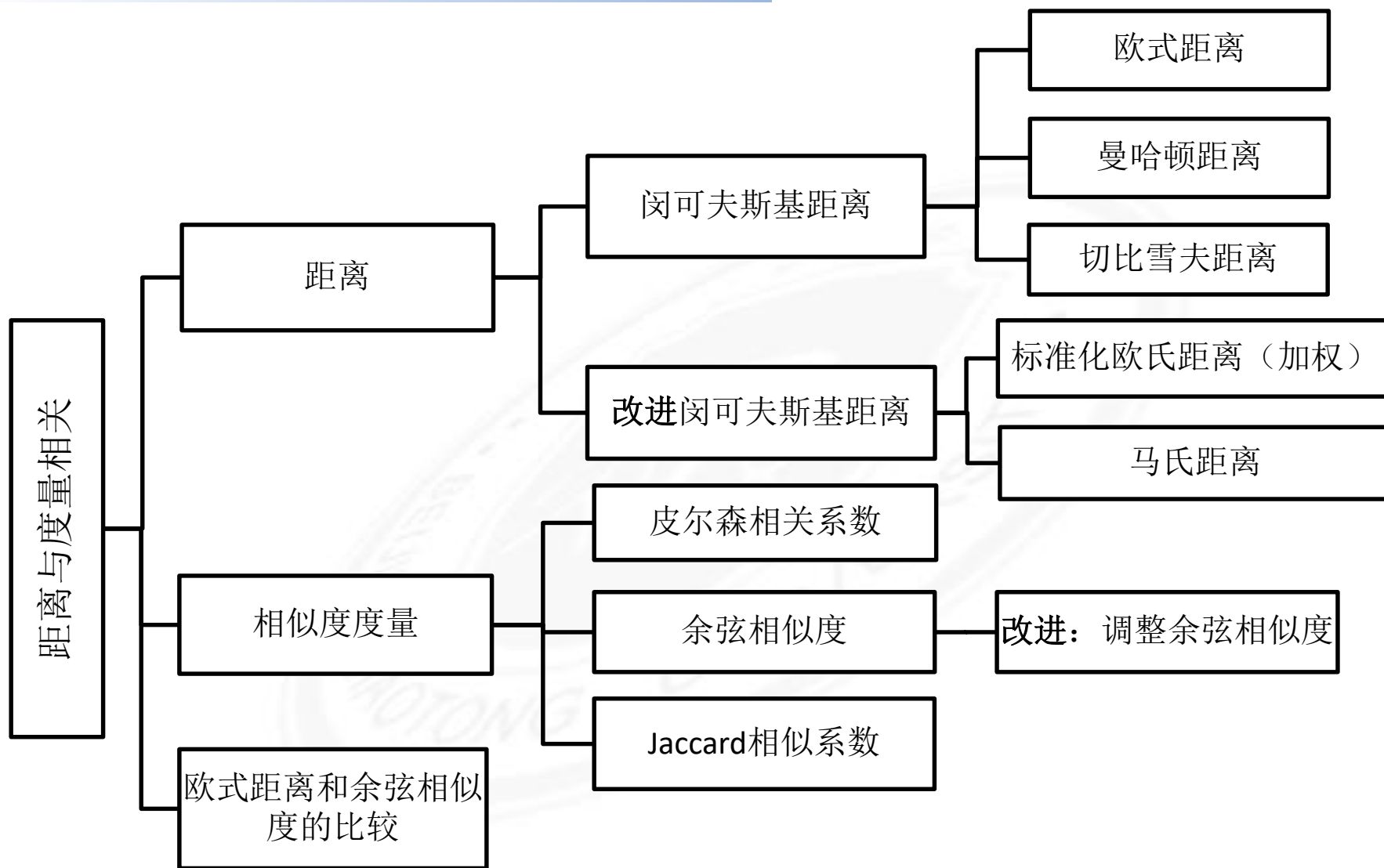


# 度量方法

- 在机器学习和数据挖掘中，常需衡量个体间差异的大小，进而评价个体的相似性和类别。
- 根据数据特性的不同，可以采用不同的度量方法。
  - 距离函数
  - 度量函数



# 度量方法





# 距离

- 定义一个距离函数  $d(x, y)$ , 需要满足下面几个基本准则:

- $d(x, x) = 0$  // 到自身的距离为0
- $d(x, y) \geq 0$  // 距离非负
- $d(x, y) = d(y, x)$  // 对称性: 若 A 到 B 距离为  $a$ , 则 B 到 A 的距离也为  $a$
- $d(x, k) + d(k, y) \geq d(x, y)$   
// 三角形法则: (两边之和大于第三边)



