Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

数据挖掘中的数据清洗方法

Data Cleaning

胡雅娴　硕士生

2018年05月13日

内容提要

- 背景简介
- 整体框架
- 预处理阶段
- 缺失值处理
- 重复数据清理
- 格式清洗
- 异常值处理
- 数据转化
- 参考文献

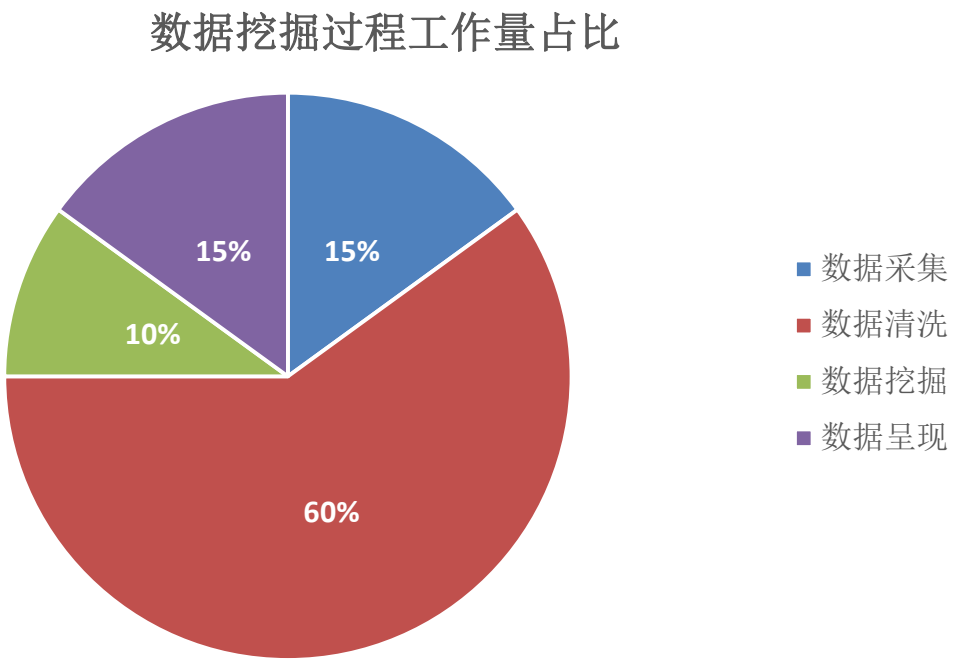
背景简介

背景简介

- 数据挖掘是从大量的、不完整的、有噪声的数据中，提取隐含在其中的潜在的有用信息的过程
- 其中最费力的事，就是**数据获取和清洗**

背景简介


数据采集
数据清洗
数据挖掘
数据呈现


数据挖掘过程工作量占比
15%
15%
10%
60%
数据采集
数据清洗
数据挖掘
数据呈现

背景简介

- 为什么要做数据清洗？
  - 解决数据质量问题
    - 数据完整性：比如人的属性中缺少性别、年龄等
    - 数据唯一性：比如不同来源的数据重复出现
    - 数据权威性：比如同一个指标出现多个来源的数据，且数值不一样
    - 数据合法性：比如数据与常识不符
    - 数据异质性：比如不同来源的不同指标，实际内涵与表示意义是一样的
  - 将“脏”数据变成标准的、干净的数据，更加适合挖掘

背景介绍

- 我们做过的数据清洗包括哪些步骤?

删除空缺值

删除重复值

这样做数据清洗真的够了吗?

