LSH检索

一、实验题目

在corel数据集上实现iDistance索引或LSH索引(二者选一即可)，并分别进行近邻搜索，查询数据集前1000点，查找前10个最近邻，统计搜索算法的性能(召回率，准确率，时间)。

二、实验原理

LSH主要用来解决高维空间中点的近似最近邻搜索问题，LSH将原始空间中的点嵌入到Hamming空间中，即原始空间中点的表达形式转换成Hamming空间中点的表达形式，原始空间中的距离度量转换成Hamming空间中的距离度量。这样，原始空间的e-NNS问题就转变成Hamming空间中的e--KNN问题了。

通过映射，距离较近的点映射到同一个位置的概率大，距离较远的点映射到同一个位置的概率小。

三、实验步骤

1、数据投影，通过构建y=ax+b，将数据投影到一维空间，这里我将设定times次投影都产生的数字，即都投影到同一个线段上才可以认为两者相似，然后将相似的数据下标存起来，放到相应的list里

|  |
| --- |
|  |
|  |

2、进行上述操作不断重复，得到许多相似的数据下标，排序，取前11个数据（一个是数据本身）。

|  |
| --- |
|  |
|  |

四、实验结果

设定桶宽0.15，and操作10组，or操作1200组时：

|  |
| --- |
|  |

设定桶宽0.15，and操作10组，or操作900组时：

|  |
| --- |
|  |