



北京交通大学
BEIJING JIAOTONG UNIVERSITY



数据理解与特征工程





目录

01 数据类型

02 数据规范

03 度量方法

04 特征工程



1. 数据类型

- 相关概念

- 数据

- 狭义：数字。
 - 广义：数据对象及其属性的集合，其表现形式可以是数字、符号、文字、图像抑或是计算机代码等等。

- 属性

- (也称为特征、维或字段)，是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)



包含电信客户信息的样本数据集

对象

属性

| 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。
 - 维归约
- **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%。
 - 文本数据集
- **分辨率(Resolution)**
 - 不同分辨率下数据的性质不同



数据集的类型

- 数据集的类别
 - 记录数据
 - 事务数据或购物篮数据
 - 数据矩阵
 - 文本数据
 - 基于图形的数据
 - 万维网
 - 化合物结构
 - 有序数据
 - 时序数据
 - 序列数据
 - 时间序列数据
 - 空间数据
 - 流数据



数据的若干特征和挑战

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特征和挑战：

- 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



2. 数据规范

数据规范化 使不同规格的数据转换到同一规格。

归一化（最大 - 最小规范化）

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

将数据映射到[0,1]区间

数据归一化的目的是使得各特征对目标变量的影响一致，会将特征数据进行伸缩变化，所以数据归一化是会改变特征数据分布的。



2. 数据规范

Z-Score标准化

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

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

处理后的数据均值为0，方差为1

数据标准化为了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。当数据特征取值范围或单位差异较大时，最好是做一下标准化处理。

