参考https://mp.weixin.qq.com/s/N4u3bKf4KwDsRxaeU4rvRg

无论在机器学习还是深度领域中,损失函数都是一个非常重要的知识点。损失函数（Loss Function）是用来估量模型的预测值 f(x) 与真实值 y 的不一致程度。我们的目标就是最小化损失函数，让 f(x) 与 y 尽量接近。通常可以使用梯度下降算法寻找函数最小值。损失函数有许多不同的类型，没有哪种损失函数适合所有的问题，需根据具体模型和问题进行选择。一般来说，损失函数大致可以分成两类：回归（Regression）和分类（Classification）。

回归模型中的三种损失函数包括：

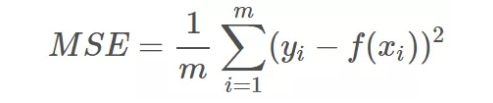
均方误差（Mean Square Error MSE）

平均绝对误差（Mean Absolute Error，MAE）、

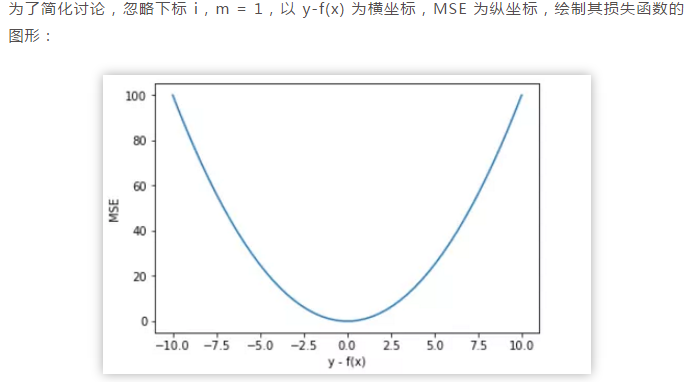
Huber Loss

1. **均方误差（Mean Square Error，MSE）**

均方误差指的就是模型预测值 f(x) 与样本真实值 y 之间距离平方的平均值。其公式如下所示：

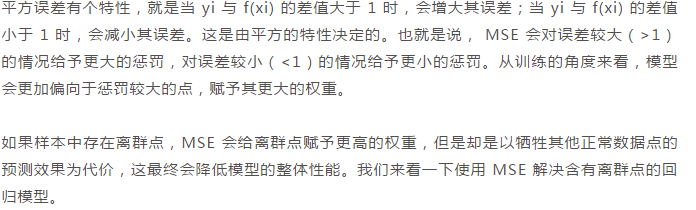


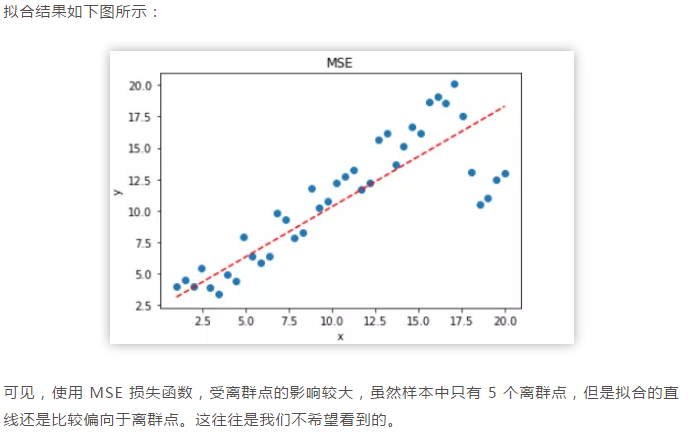
其中，yi 和 f(xi) 分别表示第 i 个样本的真实值和预测值，m 为样本个数。



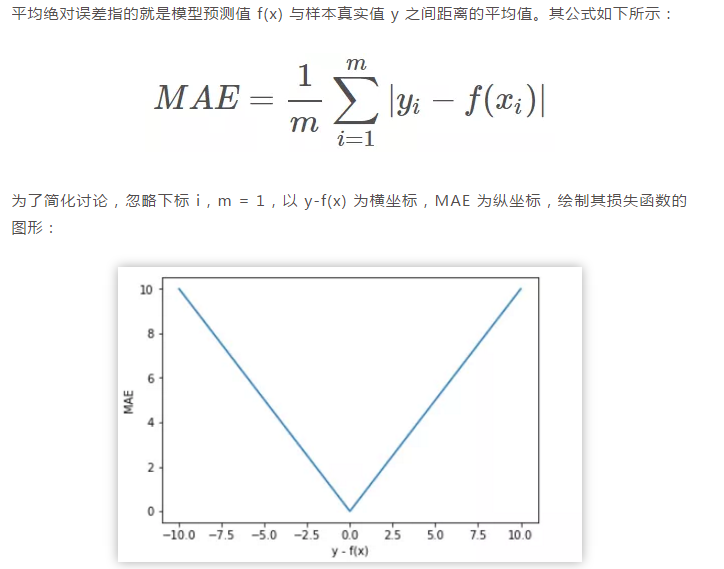
MSE 曲线的特点是光滑连续、可导，便于使用梯度下降算法，是比较常用的一种损失函数。而且，MSE 随着误差的减小，梯度也在减小，这有利于函数的收敛，即使固定学习因子，函数也能较快取得最小值，从图中可以观察到随着损失值的减小，函数的梯度也在不断的下降。