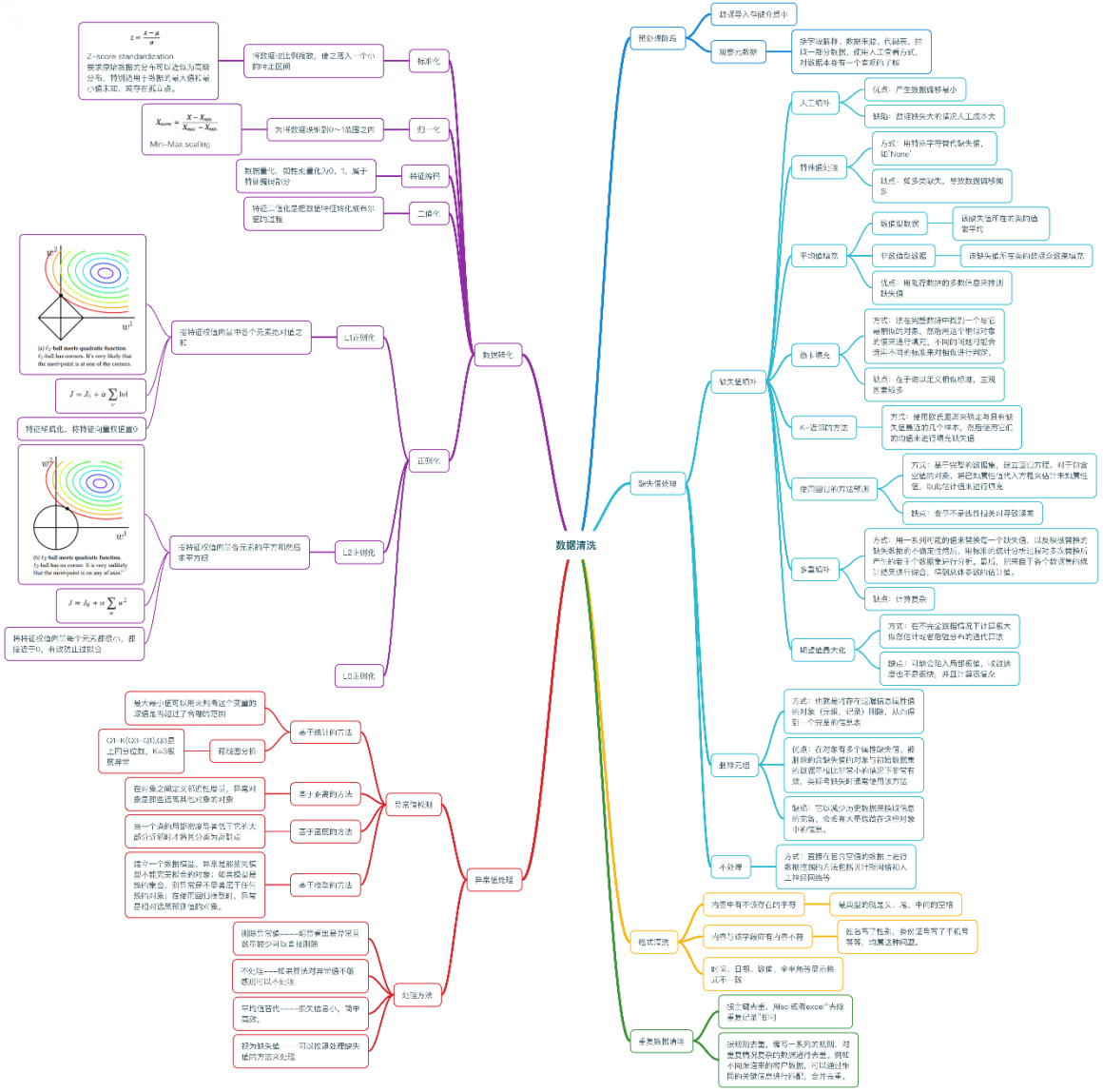
删除异常值

NO!

