

# 中心化和规范化

2019 年 7 月 19 日

## 1 利用线性模型推导中心化

假设样本有 3 个属性，则二分类线性模型为：

$$y = w_1x_1 + w_2x_2 + w_3x_3 + b$$

其中  $w_i (i = 1, 2, 3)$  是权重,  $b$  是偏置,  $y$  是模型输出。因为要得到最优化模型，所以参数只有在实数域取值才能最好地与样本数据一致。在进行样本预测时，需要根据  $y$  进行二分类，最合理的做法是将  $y$  与某个固定值比较，大于该固定值，则判断该样本为正类，否则为负类。由于  $y$  的取值是在实数域，所以该固定值最合理的取值是 0。

使变量 0 均值化的预处理过程称为中心化。中心化是每个样本的特征属性减去所有样本（正负样本）的对应特征属性的均值，是逐属性操作。

例：代码实现

```
1 D = 782 # 数据维度
2 N = 128 # 样本数量
3 X = np.random.randn(N, D) # 一行一个样本
4 X -= np.mean(X, axis=0)
```

其中，X 是所有样本数据的矩阵，每行一个样本。np.mean(X, axis=0) 用于计算所有样本中每个属性的均值。

## 2 利用属性同等重要性推导规范化

当采用迭代法进行模型优化时，需要对参数进行初始化。为了打破对称性，权重初始化为均值为 0 的小随机数。特别地偏置  $b$  基本都直接初始化

为 0.

当采用梯度下降法进行优化时，每个参数的更新方式是一致的，故希望参数的最优值范围大致相当，这样可以使所有参数同步收敛提高学习效率。要求每个变量  $x_i$  的取值范围一致，这就是规范化操作。

最常用的规范化就是除以标准差，这个操作与中心化操作一致，使逐属性的，标准差是通过计算所有样本得到的。规范化之前一定要进行中心化。例：代码实现

```
1      D = 784
2      N = 128
3      X = np.random.randn(N, D)
4      X -= np.mean(X, axis=0)
5      X /= np.std(X, axis=0)
```

中心化和规范化的组合操作就是对随机变量的标准化，变为均值方差为 0，方差为 1 的分布。数学公式为：

$$\hat{x} = (x - \mu) / \sigma$$

其中  $\mu$  为均值， $\sigma$  是标准差。

中心化和规范化组合操作的另一种常见方式是把每个属性的取值都统一到  $[-1, 1]$  这个区间内。

例：代码实现

```
1      D = 784
2      N = 128
3
4      X = np.random.randn(N, D)
5      minX = np.min(X, axis=0)
6      maxX = np.max(X, axis=0)
7
8      X = (X - minX) / (maxX - minX)
9      X = 2*X - 1
```

规范到  $[-1, 1]$  的方法比较容易受到噪声或异常值的影响。

规范化的假设认为每个属性的重要程度一致，如果有先验知识能表明某些

属性更重要或更不重要则其规范化范围需要做出调整，具体变换到多大范围由算法的超参数决定。

### 3 中心化和规范化的几何意义

中心化的几何意义是平移样本云，使样本点云的重心与原点重合，这样分类超平面就在原点附近，容易学习。

规范化的几何意义是对点云进行拉伸，使数据点云完全处于高维正方形内。物理意义是去除每个属性量纲的影响。