# 欧式距离

- 定义：所有点的对应维度之差的平方的求和再开方。
- 两个n维空间向量之间的欧式距离：
  - 向量  $a(x_{11}, x_{12}, \dots, x_{1n})$ ，向量  $b(x_{21}, x_{22}, \dots, x_{2n})$
  - $a$  与  $b$  的欧式距离：

$$d_{ab} = \sqrt{\sum_{k=1}^n (x_{1k} - x_{2k})^2}$$

- 计算原则
  - 各个维度指标在相同的刻度级别。
  - 若对身高、体重两个单位不同的指标使用欧氏距离可能使结果失效。

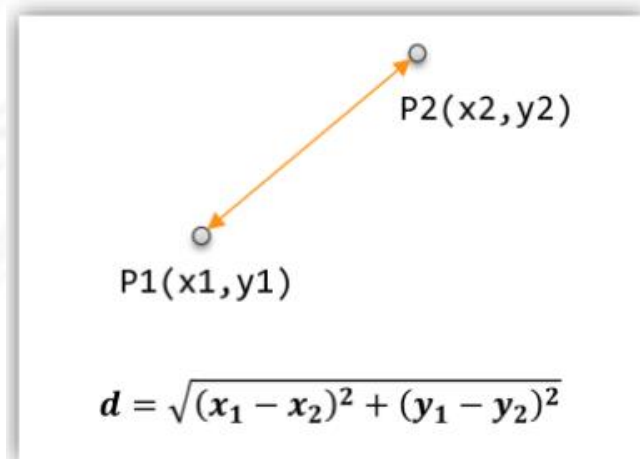


图. 二维空间中欧式距离的计算 linzhch3



# 曼哈顿距离

- 定义：将多个维度上的距离进行求和后的结果。
- 两个n维空间向量 $a$ 与 $b$ 之间的曼哈顿距离：

$$d_{ab} = \sum_{k=1}^n |x_{1k} - x_{2k}|$$

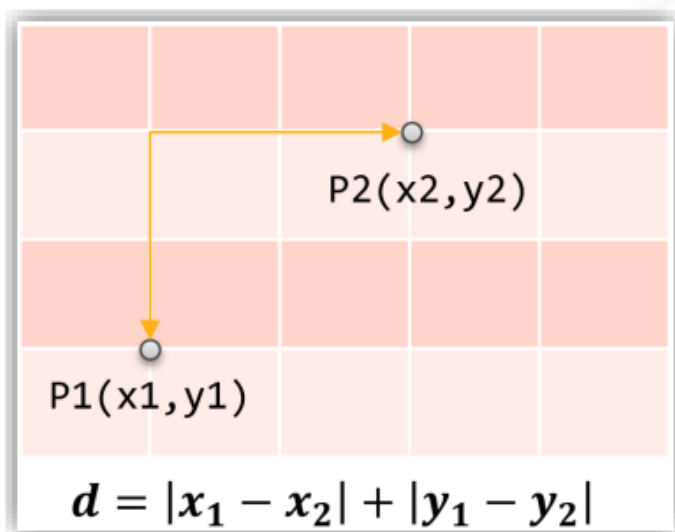
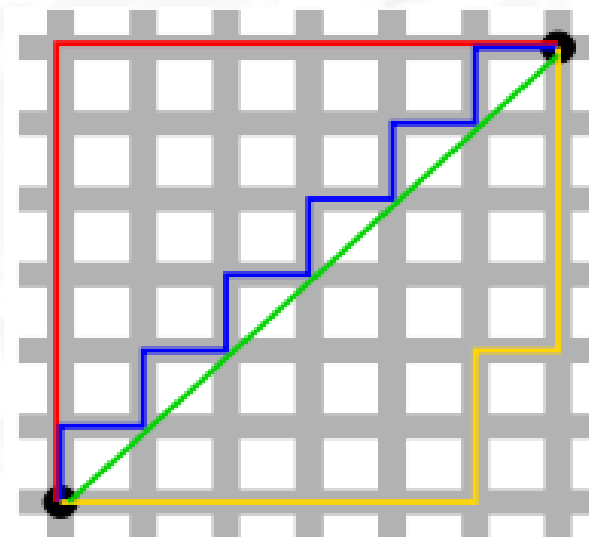


