**基本概念:**

浮点：小数点非固定的数，可表示数据范围较广，整数，小数都可表示,例如float，double.

定点：小数点固定，可表示整数，小数。例如int本质是小数点位于末尾的32位定点数；

**为什么要量化:**

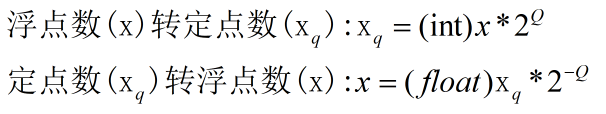
float占据4个字节,如果全部用float表示，一是占用空间大，二是读取效率不高。若将浮点数转换为定点数表示，在接受精度损失的前提下，可显著减少存储空间,提高运算效率.

对于计算机来说，浮点定点的概念是看不见的，因为它只能看到：0…00001110，至于它表示多少，是逻辑层面的设置。你如果让它是int那就按照int表示法对每个位赋予意义，如果你让它是float就按照float表示法赋予意义。

**量化过程:**

Milvus中IVF\_SQ8是把4字节float类型量化成1字节int类型.

相互转换的公式如下:



Q:量化系数,根据需求自定义,其表示的最大精度为1 / 2Q

**对0.775的量化示例:**

若Q=3, 量化后x1=(int)0.775\*23 =6,存储方式如下:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |

转换为浮点数 x2=6\*2-3=0.75, 精度损失=0.775-0.75=0.025

若Q=4, 量化后x1=(int)0.775\*24 =12,存储方式如下:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |

转换为浮点数 x2=12\*2-4=0.75, 精度损失=0.775-0.75=0.025

**对3.154的量化示例:**

若Q=3, 量化后x1=(int)3.154\*23 =25,存储方式如下:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |

转换为浮点数 x2=25\*2-3=3.125, 精度损失=3.154-3.125=0.029

若Q=4, 量化后x1=(int)3.154\*24 =50,存储方式如下:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |

转换为浮点数 x2=50\*2-4=3.125, 精度损失=3.154-3.125=0.029