删删删……

整体框架

整体框架


数据清洗
预处理阶段
数据导入存储介质中
观察元数据
字段说解释，数据来源，代码表，抽取一部分数据，使用人工查看方式，对数据本身有一个全面的了解
缺失值处理
缺失值填补
人工填补
优点：产生数据偏移量少
缺点：数据缺失大的情况人工成本大
特殊值处理
方式：用特殊字符替代缺失值，如'None'
缺点：某多类缺失，导致数据偏移很多
平均值插补
数值型数据
该缺失值所在类的数值众数来插补
非数值型数据
该缺失值所在类的数值众数来插补
优点：用现存数据的多数信息来推测缺失值
热卡填充
方式：在完整数据中找到一个与它最相似的对象，然后用这个相似对象的值来进行填充。不同的问题可能会选择不同的标准来对相似进行判定。
缺点：在于定义相似标准，主观因素较多
K-近邻的方法
方式：使用欧氏距离来确定与具有缺失值最接近的几个样本，然后使用它们的均值来进行填补缺失值
使用回归的方法预测
方式：基于完整的数据集，建立回归方程。对于包含空值的对象，将已知属性值代入方程来估计未知属性值，以此估计值来进行填充
缺点：变量不是线性相关时会导致错差
多重插补
方式：用一系列可能的值来替换每一个缺失值，以反映被替换的缺失数据的不确定性然后，用标准的统计分析过程对多次替换后产生的若干个数据集进行分析，最后，把来自于各个数据集的统计结果进行综合，得到缺失数据的估计值。
缺点：计算复杂
最值插值优化
方式：在不完全数据信息下计算极大似然估计或者后验分布的迭代算法
缺点：可能会陷入局部极值，收敛速度也不是很快，并且计算很复杂
删除元组
方式：也就是消存在这些信息属性缺失的对象（元组，记录）删除，从而得到一个完整的信息表
优点：在对象有多个属性缺失值，被删除的含缺失值的对象与初始数据集的数据量相比相差小的情况下非常有效，类标号缺失时通常使用该方法
缺点：它以减少历史数据来换取信息的丰富，会丢弃大量隐藏在这些对象中的信息。
个处理
方式：直接将含空值的数据之间进行数据挖掘的方法包括贝叶斯网络和人工神经网络等
格式清洗
内容中有不该存在的字符
某些型的如是头、尾、中间的空格
内容与该字段应有内容不符
姓名写了性别，身份证号写了手机号等等，均属这种问题。
时间、日期、数值、全半角等显示格式不一致
重复数据清理
按主键去重，用sql或者excel"去除重复记录"即可
按规则去重，编写一系列关联规则，对重复情况复杂的数据进行去重。例如：不同数据源的客户数据，可以通过相同的关键信息进行判断，合并去重。
数据转化
标准化
将数据按比例缩放，使之落入一个小的特定区间
$z=\frac{x-\mu}{\sigma}$
Z-score standardization
需要原始数据符合均值为0，标准差为1的正态分布，特别适用于当数据的最大值和最小值未知，或存在孤立点。
归一化
为将数据映射到0~1范围之内
$X_{norm}=\frac{X-X_{min}}{X_{max}-X_{min}}$
Min-Max scaling
特征编码
数据量化：如性别量化为0，1，属于特征编码部分
二值化
特征二值化是把数据特征转化成布尔值的过程
正则化
L1正则化
当特征向量的某些元素绝对值之和
$J=J_0+\alpha\sum_w|w|$
特征矩阵化，将特征向量权值置0
L2正则化
当特征向量的各元素的平方和然后求平方根
$J=J_0+\alpha\sum_w w^2$
将特征向量的某些元素都很小，都接近于0，有效防止过拟合
L0正则化
异常值处理
异常值检测
基于统计的方法
箱线图分析
基于距离的方法
基于密度的方法
基于模型的方法
处理方法
删除异常值——明显看出是异常且数量较少可以直接删除
不处理——如果算法对异常值不敏感则可以不处理
平均值替代——损失信息小，简单高效。
视为缺失值 可以按照处理缺失值的方法来处理

整体框架

预处理阶段

预处理阶段

- 数据导入存储介质中
  - 使用数据库或文本文件存储
- 观察数据
  - 看字段解释、数据来源等一切描述数据的信息
  - 抽取一部分数据，使用人工查看方式，对数据本身有一个直观的了解，并且初步发现一些问题，为之后的处理做准备