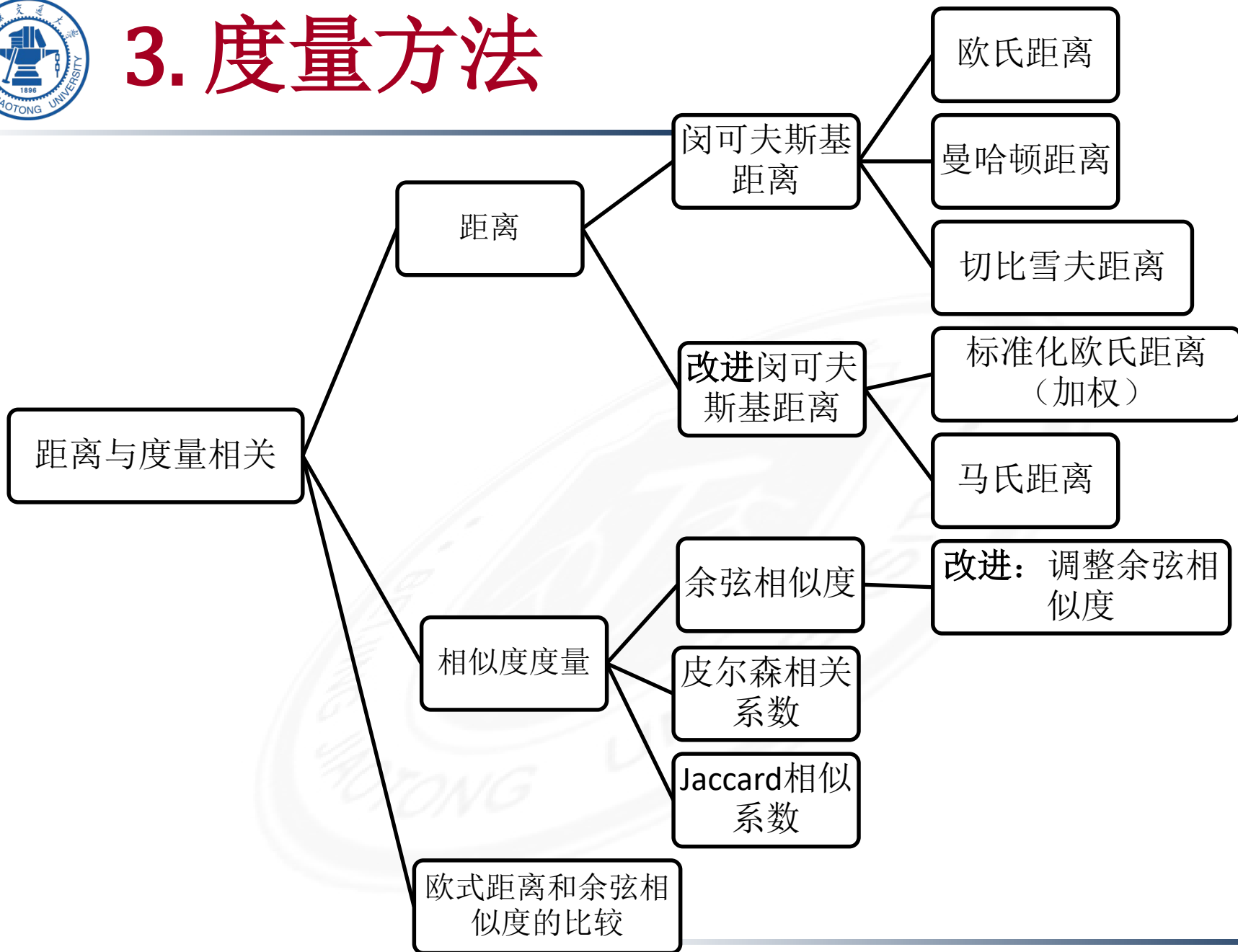


3. 度量方法

- 在机器学习和数据挖掘中，我们经常需要知道个体间差异的大小，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



3. 度量方法





距离

一般而言，定义一个距离函数 $d(x,y)$, 需要满足下面几个基本准则：

- 1) $d(x,x) = 0$ // 到自己的距离为0
- 2) $d(x,y) \geq 0$ // 距离非负
- 3) $d(x,y) = d(y,x)$ // 对称性: 如果 A 到 B 距离是 a , 那么 B 到 A 的距离也应该是 a
- 4) $d(x,k) + d(k,y) \geq d(x,y)$
 // 三角形法则: (两边之和 大于第三边)



欧式距离

- n 维空间点 $a(x_{11}, x_{12}, \dots, x_{1n})$ 与 $b(x_{21}, x_{22}, \dots, x_{2n})$ 间的欧氏距离（两个 n 维向量）：

$$d_{12} = \sqrt{\sum_{k=1}^n (x_{1k} - x_{2k})^2}$$

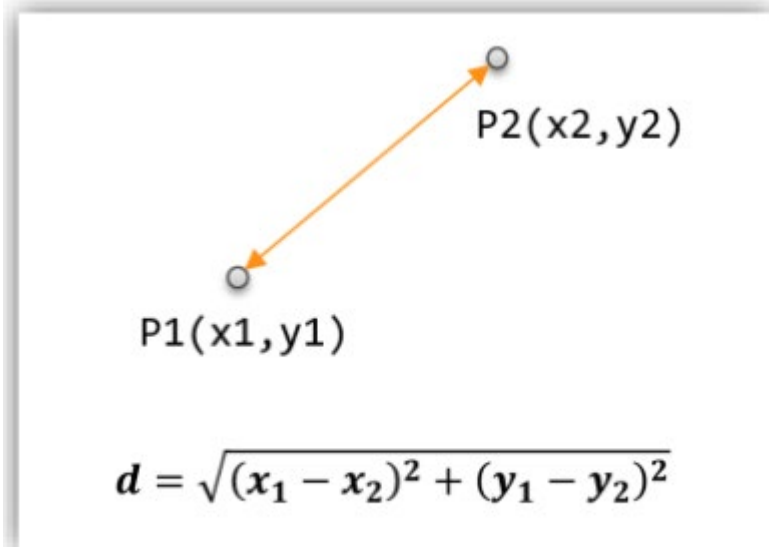


图. 二维空间中欧式距离的计算 linzh3

即：所有点的对应维度之差的平方的求和再开方。

欧式距离相似度算法需要保证各个维度指标在相同的刻度级别，比如对身高、体重两个单位不同的指标使用欧氏距离可能使结果失效。



曼哈顿距离

- n 维空间点 $a(x_{11}, x_{12}, \dots, x_{1n})$ 与 $b(x_{21}, x_{22}, \dots, x_{2n})$ 的曼哈顿距离：

