4. 模型优化

原始数据采用的是隐式评分的方式，点击量从 1 至 40 万不等，数据的标准差非常地大，会使模型不好拟合。先统计一下数据的分布情况（图一所示），以及比例分布（表 2，表 3），并将点击数转换为星级评价，从而进行显式地训练。

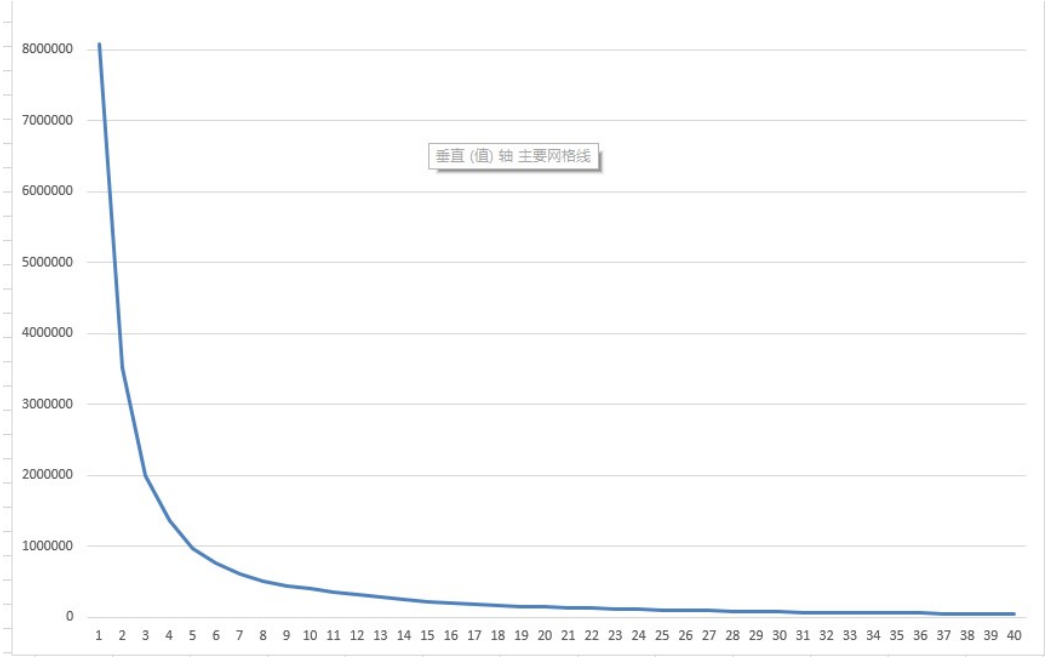

图 1 点击数-点击数的数量曲线

表 1 数据分布表	
点击次数	占所有数据的比例
1-2	47.80%
1-7	71.25%
1-15	82.69%
1-30	90.25%
表 2 点击数与星级的对应表	
点击次数	对应星级
1-2	1
3-7	2
8-15	3
16-30	4
30-N	5

这样做的好处是：将所有的数据变成了 1-5 星，训练的时候，数据减 3.0 可以得到[-2，+2]区间的数据。这是一种变相的特征值缩放（Feature Scaling）的方法，可以提高优化算法的性能。