



## 2.1 机器学习中的数据处理

HUAWEI TECHNOLOGIES CO., LTD.



[www.huawei.com](http://www.huawei.com)

# 目录

## Contents

1

样本级数据处理

2

特征级数据处理

3

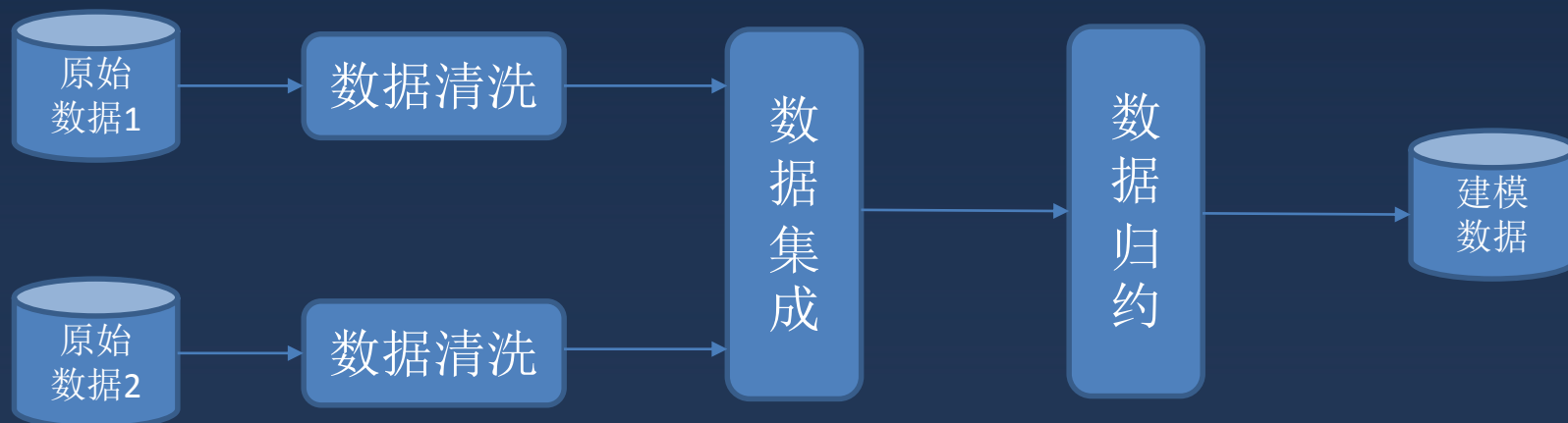
集合级数据处理

# 数据预处理的意义

机器学习建模需要高质量的数据：准确、完整、一致、时效、可信。

实际应用场景中的数据存在：不完整、不正确、含噪声。

一句略显夸张的话：数据处理3个月，建模5分钟



# 样本级数据处理

从样本的角度进行数据处理。主要是：样本选择、样本生成

**样本选择**：存在大量的无意义样本时（比如设备异常检测场景下的设备实时数据），需要对样本进行挑选，使得不同类型的样本达到均衡。



简单去重是指去掉几乎完全相同的样本（用阈值控制）；样本可视化是指将样本按特征显示出来，去掉变化规律几乎一致的样本；规则性去重是指设定一系列基于人工的先验规则，过滤样本，规则基于深厚的业务场景理解。

**样本生成**：样本缺失严重、样本对于连续性敏感时需要生成一部分的样本，目的是样本完整性。



统计值填充是指在某特征上按照统计量（均值、中位数、最大最小值）等进行缺失值填充；K最近邻填充是指用K个最近（时间上、空间上）的样本的均值进行样本生成、GAN生成是指采用GAN网络进行样本生成。