$$d_{12} = \sum_{k=1}^n |x_{1k} - x_{2k}|$$

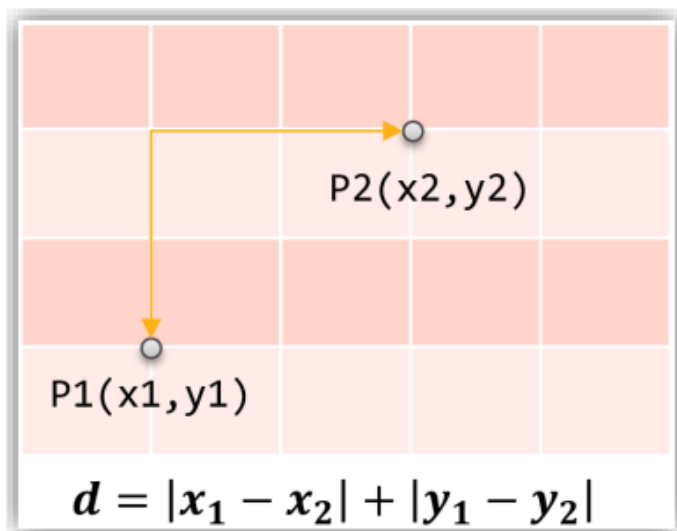
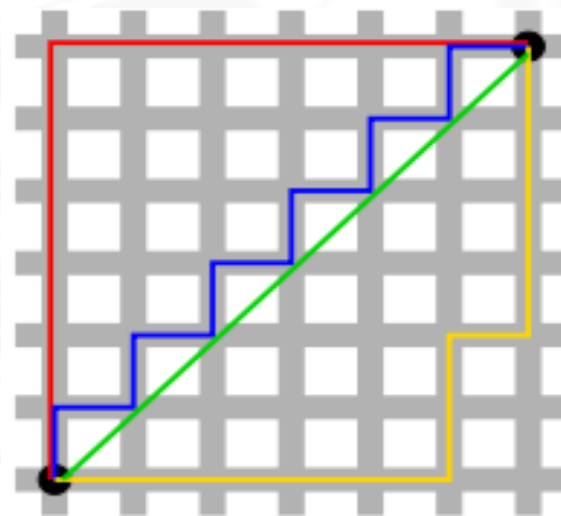


图. 二维空间中曼哈顿距离的计算



曼哈顿距离来源于城市区块距离，是将多个维度上的距离进行求和后的结果



切比雪夫距离

n维空间点a(x₁₁,x₁₂,...,x_{1n})与b(x₂₁,x₂₂,...,x_{2n})的切比雪夫距离：

$$d_{12} = \max_i (|x_{1i} - x_{2i}|)$$

| | 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 | |

切比雪夫距离（Chebyshev distance）是向量空间中的一种度量，二个点之间的距离定义为其各坐标数值差的最大值。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，因此切比雪夫距离也称为棋盘距离。



马式距离

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^T S^{-1} (x_i - x_j)}$$

若协方差矩阵是单位矩阵（各个样本向量之间独立同分布），
则公式就成了： $D(X_i, X_j) = \sqrt{(X_i - X_j)^T (X_i - X_j)}$

