数据预处理的基本概念、意义和通用方法

1. 基本概念:

1) 数据预处理过程中的归一化、标准化和中心化的基本概念。

- **归一化：** 1）把数据变成(0， 1)或者（1,1）之间的小数。主要是为了数据处理方便提出来的，把数据映射到 0～1 范围之内处理，更加便捷快速。2）把有量纲表达式变成无量纲表达式，便于不同单位或量级的指标能够进行比较和加权。归一化是一种简化计算的方式，即将有量纲的表达式，经过变换，化为无量纲的表达式，成为纯量。

- **标准化：**在机器学习中，我们可能要处理不同种类的资料，例如，音讯和图片上的像素值，这些资料可能是高维度的，资料标准化后会使每个特征中的数值平均变为 0(将每个特征的值都减掉原始资料中该特征的平均)、标准差变为 1，这个方法被广泛的使用在许多机器学习算法中(例如：支持向量机、逻辑回归和类神经网络)。

- **中心化：**平均值为 0，对标准差无要求。

2）数据预处理过程中的归一化、标准化和中心化的基本区别?

- **归一化和标准化的区别：**归一化是将样本的特征值转换到同一量纲下把数据映射到[0,1]或者[-1,1]区间内，仅由变量的极值决定，因区间放缩法是归一化的一种。标准化是依照特征矩阵的列处理数据，其通过求 z-score 的方法，转换为标准正态分布，和整体样本分布相关，每个样本点都能对标准化产生影响。它们的相同点在于都能取消由于量纲不同引起的误差；都是一种线性变换，都是对向量 X 按照比例压缩再进行平移。

- **标准化和中心化的区别：**标准化是原始分数减去平均数然后除以标准差，中心化是原始分数减去平均数。 所以一般流程为先中心化再标准化。

- 无量纲：我的理解就是通过某种方法能去掉实际过程中的单位，从而简化计算。

3） 为什么要归一化/标准化?

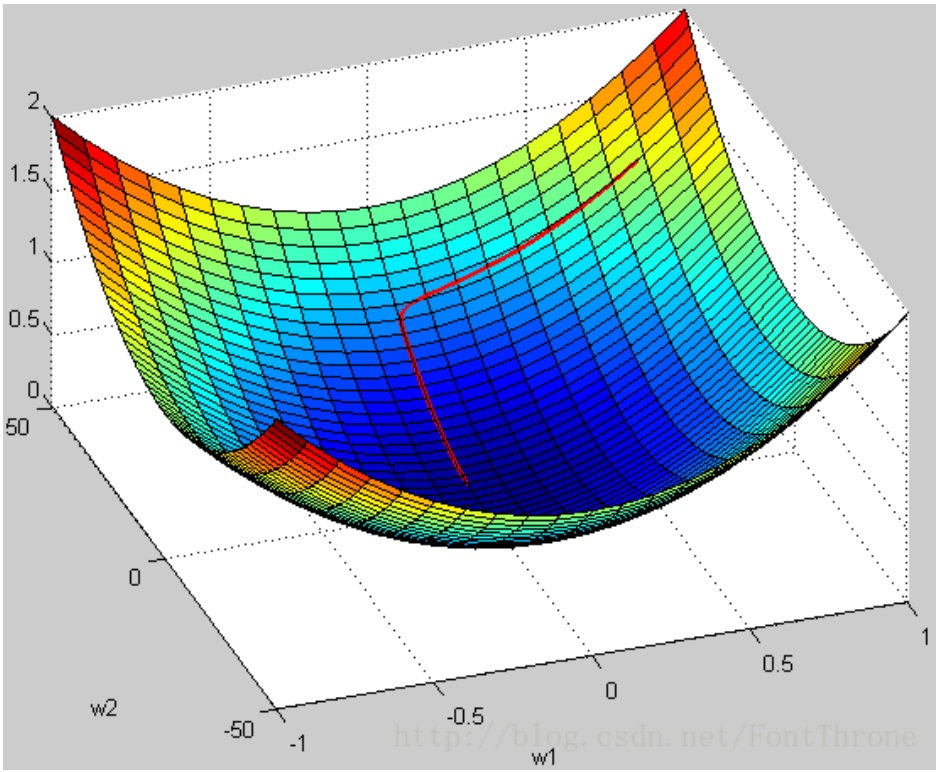
如前文所说，归一化/标准化实质是一种线性变换，线性变换有很多良好的性质，这些性质决定了对数据改变后不会造成“失效”，反而能提高数据的表现，这些性质是归一化/标准化的前提。比如有一个很重要的性质：线性变换不会改变原始数据的数值排序。

2. 意义:

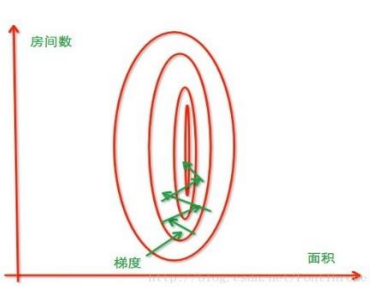
1. 求解需要

比如在 SVM 中处理分类问题是又是需要进行数据的归一化处理,不然会对准确率产生很大的影响,具体点说,比如避免出现因为数值过大导致 c,g 取值超过寻优范围。除此之外,最明显的是在神经网络中的影响,主要有四个层面:

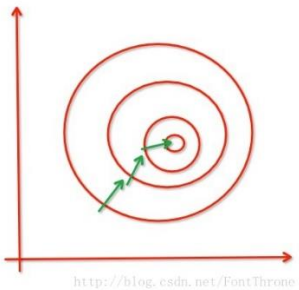
- 有利于初始化的进行
- 避免给梯度数值的更新带来数值问题
- 有利于学习率数值的调整
- 搜索轨迹:加快寻找最优解速度



2. 加快寻找最优解(加快收敛速度)



图：没有归一化前,寻找最优解的过程



图：归一化的过程

3. **无量纲化(业务上需求上的):**

指去除数据的单位限制,将其转化为无量纲的纯数值,便于不同单位或者量级的指标能够进行和加权。比如身高与体重,房子数量与收入等.

4. **数值问题**

不归一化的数值,比如浮点数可能会产生数值不相等的问题。

5. **数值范围减小对许多算法在纯粹的数值计算上都有一定加速作用(个人看法,虽然影响不大,但效果还是有的)**

3. 常用公式:

1. **min-max 标准化(Min-max normalization) ： 归一化**

又名离差标准化,是对原始数据的线性转化,公式如下:

$$x^* = \frac{x - min}{max - min}$$

含义: max : 样本最大值; min: 样本最小值;
问题: 当有新数据加入时需要重新进行数据归一化

2. **z-score 标准化(zero-mean normalization)：标准化**

又名标准差标准化,归一化后的数据呈正态分布,即均值为零,标准差为一公式如下:

$$x^* = \frac{x - \mu}{\sigma}$$

其中 μ 为所有样本数据的均值，σ 为所有样本数据的标准差。与离差标准化的不同之处在于，离差标准化仅仅对原数据的的方差与均差进行了倍数缩减，而标准差标准化则使标准化的数据方差为一。这对许多的算法更加有利，但是其缺点在于假如原始数据没有呈高斯分布，标准化的数据分布效果并不好。

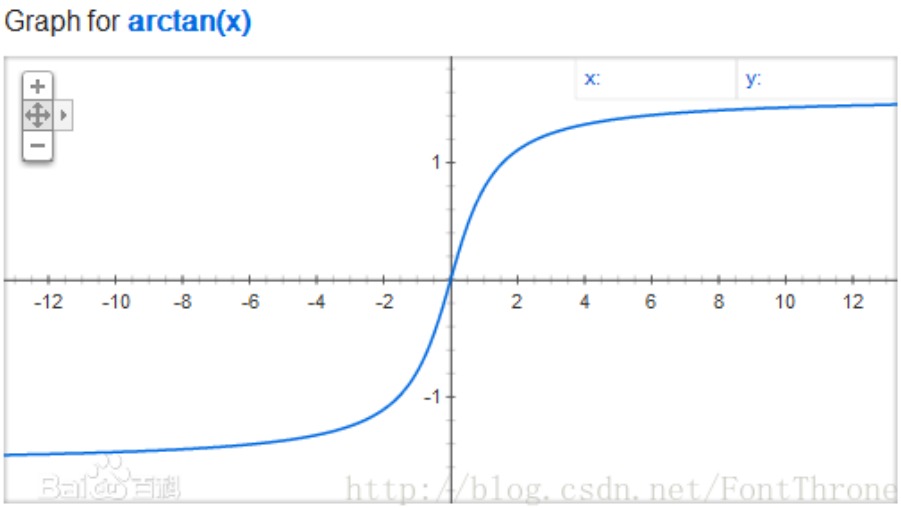
3. **atan 反正切函数标准化**

公式如下:

$$x^* = atan(x) * 2/\pi$$

问题: 数据必须大于零,大于零的函数将会被映射到[-1,0]上

atan 函数图像如下:



4. log 函数标准化

公式如下:

$$x^* = log_{10}(x),$$

问题:

a . 数据必须大于等于一

b. 如果数值大于 10**10(十的十次方),那么映射的数据将大于一

解决问题 b 的方案,改变公式以类似于 " min-max 标准化的方式 ", 如下:

$$x^* = log_{10}(x)/log_{10}(max)$$

max: 样本最大值

通过 "/log10(max)" 值得方式,可以保证所有样本能够正确的映射到[0,1]空间。

4. 文献参考

[1]. 归一化 （Normalization）、标准化 （Standardization）和中心化/零均值化 （Zero-centered）
https://www.jianshu.com/p/95a8f035c86c

[2]. 机器学习(补充知识)之归一化与标准化的概念和区别
https://blog.csdn.net/Zkangsen/article/details/90524076

[3]. 数据标准化和归一化的区别
https://www.cnblogs.com/yang520ming/p/9436954.html