# 目录

## Contents

1

样本级数据处理

2

特征级数据处理

3

集合级数据处理

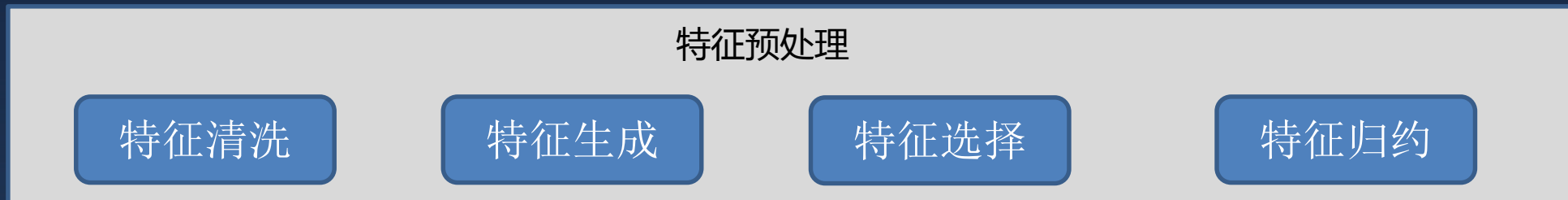
# 特征级数据处理的意义

## 特征处理本身的意义

特征是表征一个样本最重要的部分

特征是建模效果好坏的重要先行部分

特征级数据处理是数据预处理当中工作量占比最大的部分。



在长久的特征处理的战斗中形成了一系列的方法论，下面会逐个进行介绍。

- 1、特征清洗主要是进行特征上缺失值、异常值的处理
- 2、特征生成主要是进行特征的提取、组合、映射、
- 3、特征选择主要是进行特征间相关性的分析，进行特征的取舍
- 4、特征归约主要是将特征的表征（数据类型）进行一致性处理

# 特征清洗

## 特征清洗

忽略：在特征上设定规则忽略满足规则的特征值

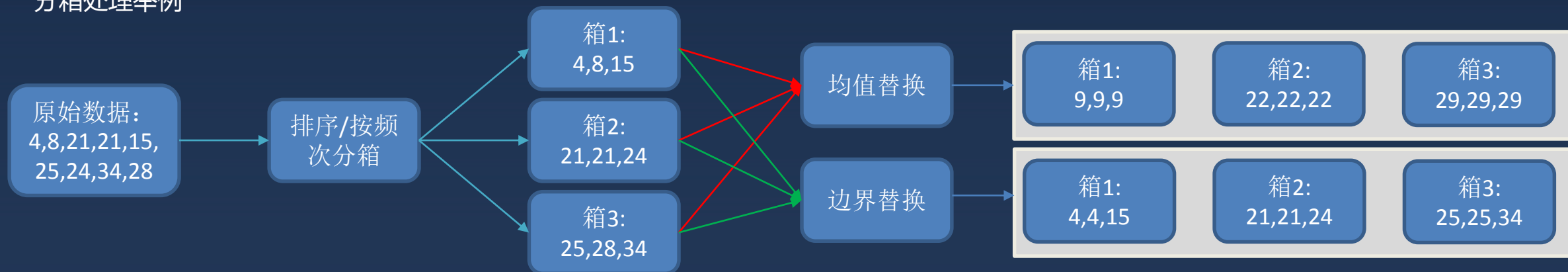
缺失值填充：使用特征上的全局统计量、K最近邻统计量进行缺失值填充

异常值处理：使用特征上的全局统计量、K最近邻统计量进行异常值的替换

分箱平滑：在特征上使用数据分箱，用每一个数据箱中的统计量代替特征本身的价值

回归平滑：使用可视化的技术、回归技术在特征上使用一个函数来进行特征拟合，特征上的离群点用函数值替换

### 分箱处理举例



# 特征生成

## 特征生成

组合：多个特征按照一定规则形成新的特征，比如加和、相乘等

特征统计：在某个特征上采用一些统计量或者特殊的计算规则生成新的特征

特征移位：在时间相关的一些场景，可以将特征进行移位操作，比如周1的A特征和周2的B特征组合成为新的特征。

## 特征组合举例

原始特征 $\langle x_1, x_2, x_3 \rangle$ ，组合特征 $\langle x_1x_2, x_2x_3, x_1x_3, x_1^2, x_2^2, x_3^2 \rangle$

## 特征统计举例

原始特征 $\langle x_1 \rangle$ ，将样本分片，统计特征 $\langle \text{mean}(x_1), \text{max}(x_1), \text{min}(x_1), \text{variance}(x_1), \text{median}(x_1), \dots \rangle$

