特征处理方法

在本章中我们学习了如何对原始数据进行特征工程的处理。针对不同类型的数据，使用的处理方式往往也不尽相同，下面我们针对**数值型特征**和**离散型特征**分别总结常见的两种特征处理方式：

无量纲化

在处理数值类特征的时候最常见的需求就是对特征进行无量纲化处理。因为不同的原始特征往往有各自不同的单位，数据的分布也不同。对于任何机器学习模型而言，想要同时处理这些特征就要求它们尽量有相似的分布，并去掉各自独特量纲的影响。

无量纲化有以下两个常用方法：

- 标准化 Standardization

标准化会试图将数据转化为标准正态分布，从而达到无量纲化的目的。具体算法为：

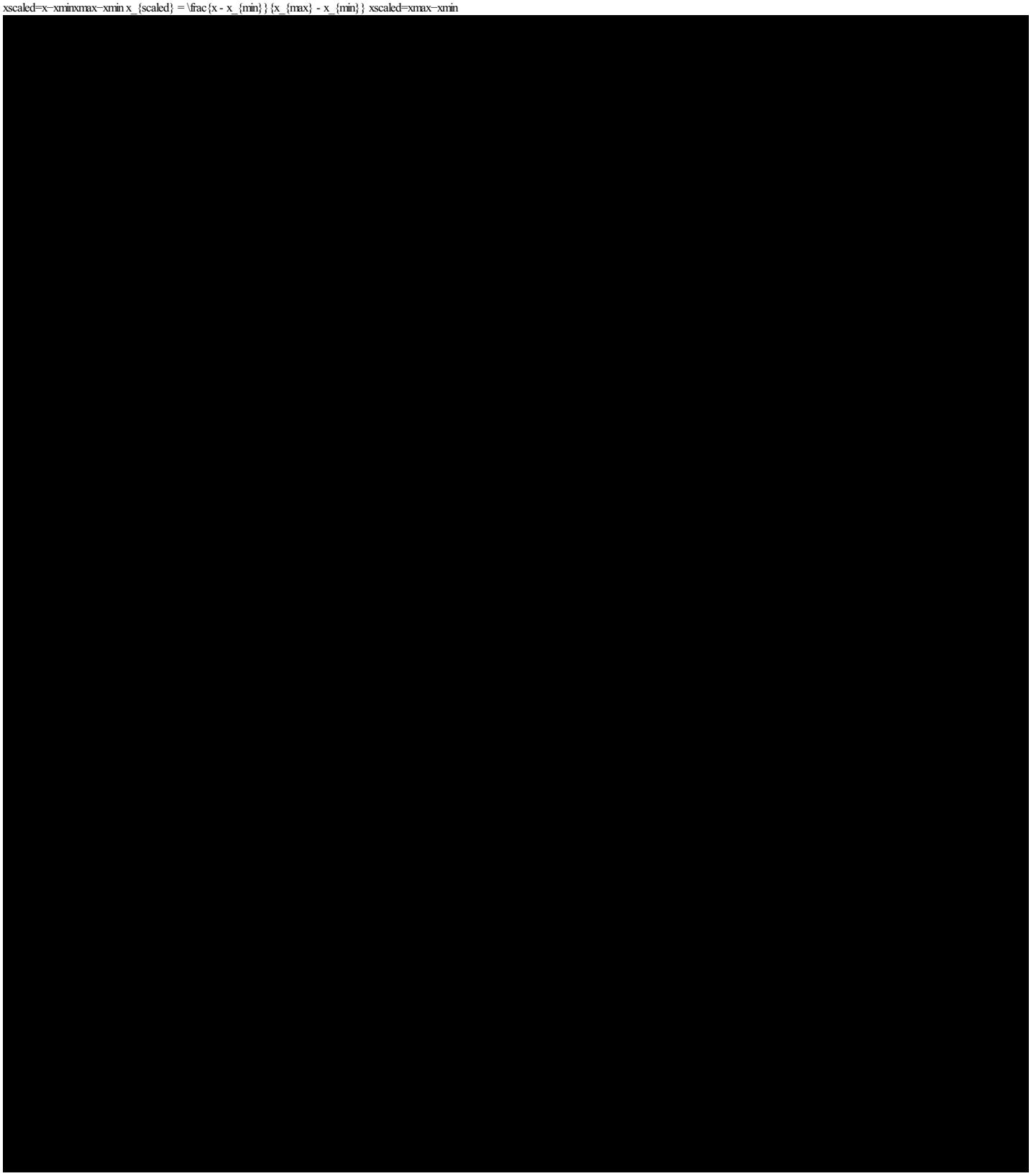
KaTeX parse error: Expected group after '^' at position 3: x^、_i = \frac{x_i…