则 X_i 与 X_j 之间的马氏距离等于他们的欧氏距离。

即：若协方差矩阵是对角矩阵，公式变成了标准化欧氏距离。

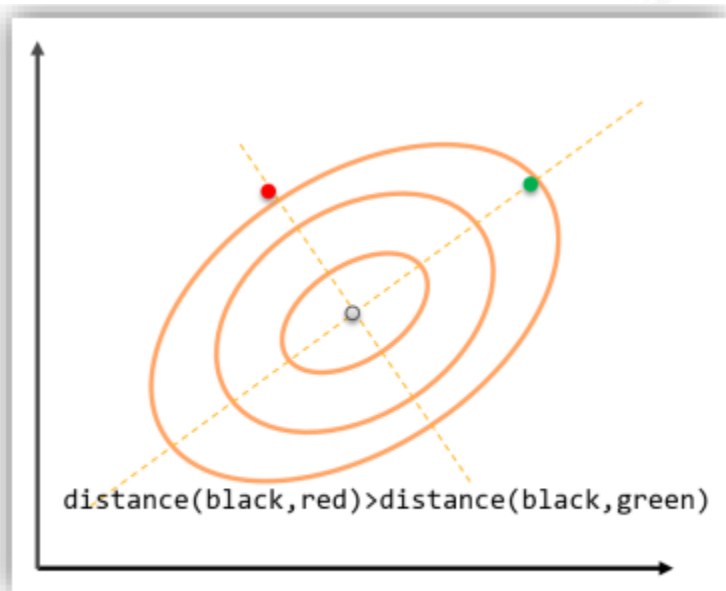


图. 二维空间中的马氏距离

标准化欧氏距离是在假设数据各个维度不相关的情况下，利用数据分布的特性计算出不同的距离。如果维度相互之间数据相关（例如：身高较高的信息很有可能会带来体重较重的信息，因为两者是有关联的），就要用到马氏距离



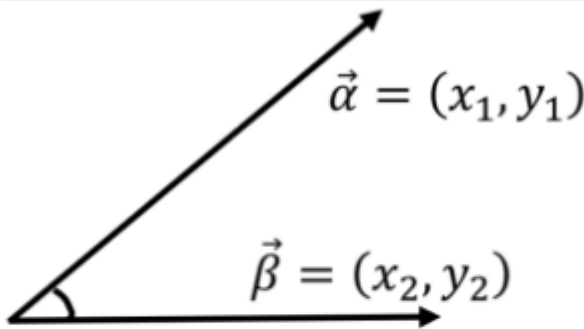
相似度度量

- 相似度度量（Similarity），即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，说明个体间相似度越小，差异越大。
 - 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。



The diagram shows two vectors, $\vec{\alpha} = (x_1, y_1)$ and $\vec{\beta} = (x_2, y_2)$, originating from the same point. Vector $\vec{\beta}$ is horizontal, and vector $\vec{\alpha}$ is at an angle θ to it. The angle is marked with an arc.

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

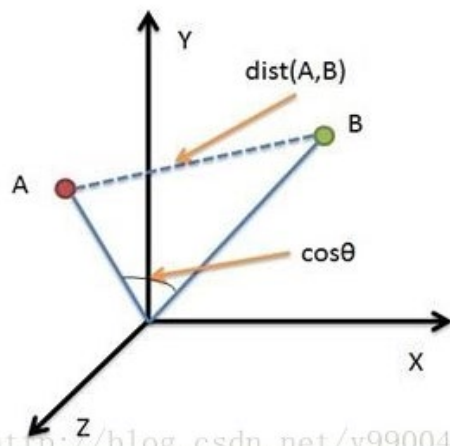
图. 二维空间中的夹角余弦



欧式距离和余弦相似度

(1) 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的**方向夹角**区分差异，对向量各个维度内的**数值特征不敏感**，所以同时修正了用户间可能存在的度量标准不统一的问题。

(2) 余弦夹角的值域区间为 $[-1,1]$ ，相对于欧式距离的值域范围 $[0, \text{正无穷大}]$ ，能够很好的对向量间的相似度值进行了量化。





皮尔森相关系数

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

上面是总体相关系数，估算样本的协方差和标准差，可得到样本相关系数(样本皮尔逊系数)，常用英文小写字母 r 代表：

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



皮尔森相关系数

- pearson是一个介于-1和1之间的值，用来描述两组线性的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 Coefficient)