缺点：



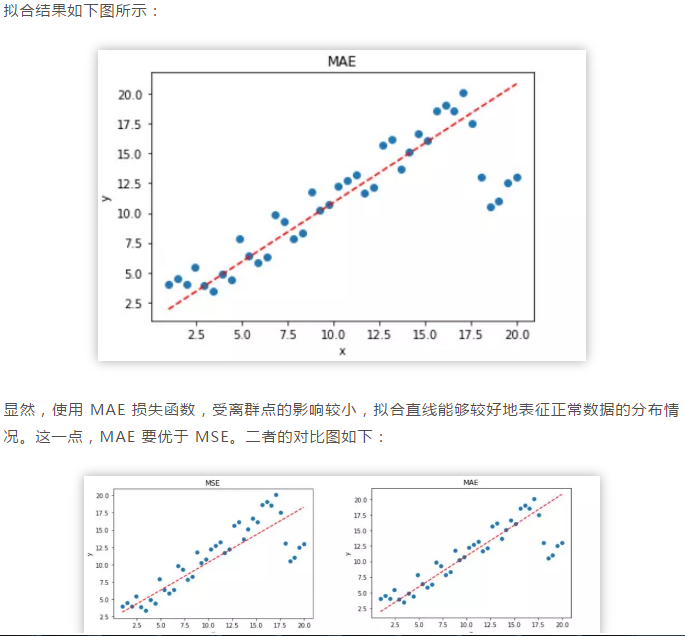


1. **平均绝对误差（Mean Absolute Error，MAE）**



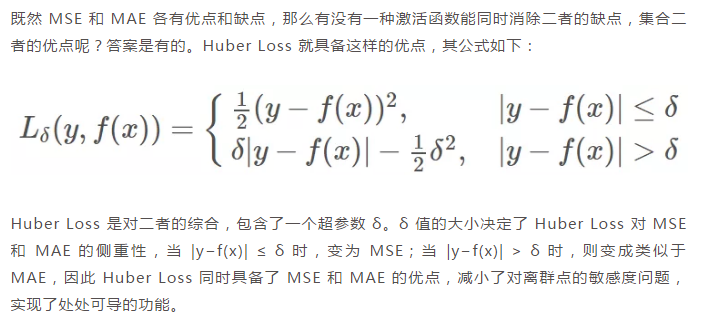
直观上来看，MAE 的曲线呈 V 字型，连续但在 y-f(x)=0 处不可导，计算机求解导数比较困难。而且 MAE 大部分情况下梯度都是相等的，这意味着即使对于小的损失值，其梯度也是大的。这不利于函数的收敛和模型的学习。

值得一提的是，MAE 相比 MSE 有个优点就是 MAE 对离群点不那么敏感，更有包容性。因为 MAE 计算的是误差 y-f(x) 的绝对值，无论是 y-f(x)>1 还是 y-f(x)<1，没有平方项的作用，惩罚力度都是一样的，所占权重一样。针对 MSE 中的例子，我们来使用 MAE 进行求解，看下拟合直线有什么不同。