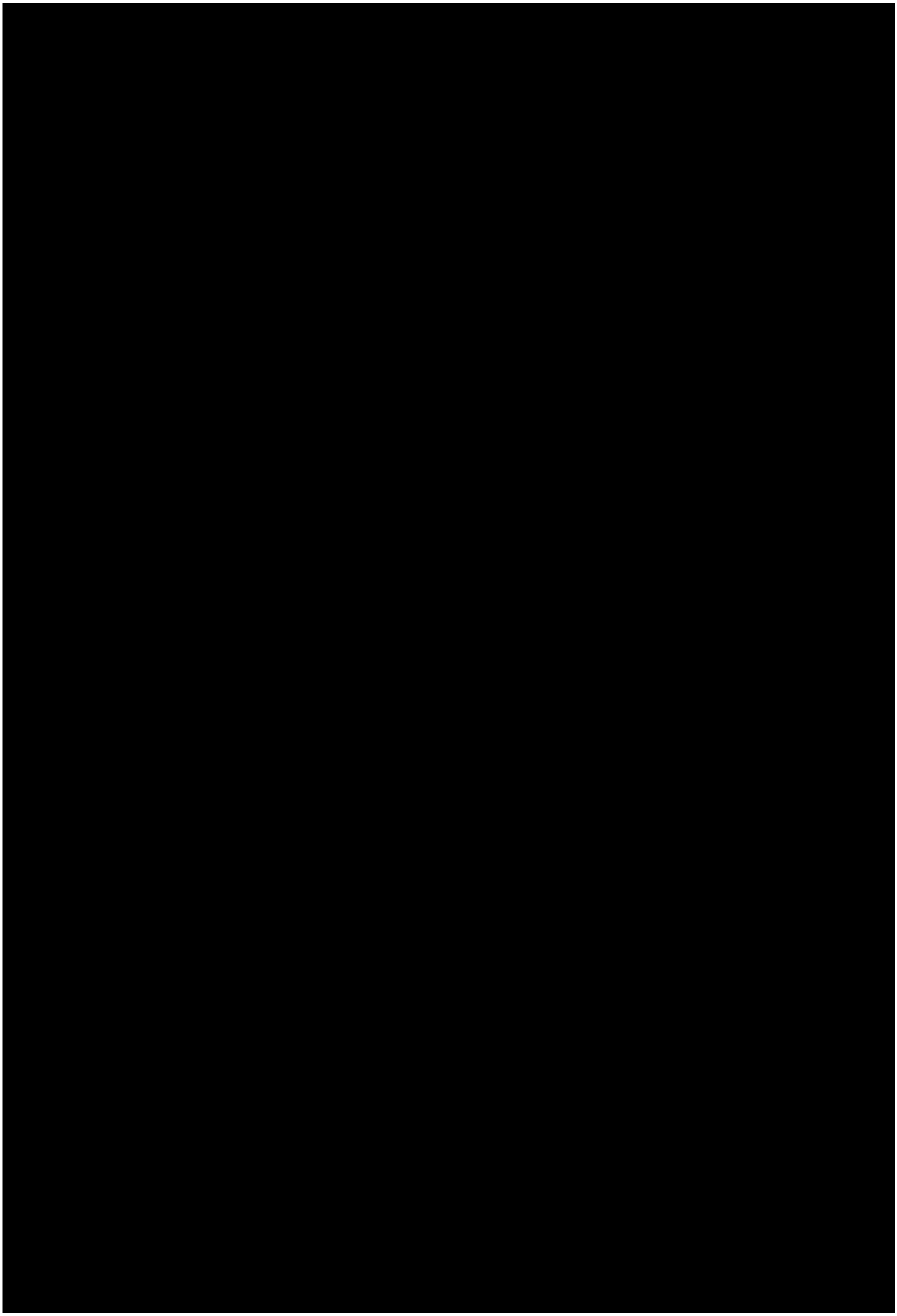
其中μX\mu_XμX为所有样本的平均值，而σX\sigma_XσX为样本的标准差。