图. 二维空间中曼哈顿距离的计算



城市区块距离




# 切比雪夫距离

- 定义：所有点的各坐标数值差的最大值。
- 两个n维空间向量 $a$ 与 $b$ 之间的切比雪夫距离：

$$d_{ab} = \max_i (|x_{1i} - x_{2i}|)$$

- 棋盘距离

- 从一个位置走到其他位置需要的步数恰为二个位置的切比雪夫距离（Chebyshev distance），该距离也被称为棋盘距离。

	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1		1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	





# 相似度度量

- 相似度度量 (Similarity)
  - 概念：计算个体间的相似程度。
  - 与距离度量相反，相似度度量的值越小，表明个体间相似度越小，差异越大。
- 根据数据特性的不同，可采用不同的相似度度量方法。
  - 余弦相似度
  - 皮尔森相关系数
  - Jaccard相似系数(Jaccard Coefficient)



# 余弦相似度

- 两个向量越相似，向量夹角越小，余弦值的绝对值越大；值为负，两向量负相关。
- 应用：文本的相似度和推荐系统等。

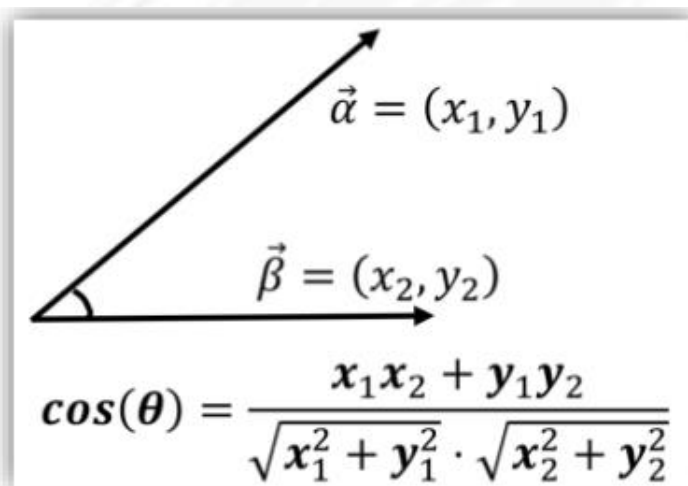
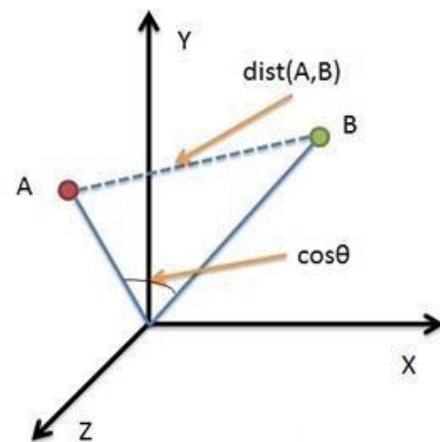

$$\vec{\alpha} = (x_1, y_1)$$
$$\vec{\beta} = (x_2, y_2)$$
$$\cos(\theta) = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} \cdot \sqrt{x_2^2 + y_2^2}}$$

图. 二维空间中的夹角余弦



# 欧式距离 VS 余弦相似度

- 欧氏距离从向量间的绝对距离区分差异，计算得到的相似度值对向量各个维度内的数值特征非常敏感，而余弦夹角从向量间方向夹角区分差异，对向量各个维度内的数值特征不敏感，同时修正了用户间可能存在的度量标准不统一的问题。
- 余弦夹角的值域区间为 $[-1, 1]$ ，相对于欧式距离的值域范围 $[0, +\infty]$ ，可对向量间的相似度值进行量化。





# 皮尔森相关系数

- 反映两个变量之间的线性相关程度。
- 定义：总体相关系数  $\rho$  为两个变量X、Y之间的协方差和两者标准差乘积的比值。

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

- 估算样本的协方差和标准差，可得到样本相关系数（即样本皮尔森相关系数），常用r表示。

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$



# 皮尔森相关系数

- Pearson相关系数是一个介于-1和1之间的值，用来描述两组线性数据一同变化移动的趋势。
  - 相关系数 $>0$ ，表明它们之间是正相关的。即当一个变量增大，另一个变量也增大；
  - 相关系数 $<0$ ，表明它们之间是负相关的，如果一个变量增大，另一个变量却减小；
  - 如果相关系数 $=0$ ，表明它们之间不存在线性相关关系。