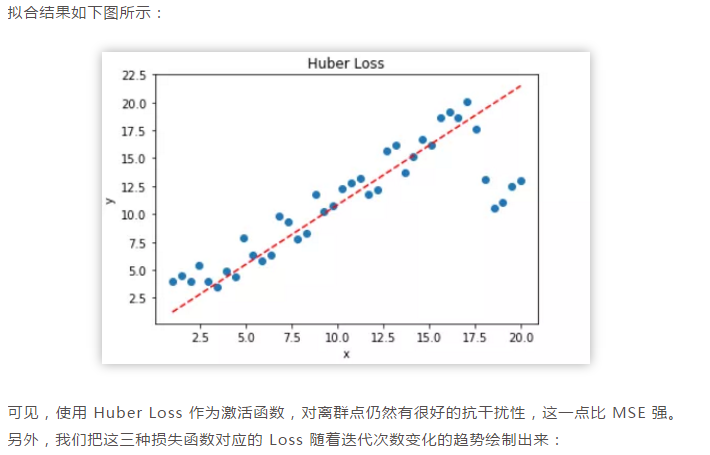


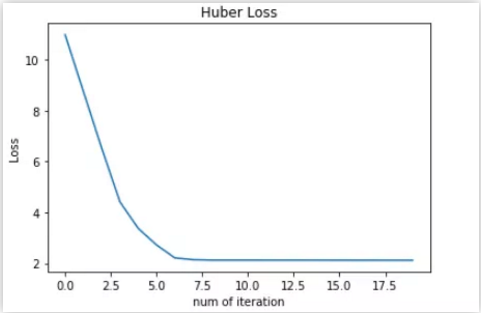
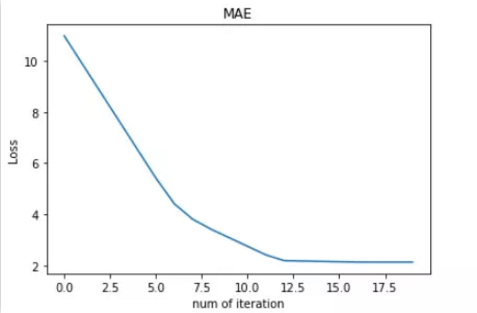
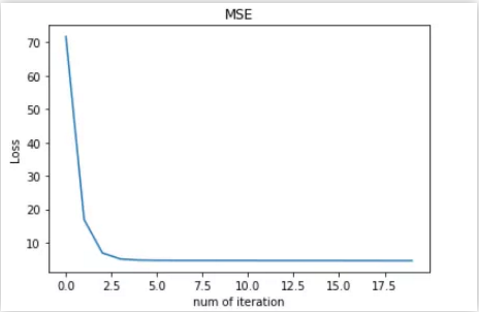
实际应用中，我们应该选择 MSE 还是 MAE 呢？从计算机求解梯度的复杂度来说，MSE 要优于 MAE，而且梯度也是动态变化的，能较快准确达到收敛。但是从离群点角度来看，如果离群点是实际数据或重要数据，而且是应该被检测到的异常值，那么我们应该使用MSE。另一方面，离群点仅仅代表数据损坏或者错误采样，无须给予过多关注，那么我们应该选择MAE作为损失。

1. **Huber Loss**

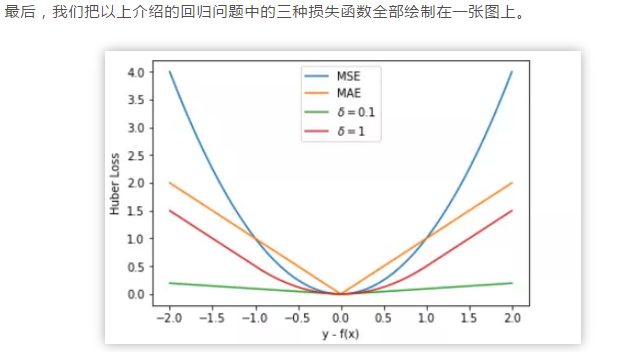


利用了均方误差在接近极值点的优势，快速收敛，处处可导，在误差较大的区域则选用近似MAE的做法，使得离群点的影响更小



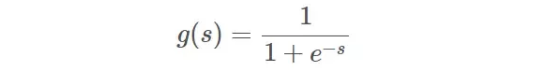


对比发现，MSE 的 Loss 下降得最快，MAE 的 Loss 下降得最慢，Huber Loss 下降速度介于 MSE 和 MAE 之间。也就是说，Huber Loss 弥补了此例中 MAE 的 Loss 下降速度慢的问题，使得优化速度接近 MSE。



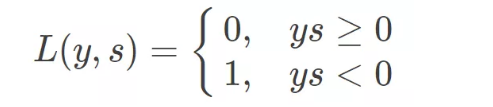
分类损失函数

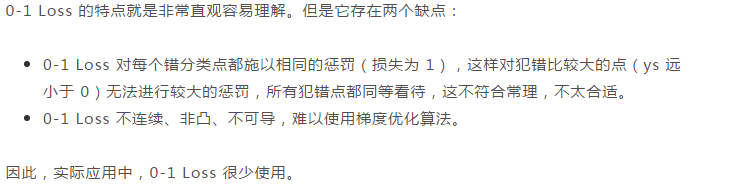
在讨论分类问题的损失函数之前，我想先说一下模型的输出 g(s)。一般来说，二分类机器学习模型包含两个部分：线性输出 s 和非线性输出 g(s)。其中，线性输出一般是模型输入 x 与 参数 w 的乘积，简写成：s = wx；非线性输出一般是 Sigmoid 函数，其表达式如下：



经过 Sigmoid 函数，g(s) 值被限定在 [0,1] 之间，若 s ≥ 0，g(s) ≥ 0.5，则预测为正类；若 s < 0，g(s) < 0.5，则预测为负类。

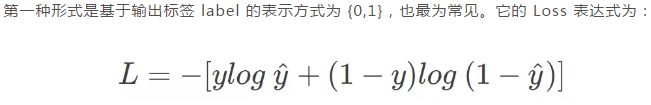
**0-1 Loss**





1. **Cross Entropy Loss**

Cross Entropy Loss 是非常重要的损失函数，也是应用最多的损失函数之一。二分类问题的交叉熵 Loss 主要有两种形式，下面分别详细介绍。



Y为样本真实值，y’为算法预测值，Y为正样本那么 1-Y为负样本，基于极大似然