缺失值处理

缺失值处理

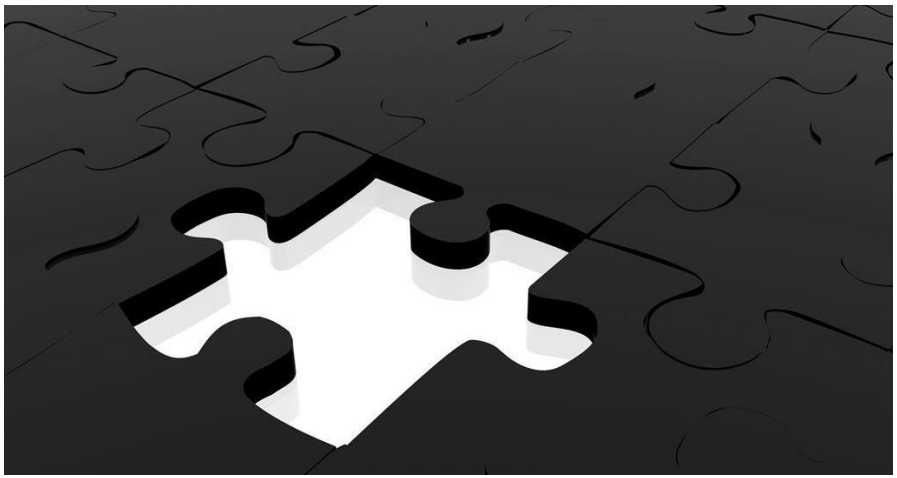
- 造成数据缺失的原因
  - 信息暂时无法获取，如医疗数据库中，并非所有病人的所有临床检验结果都能在给定时间内得到
  - 信息被遗漏，可能由于输入时认为不重要，或由于数据采集设备的故障等原因导致的数据丢失
  - 有些对象的某个属性或某些属性是不可用的，如一个未婚者的配偶姓名、一个儿童的固定收入状况等
  - 获取信息的代价太大

缺失值处理

- 缺失的类型
  - 完全随机缺失：数据的缺失是完全随机的，不依赖于任何不完全变量或完全变量，不影响样本的无偏性，如家庭住址
  - 随机缺失：数据的缺失不是完全随机的，即该类数据的缺失依赖于其他完全变量，如是否有配偶姓名取决于是否已婚
  - 非随机缺失：数据的缺失与不完全变量自身的取值有关，如高收入人群不愿意提供家庭收入

缺失值处理

- 缺失值的存在所造成的影响
  - 系统丢失了大量的有用信息
  - 系统中所表现出的不确定性更加显著
  - 包含空值的数据会使挖掘过程陷入混乱，导致不可靠的输出



缺失值处理

- 缺失值填补
  - 人工填补：当数据缺失比例很小时，可以直接对缺失记录进行手工处理
    - 优点：产生数据偏移最小
    - 缺点：数据缺失大时人工成本大，一般不推荐使用
  - 特殊值处理：将空值作为一种特殊的属性值来处理，不同于其他的任何属性值，如’None’
    - 缺点：可能导致严重的数据偏离，一般不推荐使用
  - 平均值填充
    - 数值型数据：用该缺失值所在类的平均值来填充
    - 非数值型数据：用该缺失值所在类的数据众数来填充
    - 优点：用现存数据的多数信息来推测缺失值

缺失值处理

- 缺失值填补
  - 热卡填充：在完整数据中找到一个与它最相似的对象，然后用这个相似对象的值来进行填充。不同的问题可能会选用不同的标准来对相似进行判定。
    - 缺点：难以定义相似标准，主观因素较多
  - K-近邻方法：使用欧式距离来确定与具有缺失值最近的几个样本，然后使用它们的均值来填充
    - 二维平面上两点a(x1,y1)与b(x2,y2)间的欧氏距离：

$$d_{12} = \sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$

缺失值处理

- 缺失值填补
  - 回归：基于完整的数据集，建立回归方程，对于包含空值的对象，将已知属性代入方程来估计未知属性值，以此估计值来填充
    - 缺点：变量不是线性相关时容易导致误差
  - 多重填补：用空缺属性的所有可能的属性值来填充，产生若干个完整数据集合，每个填补数据集合都用针对完整数据集的统计方法进行统计分析，最后对各个结果进行综合
    - 缺点：当数据量很大或者遗漏的属性值较多时，其计算的代价很大

缺失值处理

- 删除数据
  - 将存在遗漏信息属性值的数据删除，从而得到一个完备的信息表
  - 优点：在对象有多个属性缺失值、被删除的含缺失值的对象与初始数据集的数据量相比非常小的情况下，该方法很有效
  - 缺点：以减少历史数据来换取信息的完备，会丢弃大量隐藏在这些数据中的信息

