凯读投资笔试思路

崔晏菲

1. 思路：

首先最基本的，要去掉过期的尾部数据。

其次，若去掉尾部数据之后仍然内存不够，那么就将已存储的数据中时间间隔最近的两个数据点做加权平均，并记录下它们所占的权重。

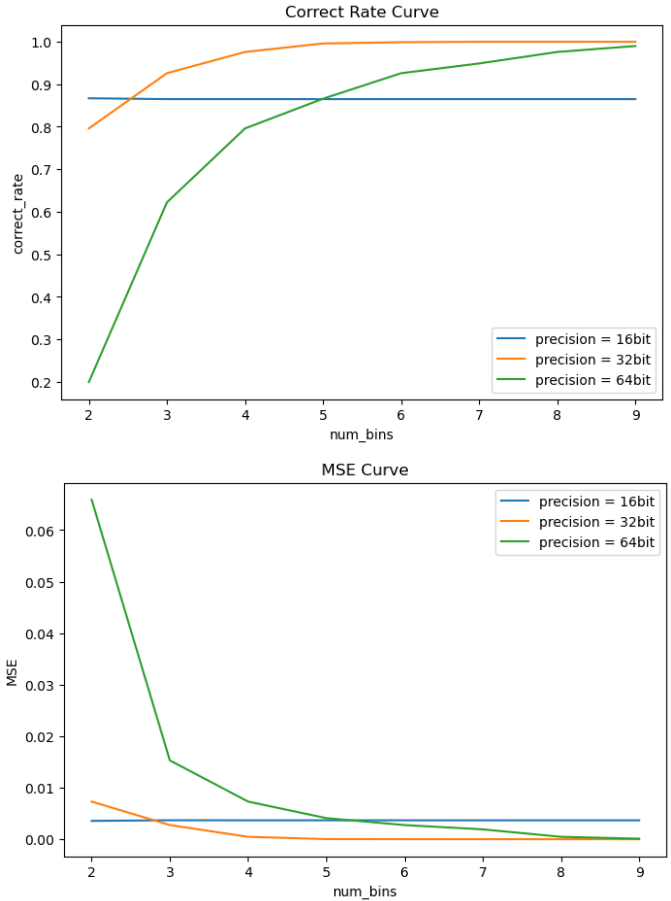
这样，每个数据点在内存中就需要保存三个属性：时间点、值和权重

2. 优化：

注意到，题目中提到存储空间是有限的，因此在有限的空间中对数据如何进行压缩是非常重要的。而在 python 中，默认 float 都是 64 位的，因此如果将存储的数据精度降低，就可以存储更多的数据，从而提升结果的精确度。因此我可以用 float32 甚至 float16 来存储时间点和值，权重使用 uint8 就可以。

3. 结果：

下图是我在随机生成的 1000 个实验数据上测试，得到的正确率和误差结果：



可见，当使用 float32 时，兼顾了数据精度和存储空间的平衡，仅需要 4 个 num_bins 就可以达到 100%的准确率。而如果使用 float16，num_bins 仅为 2

就可以达到最优准确率，高达 88.4%，且 MSE 仅为 0.008，这在绝大部分情况下都足够使用了。

具体的代码和运行结果见 code.ipynb