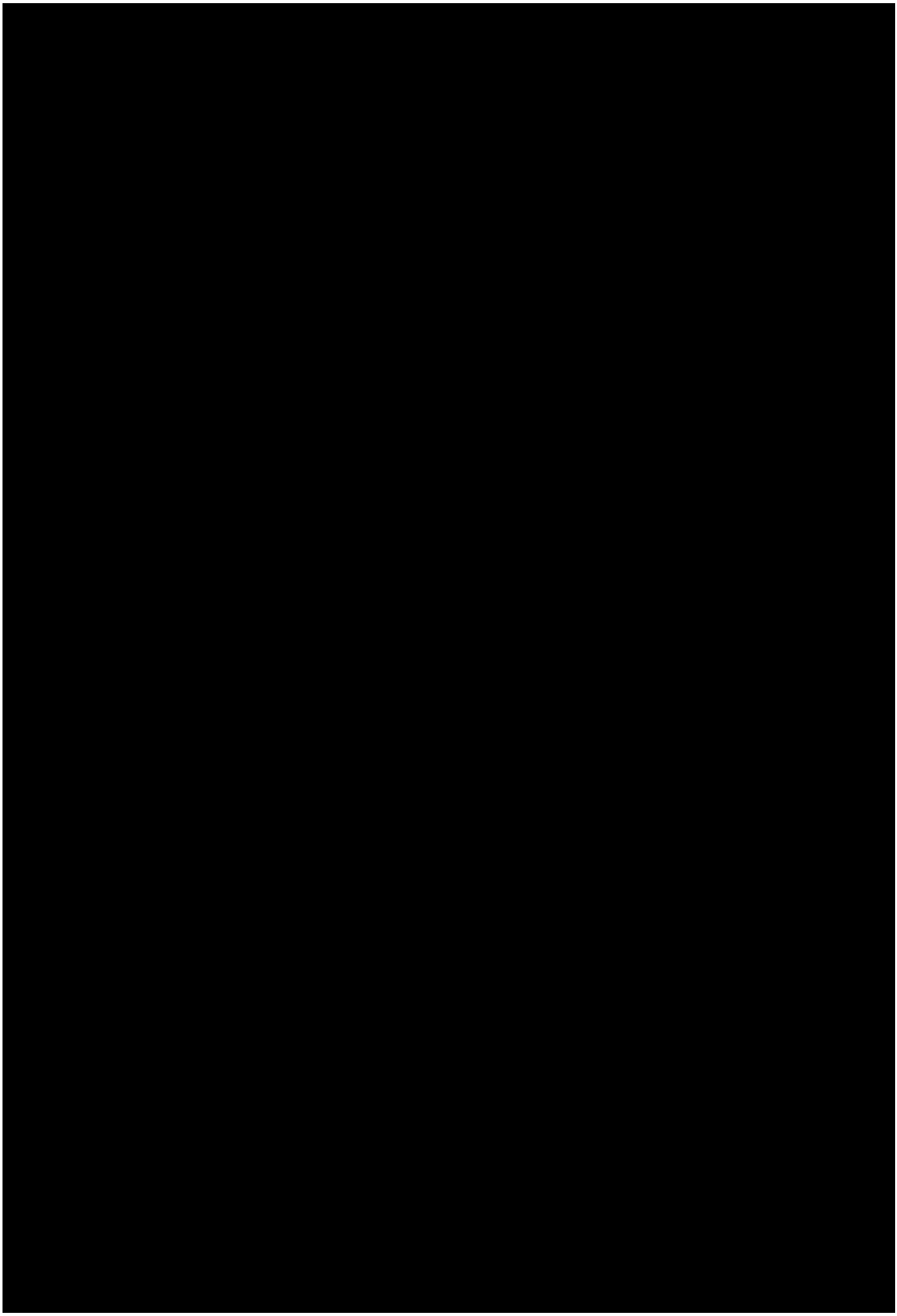
- 归一化 Normalization

归一化会将所有样本数据都缩放至一个特定大小的区间内（如0-1），具体算法为：

xscaled=x−xminxmax−xmin x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} xscaled=xmax−xmin