缺失值处理

- 不处理
  - 对空缺值的不正确填充往往会将新的噪声引入数据中，使挖掘任务产生错误的结果
  - 有些情况下，希望在保持原始信息不发生变化的情况下对数据进行分析
  - 直接在包含空值的数据上进行数据挖掘的方法包括贝叶斯网络等

缺失值处理

- 对缺失值的处理要具体问题具体分析
  - 年收入
    - 商品推荐：填充平均值
    - 借贷额度：填充最小值
  - 人体寿命
    - 保险费用估计：填充最大值
    - 人口估计：填充平均值

重复数据清理

重复数据清理

- 按关键信息去重
  - 例如，以姓名、ID等唯一关键信息去除重复数据
- 按规则去重
  - 编写一系列的规则，对重复情况复杂的数据进行去重
  - 例如，从不同渠道来的数据，可以通过相同的关键信息进行匹配，合并去重

格式清洗

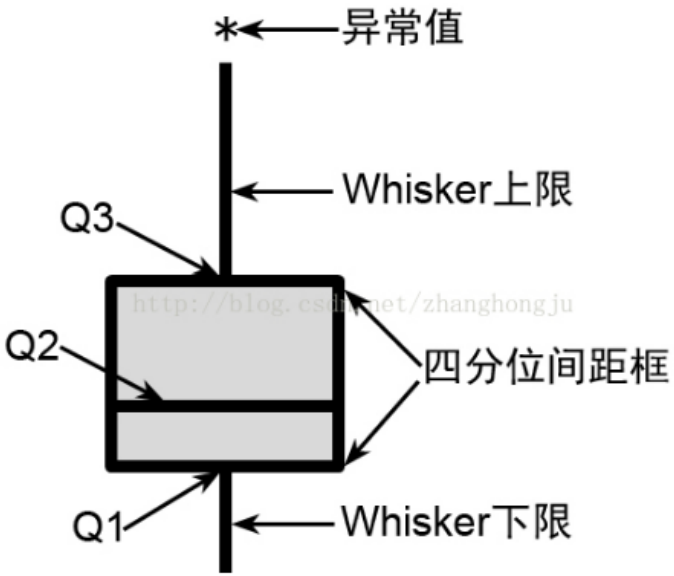
格式清洗

- 内容中有不该存在的字符
  - 头、尾、中间的空格，需要以半人工方式来找出可能存在的问题，并去除不需要的字符
- 内容与该字段应有内容不符
  - 姓名写了性别，身份证号写了手机号等，不能简单的以删除来处理，需要详细识别问题类型
- 时间、时期、数值、全半角等显示格式不一致
  - 在整合多来源数据时可能遇到，将其处理成一致的某种格式即可