- Jaccard系数主要用于计算符号度量或布尔值度量的个体间的相似度，因为个体的特征属性都是由符号度量或者布尔值标识，因此无法衡量差异具体值的大小，只能获得“是否相同”这个结果，所以Jaccard系数只关心个体间共同具有的特征是否一致这个问题。如果比较X与Y的Jaccard相似系数，只比较 x_n 和 y_n 中相同的个数，公式如下：

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



4. 特征工程

特征工程相关概念

定义



是把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



- 使模型的性能得到提升
- 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，特征表示无歧义表示
 - 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

在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、属性分割和结合



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的联立。
- 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，通常依赖于对于特征本身的变换。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。



特征构建

转换特征构造

此外，由于业务的需求，一些指标特征也需要基于业务理解进行特征构造。

- 基于单价和销售量计算销售额.
- 基于原价和售价计算利润.
- 基于不同月份的销售额计算环比或同比销售额增长/下降率.
-



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。

常用方法

降维方面的PCA、ICA、LDA等

图像方面的SIFT、Gabor、HOG等

文本方面的词袋模型、词嵌入模型等



特征提取

降维

1.PCA(Principal Component Analysis, 主成分分析)

PCA 是降维最经典的方法，它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。

PCA 的思想是通过坐标轴转换，寻找数据分布的最优子空间。

步骤





特征提取

降维

2. ICA(Independent Component Analysis, 独立成分分析)

ICA独立成分分析，获得的是相互独立的属性。ICA算法本质寻找一个线性变换 $z = Wx$ ，使得 z 的各个特征分量之间的独立性最大。

步骤

PCA 对数据
进行降维



ICA 来从多
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

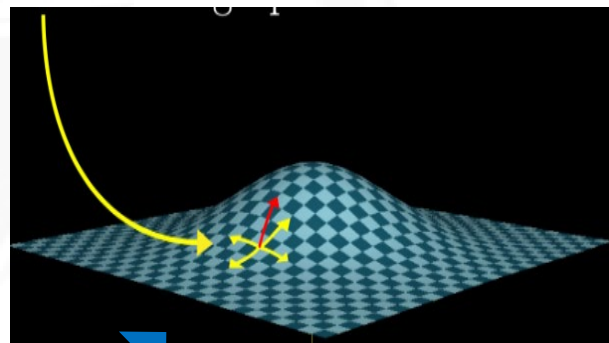
图像特征提取

1. SIFT 特征

优点:

- 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；
- SIFT 特征对参数调整鲁棒性好，可以根据场景需要调整适宜的特征点数量进行特征描述，以便进行特征分析。

缺点：不借助硬件加速或者专门的图像处理器很难实现。



步骤

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

2. HOG特征

方向梯度直方图(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。

步骤

归一化处
理

计算图像
梯度

统计梯度
方向

特征向量
归一化

生成特征
向量



特征提取

文本特征提取

1. 词袋模型

将整段文本以词为单位切分开，然后每篇文章可以表示成一个长向量，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，公式为 $TF - IDF(t, d) = TF(t, d) \times IDF(t)$

$TF(t, d)$ 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 对表达语义所起的重要性，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

2. 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。
- 另外, 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): 从给定的特征集合中选出相关特征子集的过程。

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

相关特征

- 对当前学习任务有用的属性或者特征

无关特征

- 对当前学习任务没用的属性或者特征



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
- 但同时模型就变复杂，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
- 但模型就变简单，也就降低计算复杂度



特征选择的三种方法

筛选器(Filter):

先对数据集进行特征选择，其过程与后续学习器无关，即设计一些统计量来过滤特征，并不考虑后续学习器问题

封装器(Wrapper):

就是一个分类器，它是将后续的学习器的性能作为特征子集的评价标准

嵌入式(Embedding):