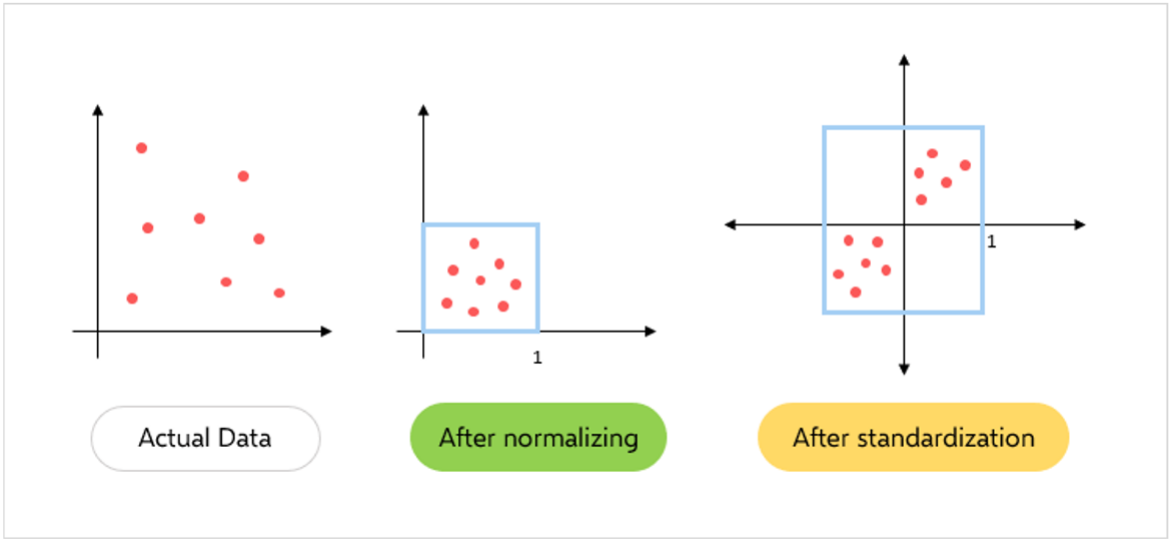




x−xmin

其中xminx_{min}xmin为样本最小值，xmaxx_{max}xmax为样本最大值。

标准化和归一化的区别可以很直观的用以下这张图看出来：



哑编码

哑编码的目的是讲离散型特征转变为数值型特征，这样才能被机器学习模型利用。常见的哑编码分为One-hot Encoding和Multi-hot Encoding.

One-hot Encoding

one-hot哑编码会将每种可能的特征取值映射为一个由0和1组成的向量，并且保证 1)每一个特征取值都有唯一的映射，2)每个特征向量只有一位是1，其余各位均为0.

比如对于用户的性别，我们可以简单的对其进行如下映射：

```
男 => [0, 1]
女 => [1, 0]
```

再比如对于数码相机品牌，我们也可以有类似的映射：

```
佳能 => [1, 0, 0, 0, 0]
索尼 => [0, 1, 0, 0, 0]
尼康 => [0, 0, 1, 0, 0]
富士 => [0, 0, 0, 1, 0]
理光 => [0, 0, 0, 0, 1]
```

Multi-hot Encoding

相信大家都已经发现了，ont-hot编码适合于那些单一取值可能的特征，比如某个用户的性别，某个商品的品牌。而对于那些有多种取值可能的特征（比如电影的类型），我们就需要用multi-hot编码对其进行处理了。
multi-hot编码的过程和one-hot非常类似，很像对于原始的one-hot编码取并集，或者是做or操作。

比如电影的类型可以被如下编码：

```
动作、科幻        => [1, 1, 0, 0]
爱情、动画        => [0, 0, 1, 1]
动画、科幻        => [0, 1, 0, 1]
动作、科幻、动画  => [1, 1, 0, 1]
```

经过这些变化后，我们就可以将原本离散的分类型特征转化为一个数值型的特征向量，从而输入模型进行训练了。