## 特征移位举例

原始数据

x1	x2	x3
1	0.11	123
2	0.23	456
3	0.45	789
4	0.67	1001

特征移位后数据

x1	x2	x3_new
1	0.11	456
2	0.23	789
3	0.45	1001
4	0.67	null



# 特征选择

**特征选择**：面对成百上千维的特征，大概率存在着与目标工作无关的特征，是冗余的，所以需要一些方法进行特征选择。

主成分分析：将n维特征经过principal compos analysis或PCA搜索k个最能代表数据信息的维度，在搜索、计算的过程中维度本身发生了改变，产生的是原始维度的映射，将映射后的特征进行重要性排序。

特征子集：当存在n个原始特征时，即有 $2^n$ 个特征子集。在n个特征进行筛选时，可以使用统计显著性检验的方法、信息增益的方法、决策树的方法。

相关性分析：用于检验2个特征之间的相关性，主要指标有Pearson相关系数、Spearman相关系数、卡方检验等。

**主成分分析步骤**：

1、在每个特征上减去这个特征的平均值。2、计算数据集的协方差矩阵。3、计算协方差矩阵的特征值和特征向量。4、将特征值排序保留最大的N个特征值。5、将数据转换到特征向量构建的新空间中。

**特征子集主要方法**：

- 1、逐步向前，从空集开始，每一次添加当前特征集中最有价值的特征，特征价值来自于统计显著性检验、信息增益等。
- 2、逐步向后，从特征全集开始，每一次删除当前特征集中最没有价值的特征，特征价值来自于统计显著性检验、信息增益等。
- 3、决策树，构建一棵规定深度的决策树，将没有出现在树中的特征删除。

# 特征归约

**特征归约**：将特征划分到类似的空间当中，消除彼此度量差异造成的影响。

离散化：将连续型特征映射到一个类别当中，比如收入的具体数值映射为“高、中、低”

标准化：将所有的连续型特征都归约到同一个区间上。常用的区间是 $[-1, 1]$ 和 $[0, 1]$

光滑：与特征清洗类似，采用回归等技术去除特征上的噪声数据。

聚集：不关注每一条样本的特征信息，关注样本集分片后的特征信息，等同于在特征上分片。比如原始样本是日销售额，而建模可以只关注周销售额。

**离散化方法**：

- 1、等宽分箱（连续值的区间划分是宽度一致的）
- 2、等频分箱（连续值的划分是出现次数一致的）
- 3、聚类分箱（按照特征上聚类的结果进行分箱）
- 4、邻近分箱（在分类问题中选择跟自己最近的类别标签相同的值进行合并）

**标准化方法**：

- 1、最大最小值标准化
- 2、Z-score标准化
- 3、和值比例标准化

# 目录

## Contents

1

样本级数据处理

2

特征级数据处理

3

集合级数据处理

# 集合级数据处理

**What's 集合级数据处理**，两个方面：

- 1、很多数据集是一个二维表，行是样本，列是特征，但是还有一些场景下的数据是一个样本就是一个二维表，将本来由一个二维表表示一个样本转换为一行数据表示一个样本。
- 2、当样本处理、特征处理完成之后，进入到建模阶段，需要考虑使用怎样的验证集和测试集去判定模型的好坏，涉及到样本集合的划分。

## 二维表→样本

- 1、将二维表分片
- 2、在每个分片上在原始每个特征上进行基于预先规则的计算，每个特征对应多个结果
- 3、每个分片变成了一行数据
- 4、二维表变成了多条行数据

## 训练数据划分

- 1、随机划分训练集、验证集和测试集
- 2、指定测试集，其余部分随机划分训练集和验证集

训练数据的划分方式需要根据具体的业务场景而定



# Thank You.

**Copyright©2016 Huawei Technologies Co., Ltd. All Rights Reserved.**

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

华为云机器学习服务MLS  
[www.huaweicloud.com/product/mls.html](http://www.huaweicloud.com/product/mls.html)