是学习器自主选择特征



特征选择-筛选器

原理：先对数据集进行特征选择，然后再训练学习器
特征选择过程与后续学习器无关
也就是先采用特征选择对初始特征进行过滤，然
后用过滤后的特征训练模型

优点：计算时间上比较高效，而且**对过拟合问题**有较高的鲁棒性
缺点：倾向于选择冗余特征，即没有考虑到特征之间的相关性



特征选择-筛选器

1、Relief 方法



- ◆定义：Relevant Features是一种著名的筛选器特征选择方法。该方法设计了一个相关统计量来度量特征的重要性。
 - 该统计量是一个向量，其中每个分量都对应于一个初始特征。
 - 特征子集的重要性则是由该子集中每个特征所对应的相关统计量分量之和来决定的。
 - 最终只需要指定一个阈值 k ，然后选择比 k 大的相关统计量分量所对应的特征即可。也可以指定特征个数 m ，然后选择相关统计量分量最大的 m 个特征。
- ◆Relief 是为二分类问题设计的，其拓展变体 Relief-F 可以处理多分类问题。



特征选择-筛选器

2、方差选择法



先要计算各个特征的方差，然后根据阈值，选择方差大于阈值的特征。

3、相关系数法



先要计算各个特征对目标值的相关系数以及相关系数的 P 值。

4、卡方检验



检验定性自变量对定性因变量的相关性。假设自变量有 N 种取值，因变量有 M 种取值，考虑自变量等于 i 且因变量等于 j 的样本频数的观察值与期望的差距，构建统计量：

$$X^2 = \sum \frac{(A - E)^2}{E}$$



特征选择-筛选器

5、互信息法

概念：经典的互信息也是评价定性自变量对定性因变量的相关性的。

为了处理定量数据，最大信息系数法被提出。

互信息计算公式如下：

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$



特征选择-封装器



原理：封装器特征选择直接把最终将要使用的学习器的性能作为特征子集的评价原则。其目的就是为给定学习器选择最有利于其性能、量身定做的特征子集。



优点：直接针对特定学习器进行优化，考虑到特征之间的关联性，因此通常封装器特征选择比筛选器特征选择能训练得到一个更好性能的学习器。

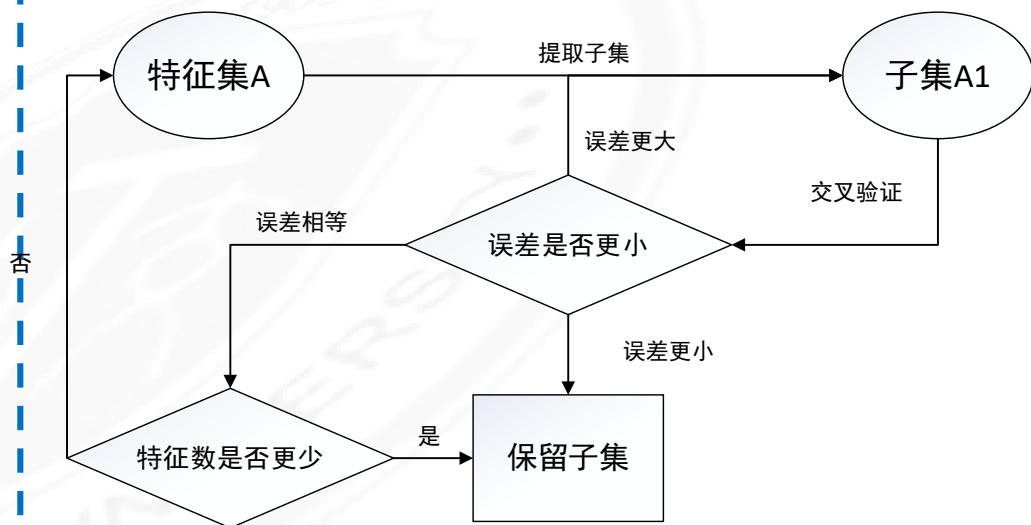
缺点：由于特征选择过程需要多次训练学习器，故计算开销要比筛选器特征选择要大得多。