异常值处理

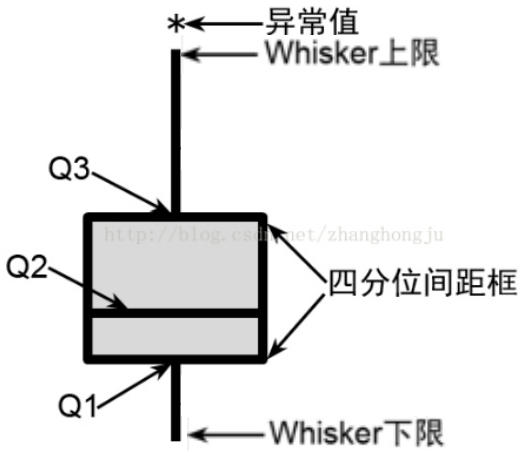
异常值处理

- 异常值检测
  - 基于统计的方法
    - 最大最小值可以用来判断这个变量的取值是否超过了合理的范围
    - 箱线图分析



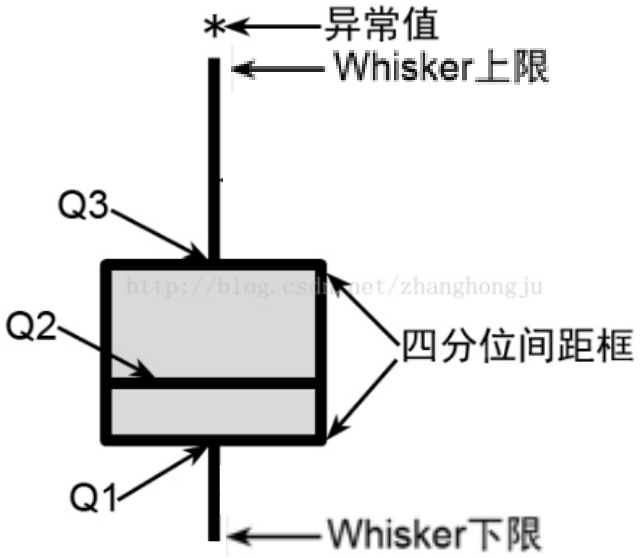
异常值处理

- 箱线图分析
  - 用一组数据中的最小值、第一四分位数、中位数、第三四分位数和最大值来反映数据分布的中心位置和散布范围
  - Q1的位置=(n+1)/4
  - Q2的位置=(n+1)/2，中位数的位置
  - Q3的位置=3(n+1)/4
  - Whisker上限是延伸至距框顶部、1.5倍框高范围内的最大数据点
  - Whisker下限是延伸至距框底部、1.5倍框高范围内的最小数据点
  - 超出Whisker上限或下限的数值将使用星号“*”表示



异常值处理

- 箱线图分析
  - 例如有1, 2, 5, 4, 3, 7, 15五个数据
  - 按顺序排列为1, 2, 3, 4, 5, 7, 15
  - Q2=(7+1)/2=4
  - Q1=(7+1)/4=2
  - Q3=3(7+1)/4=6
  - 四分位间距框高度：6−2=4
  - 异常值：15

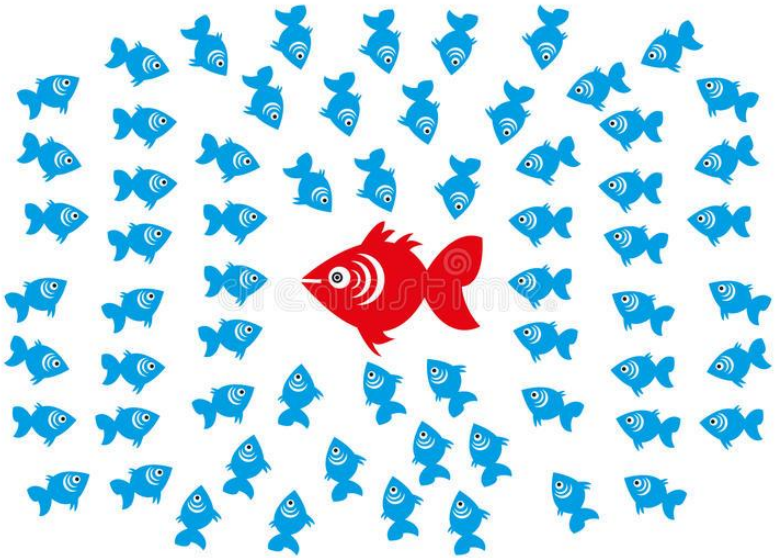


异常值处理

- 异常值检测
  - 基于距离的方法：在对象之间定义邻近性度量，异常对象是那些远离其他对象的对象
  - 基于密度的方法：当一个点的局部密度显著低于它的大部分近邻时，认为其是离群点
  - 基于模型的方法
    - 建立一个数据模型，异常值是同模型不能完美拟合的对象
    - 模型是簇的集合，则异常值是不显著属于任何簇的对象
    - 回归模型，异常值是相对远离预测值的对象

异常值处理

- 处理方法
  - 删除异常值：明显看出是异常且数量较少可以直接删除
  - 不处理：如果算法对异常值不敏感则可以不处理
  - 平均值代替：损失信息少，简单高效
  - 视为缺失值：可以按照处理缺失值的方法来处理



