## 数据归一化

### 作用

1.提高归一化速度

就是将数据除以该列的最大值，是该列<=1，数据归一化可以是等高线显得很圆，在梯度下降时能够较快收敛。（梯度下降时沿垂直等高线的方向，如果等高线很圆，就能很快指向圆心）

2.有可能提高精度

一些分类器需要计算样本间的距离，如果一个特征值范围非常大，那么计算距离就主要靠这个特征，可能与实际情况相悖。归一化可以有效解决这个问题。

3.试用范围

概率模型不需要归一化，因为他不关心变量的值，而是关心变量的分布和变量之间的条件概率，如决策树、rf。

而像adaboost、gbdt、xgboost、svm、lr、KNN、Kmeans之类的最优化问题需要归一化。

### 方法

#### 1.线性归一化

x’= 

#### 2.标准差归一化（0均值）





其中，μ、σ分别为原始数据集的均值和方法。该种归一化方式要求原始数据的分布可以近似为高斯分布，否则归一化的效果会变得很糟糕。

或者



#### 3.非线性归一化

应用于特征分布跨度很大，采用非线性归一化。通过一些函数，将原始值进行映射。一般采用log、指数函数、正切等函数。

#### 方法比较

1、在分类、聚类算法中，需要使用距离来度量相似性的时候、或者使用PCA技术进行降维的时候，第二种方法(Z-score standardization)表现更好。

2、在不涉及距离度量、协方差计算、数据不符合正太分布的时候，可以使用第一种方法或其他归一化方法。比如图像处理中，将RGB图像转换为灰度图像后将其值限定在[0 255]的范围。

## 梯度下降GD

梯度下降主要有三种

### 1.BGD

BGD 采用整个训练集的数据来计算 cost function 对参数的梯度。

**缺点：**

**由于这种方法是在一次更新中，就对整个数据集计算梯度，所以计算起来非常慢，遇到很大量的数据集也会非常棘手，而且不能投入新数据实时更新模型。**

### 2.SGD

和 BGD 的一次用所有数据计算梯度相比，SGD 每次更新时对每个样本进行梯度更新，对于很大的数据集来说，可能会有相似的样本，这样 BGD 在计算梯度时会出现冗余，而**SGD 一次只进行一次更新，就没有冗余，而且比较快，并且可以新增样本。**

随机梯度下降是通过每个样本来迭代更新一次，如果样本量很大的情况，那么可能只用其中部分的样本，就已经将theta迭代到最优解了，对比上面的批量梯度下降，迭代一次需要用到十几万训练样本，一次迭代不可能最优，如果迭代10次的话就需要遍历训练样本10次。

**缺点是SGD的噪音较BGD要多，使得SGD并不是每次迭代都向着整体最优化方向**。**所以虽然训练速度快，但是准确度下降，并不是全局最优**。**虽然包含一定的随机性，但是从期望上来看，它是等于正确的导数的。**

**缺点：**

**SGD 因为更新比较频繁，会造成 cost function 有严重的震荡。**

**BGD 可以收敛到局部极小值，当然 SGD 的震荡可能会跳到更好的局部极小值处。**

**当我们稍微减小 learning rate，SGD 和 BGD 的收敛性是一样的。**

### 3.MBGD

**MBGD 每一次利用一小批样本，即 n 个样本进行计算，这样它可以降低参数更新时的方差，收敛更稳定，另一方面可以充分地利用深度学习库中高度优化的矩阵操作来进行更有效的梯度计算。**

**超参数设定值:  n 一般取值在 50～256**

**缺点：（两大缺点）**

1. **不过 Mini-batch gradient descent 不能保证很好的收敛性，learning rate 如果选择的太小，收敛速度会很慢，如果太大，loss function 就会在极小值处不停地震荡甚至偏离。（有一种措施是先设定大一点的学习率，当两次迭代之间的变化低于某个阈值后，就减小 learning rate，不过这个阈值的设定需要提前写好，这样的话就不能够适应数据集的特点。）**对于非凸函数，还要避免陷于局部极小值处，或者鞍点处，因为鞍点周围的error是一样的，所有维度的梯度都接近于0，SGD 很容易被困在这里。（**会在鞍点或者局部最小点震荡跳动，因为在此点处，如果是训练集全集带入即BGD，则优化会停止不动，如果是mini-batch或者SGD，每次找到的梯度都是不同的，就会发生震荡，来回跳动。**）
2. SGD对所有参数更新时应用同样的 learning rate，如果我们的数据是稀疏的，**我们更希望对出现频率低的特征进行大一点的更新。LR会随着更新的次数逐渐变小。**