机器学习基础知识集锦

损失函数

- 均方误差(MSE)，也称为L2损失

$$MSE = \frac{\sum_{i=1}^{n}(y_i - f(x))^2}{n}$$

  均方误差问题在于对离群值比较敏感，一旦训练样本存在离群值对整个Loss的影响比较大。

- 平均绝对误差(MAE)，也称为L1损失函数

$$MAE = \frac{\sum_{i=0}^{n}|y_i - f(x)|}{n}$$

- Huber损失函数

$$Huber = \begin{cases} \frac{1}{2}(y - f(x))^2 & if |y - f(x)| <= \delta \\ \delta|y - f(x)| - \frac{1}{2}\delta^2 & otherwise \end{cases}$$

  可以看到HuberLoss结合了MSE和MAE两者，因此HuberLoss可以避免MSE对离群值过于敏感的问题，同时也避免MAE导数不连续导致寻找最优解低效的问题。

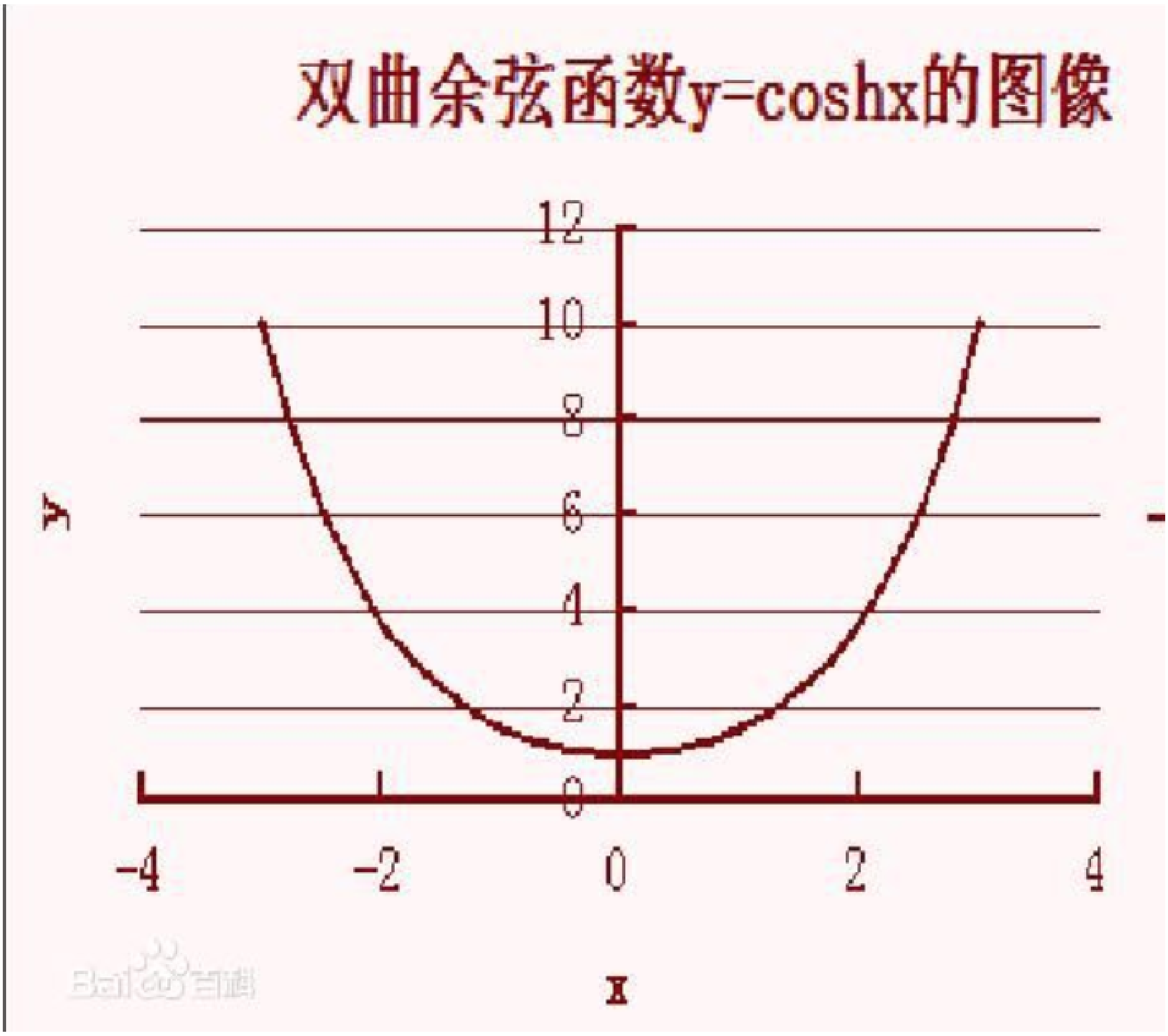
- Log-Cosh Loss函数

$$L(y, y^p) = \sum_{i=1}^{n} log(cosh(f(x) - y_i))$$

  双曲余弦函数公式

$$cosh(x) = \frac{e^x + e^{-x}}{2}$$

  双曲余弦函数图像



Log-Cosh Loss函数的优点: 对于小的x，其大约等于$\frac{x^2}{2}$，对于大的x，其大约等于$abs(x) - log2$。因此可以看到该函数和HuberLoss的作用是比较相似的，可以避免离群值的影响，又可以避免MAE不连续导致寻找最优解低效的问题。并且其处处二阶可导。

- 分位数损失函数(Quantile Loss)

参考文章: https://yq.aliyun.com/articles/602858?utm_content=m_1000002415

距离度量

- 欧式距离

$$Distance = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- 闵可夫斯基距离

$$Distance = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

当p为1时即为曼哈顿距离，当p=2时即为欧式距离。

- 互信息

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$

- 余弦相似度

$$Cos(x,y) = \frac{x \cdot y}{|x||y|}$$

- 皮尔逊相关系数

$$\rho(X;Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

其中$Cov(X,Y)$为X、Y的协方差，协方差的计算公式为：

$$Cov(X,Y) = E((X - E(X))(Y - E(Y)))$$

$\sigma(X)$和$\sigma(Y)$分别为X和Y的标准差。

- Jaccard相关系数

$$J = \frac{X \cap Y}{X \cup Y}$$