数据转化

数据转化

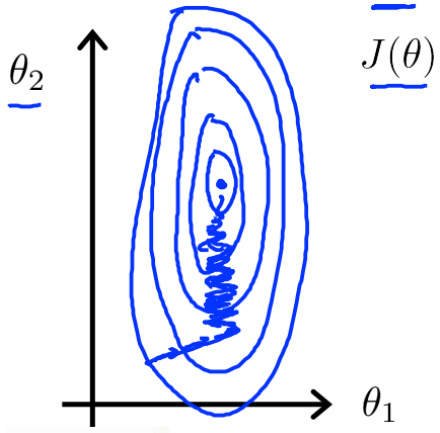
- 标准化
  - 将数据按比例缩放，使之落入一个小的特定区间
  - Z-score standardization: $z = \frac{x-\mu}{\sigma}$
  - 适用于属性的最大值和最小值未知的情况，或有超出取值范围的离群数据的情况
- 归一化
  - 将数据映射到0-1范围之内
  - Min-Max scaling: $X_{norm} = \frac{X-X_{min}}{X_{max}-X_{min}}$

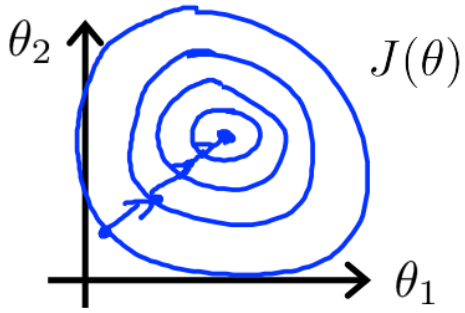
数据转化

- 标准化和归一化可以统称为标准化
- 为什么要做标准化?
  - 去除数据的单位限制，将其转化为无量纲的纯数值，便于不同单位或量级的指标能够进行比较和加权
  - 提升模型的精度，在涉及到距离计算的算法时效果显著
    - 如需要计算欧氏距离，取值范围小的属性比取值范围大的属性对结果的影响较小

数据转化

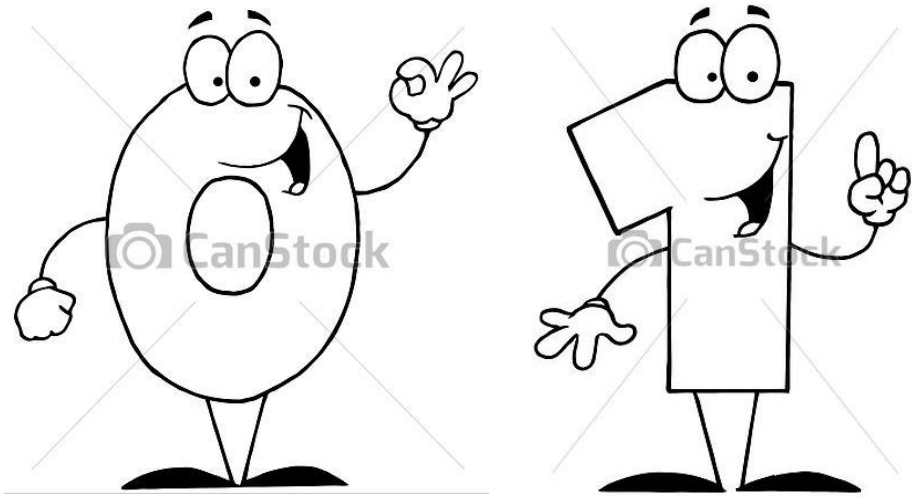
- 标准化和归一化可以统称为标准化
- 为什么要做标准化?
  - 提升模型的收敛速度
    - θ1的取值为1-5，θ2的取值为0-2000，左图是一个窄长的椭圆形，导致在梯度下降时，梯度的方向为垂直等高线的方向而走之字形路线，这样会使迭代很慢，相比之下，右图（归一化[0,1]）迭代更快





数据转化

- 特征编码
  - 数据量化，如性别量化为0和1，等级高中低量化为0, 1, 2
- 二值化
  - 把数据特征转化为布尔值，即设定一个阈值，大于阈值则为1，小于等于阈值则为0



参考文献

参考文献

[1] https://blog.csdn.net/lujiandong1/article/details/52654703
[2] https://blog.csdn.net/zhanghongju/article/details/18446131
[3]https://blog.csdn.net/pipisorry/article/details/52247379
[4] https://www.deeplearn.me/1389.html
[5] http://www.ituring.com.cn/book/tupubarticle/9676

道德经

知人者智，自知者明。
胜人者有力，自胜者
强。知足者富。强行
者有志。不失其所者
久。死而不亡者，寿。

谢谢！