# Jaccard相似系数

- Jaccard系数( Jaccard Coefficient)
  - 用于计算符号度量或布尔值度量的个体间的相似度。
  - 若个体的特征属性是由符号度量或者布尔值标识，只可得到“是否相同”的结果，无法衡量差异具体值的大小。Jaccard系数可衡量个体间共同具有的特征是否一致。
  - 若比较 X 与 Y 的Jaccard系数，只需比较  $x_n$  和  $y_n$  中相同的个数，公式如下：

$$Jaccard(X, Y) = \frac{X \cap Y}{X \cup Y}$$



# 特征工程

定义



将原始数据转变为模型的训练数据的过程

目的



获取更好的训练数据特征，使得数据挖掘模型逼近这个上限

作用



使模型性能得到提升  
在数据挖掘中占有非常重要的作用

构成



1. 特征表示
2. 特征提取
3. 特征选择



# 特征表示

- 特征表示，是将数据转换为有利于后续分析和处理的形式而进行的一种**形式化表示和描述**。
  - **不同类型数据使用不同特征表示方法**
  - 特征表示有利于后续的分析处理
  - 模型输出为**可计算向量**
  - 借鉴**专家知识**，能够提高特征表示质量
  - 对原始数据数字化后的特征表示可以**描述原始对象**





# 特征表示

- 当原始数据集中的特征的形式不适合直接进行建模时，使用一个或多个原特征构造新的特征可能会比直接使用原有特征更为有效。

特征构建：是指从原始数据中人工的找出一些具有物理意义的特征。

操作：使用混合属性或者组合属性来创建新的特征，或是分解或切分原有的特征来创建新的特征

方法：经验、属性分割



# 特征构建

- 聚合特征构造

- 聚合特征构造主要通过对多个特征的**分组聚合**实现，这些特征通常来自同一张表或者多张表。
- 聚合特征构造使用**一对多**的关联来对观测值分组，然后**计算统计量**。
- 常见的分组统计量有**中位数、算术平均数、众数、最小值、最大值、标准差、方差和频数**等



# 特征构建

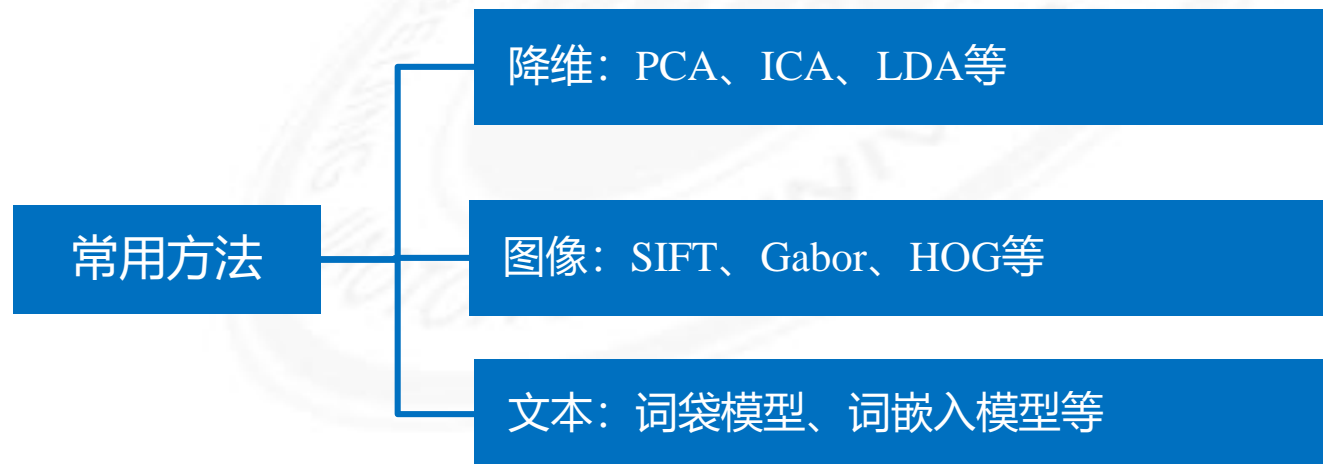
- 转换特征构造

- 相对于聚合特征构造依赖于多个特征的分组统计，转换特征构造通常依赖于特征本身的变换。
- 转换特征构造使用单一特征或多个特征进行变换后的结果作为新的特征。
- 常见的转换方法有单调转换（幂变换、log变换、绝对值等）、线性组合、多项式组合、比例、排名编码和异或值等。



# 特征提取

- 提取对象：原始数据（特征提取一般是在特征选择之前）
- 提取目的：自动地构建新的特征，将原始数据转换为一组具有明显物理意义（比如几何特征、纹理特征）或者统计意义的特征。