特征选择-封装器

1. LVW

- Las Vegas Wrapper是一个典型的封装器特征选择方法。使用随机策略来进行子集搜索，并以**最终分类器的误差**作为特征子集的评价标准。
- 由于 LVW 算法中每次特征子集评价都需要训练学习器，计算开销很大，因此它会设计一个停止条件控制参数 T 。但是如果初始特征数量很多、 T 设置较大、以及每一轮训练的时间较长，则很可能算法运行很长时间都不会停止。

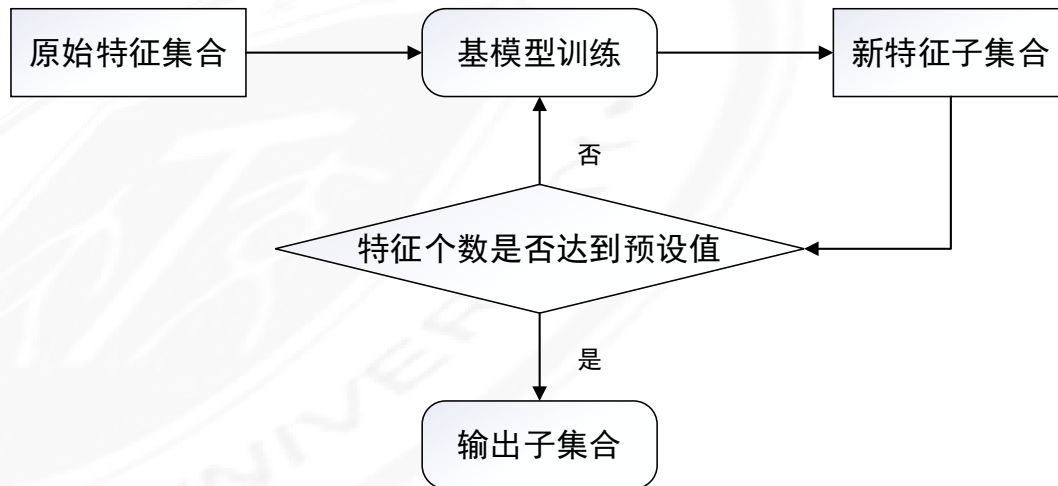




特征选择-封装器

2. 递归特征消除法

- 使用一个基模型来进行多轮训练，每轮训练后，消除若干权值系数的特征，再基于新的特征集进行下一轮训练。





特征选择-嵌入式



原理：嵌入式特征选择是将特征选择与学习器训练过程融为一体，两者在同一个优化过程中完成的。即学习器训练过程中自动进行了特征选择。



常用的方法包括：

- 利用**正则化**，如L1, L2 范数，主要应用于如线性回归、逻辑回归以及支持向量机(SVM)等算法；优点：降低过拟合风险；求得的 w 会有较多的分量为零，即：它更容易获得稀疏解。
- 使用决策树思想，包括决策树、随机森林、Gradient Boosting 等。



特征选择-嵌入式

常见的嵌入式选择模型：



在 Lasso 中， λ 参数控制了稀疏性：

- 如果 λ 越小，则稀疏性越小，被选择的特征越多
- 相反 λ 越大，则稀疏性越大，被选择的特征越少



在 SVM 和 逻辑回归中，参数 C 控制了稀疏性：

- 如果 C 越小，则稀疏性越大，被选择的特征越少
- 如果 C 越大，则稀疏性越小，被选择的特征越多



特征提取VS特征选择

| 项目 | 特征提取 | 特征选择 |
|-----|--|---|
| 共同点 | 都从原始特征中找出最有效的特征 都能帮助减少特征的维度、数据冗余 | |
| 区别 | <ul style="list-style-type: none">➤ 强调通过特征转换的方式得到一组具有明显物理或统计意义的特征➤ 有时能发现更有意义的特征属性 | <ul style="list-style-type: none">➤ 从特征集合中挑选一组具有明显物理或统计意义的特征子集➤ 能表示出每个特征对于模型构建的重要性 |