# 数据预处理

数据预处理是指进行模型求解前对数据的处理过程，主要包括有数据清洗、数据集成、数据变换、数据降维这几类步骤或方法。

## 数据清洗

数据清洗的目的是为了避免GIGO（Garbage In Garbage Out）,如果输入本身就是有问题的，那不论模型多么强大，得到的也只是有问题的输出。数据清洗的内容包括但不限于：恶意数据清洗、去噪、异常值处理、缺失值处理、记录去重等。数据清洗一般在log级数据就要进行。

### 恶意数据清洗

恶意数据清洗和业务场景密切相关，一般会直接或间接地涉及到金钱关系。比如频繁刷量、点赞等。在腾讯“神盾”推荐系统中采用Guid来标识每一个设备的思路进行恶意数据源的判断。

### 去噪

去噪的目的是去除偶然触发，并非被用户有意操作的行为日志。比如用户点了某个广告，并不一定代表用户对这个广告感兴趣，可能只是不小心的点击行为。

### 异常值处理

不排除会出现数据中的某些字段或某些部分严重偏离该字段或该部分所具有的统计分布（比如均值、方差）。这时候就需要对这些异常记录进行检测并处理。检测的方法有人工设置规则、聚类、分箱法、回归分析等。

### 缺失值处理

由于各种原因，上报的数据中某些字段存在缺失。处理这种情况的方式有删除整条记录、填充（比如属性平均值、默认值）、预测（用其他字段的信息预测缺失字段的值）、用协同过滤（矩阵补全）的方式进行补全等。

### 记录去重

因网络环境或者代码bug等一些原因，存在某些数据重复上报的情况。重复记录的比率较低时，可以根据某一字段或某些字段的组合进行去重。

## 数据集成

数据集成是指把来自不同数据源的多组数据组合成一个整体，主要的处理对象是已经结构化的数据。集成过程中会碰到的问题比如不同字段的含义不清、字段的冗余等。

## 数据变换

数据变换的目的是构造更加合适模型训练的输入数据。主要的做法有：归一化、标准化、平滑、离散化&二值化。

### 归一化

归一化的目的是去除字段中不同量纲的影响，使处理后的数据落在[0,1]范围内。常用的方法有最小-最大法、最大值法（除以本字段的最大值）等。

### 标准化

标准化的目的是使每一维度的数据具有零均值和单位方差。标准化要求原始数据的分布可以近似为高斯分布，否则效果比较差。计算公式为：

在分类、聚类算法中、需要使用距离度量的时候、或者使用降维技术进行降维的时候，数据的标准化应用的比较多。

### 平滑

平滑一般应用在求比率的场景中，目的是为了让求得的比率更加接近实际情况。

贝叶斯平滑、拉普拉斯平滑、log方式平滑等。

### 离散化&二值化

离散化是指把不同的连续值划分到不同的区间中，比如把年龄分到不同的年龄段中。离散化有助于拟合数据中的非线性关系，因为有时候连续值所在区间的差异是更有意义的。

在对一些id、类目做处理时，经常使用二值化进行表达。因为在模型中这些特征的相加相减是没有意义的，但对应的是哪一个id、类目这种信息是有意义的。

并不是所有的模型在训练之前都要先对输入数据做变换处理，和具体模型的训练原理有关。比如在GBDT/RF等基于决策树的模型中，可以不做或少做数据的归一化、标准化处理，训练LR模型的时候就需要对id、类目做二值化，训练神经网络、k-means的时候就需要对数据做标准化处理等。