# 特征提取

- 降维方法—PCA

- 主成分分析法(Principal Component Analysis, PCA)。
- PCA 是降维最经典的方法，旨在找到数据中的主成分，并利用这些主成分来表征原始数据，从而达到降维的目的。
- 核心思想：通过坐标轴转换，寻找数据分布的最优子空间。





# 特征提取

- 降维方法—ICA

- 独立成分分析法 (Independent Component Analysis, ICA)
- ICA分析法可获得相互独立的属性。
- 核心思想：寻找一个线性变换  $z = Wx$ ，使得 $z$ 的各个特征分量之间独立性最大。

PCA 对数据  
进行降维



ICA 从多个  
维度分离出  
有用数据

PCA 是 ICA 的数据预处理方法



# 特征提取

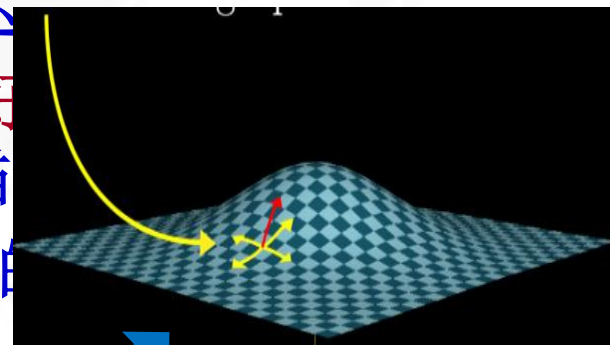
## • 图像特征提取—SIFT 特征

– 优点:

(1) 具有旋转、尺度、平移、视角及亮度不变性，有利于对目标特征信息进行有效表达

(2) SIFT 特征对参数调整鲁棒性好，通过调整适宜的特征点数量进行特征描述

– 缺点：不借助硬件加速或者专门的



疑似特征点检测

去除伪特征点

特征点梯度  
与方向匹配

特征描述向量的  
生成



# 特征提取

- 图像特征提取—HOG特征
  - 方向梯度直方图(Histogram of oriented gradient, HOG)特征是**2005** 年针对行人检测问题提出的直方图特征，它通过计算和统计图像局部区域的梯度方向直方图来实现特征描述。







# 特征提取

- 文本特征提取—词袋模型

- 将整段文本以词为单位切分开，每篇文章可表示为一个长向量，向量的每一个维度代表一个单词，而该维度的权重反映了该单词在原来文章中的重要程度。
- 采用 TF-IDF 计算权重，公式如下：

$$TF - IDF(t, d) = TF(t, d) \times IDF(t)$$

- $TF(t, d)$  表示单词  $t$  在文档  $d$  中出现的频率。
- $IDF(t)$  是逆文档频率，用来衡量单词  $t$  对表达语义的重要性，其表示为：

$$IDF(t) = \log \frac{\text{文章总数}}{\text{包含单词 } t \text{ 的文章总数} + 1}$$



# 特征提取

- 文本特征提取—N-gram 模型
  - 将连续出现的  $n$  个词 ( $n \leq N$ ) 组成的词组(N-gram)作为一个单独的特征放到向量表示, 构成了 N-gram 模型。
  - 同一个词可能会有多种词性变化, 但却具有相同含义, 所以实际应用中还会对单词进行词干抽取(Word Stemming)处理, 即将不同词性的单词统一为同一词干的形式。



# 特征选择

- 特征选择（feature selection）
  - 定义：从给定的特征集合中选出相关特征子集的过程。
  - 相关特征：对当前学习任务有用的属性或者特征。
  - 原因：维数灾难问题；去除无关特征可以降低学习任务的难度，简化模型，降低计算复杂度。
  - 目的：确保不丢失重要的特征。



# 特征选择

## 模型性能

- 保留尽可能多的特征，模型的性能会提升
- 但同时模型变复杂，计算复杂度也同样提升

VS

## 计算复杂度

- 剔除尽可能多的特征，模型的性能会有所下降
- 但模型变简单，计算复杂度随之降低



# 特征提取 VS 特征选择

项目	特征提取	特征选择
共同点	<ul style="list-style-type: none"><li>➤ 都从原始特征中找出最有效的特征</li><li>➤ 都能帮助减少特征的维度、数据冗余</li></ul>	
区别	<ul style="list-style-type: none"><li>➤ 强调通过<b>特征转换</b>的方式得到一组具有明显物理或统计意义的特征</li><li>➤ 有时能发现更有意义的特征属性</li></ul>	<ul style="list-style-type: none"><li>➤ 从特征集合中<b>挑选</b>一组具有明显物理或统计意义的特征子集</li><li>➤ 能表示出每个特征对于模型构建的重要性</li></ul>