本周学习内容：机器学习之K-邻近算法

K-邻近算法（K-nearest neighbor）：

由Cover T和Hart P提出的属于基本分类方法。工作原理：有一个样本数据集合，并且每个数据都含有标签，即知道每个数据与所属分类的对应关系。当输入新的没有标签的数据时，将新的数据的每个特征与样本集中数据相应特征进行比较，然后算法提取样本最相似的数据的分类标签。一般来说，只选取样本数据集中前K个最相似的数据，然后选取前K个数据中出现次数最多分类作为新数据的分类。

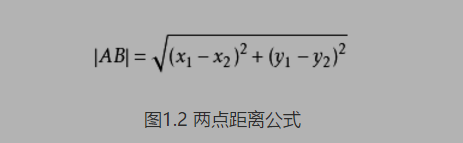
例如，下面是样本数据和标签



那么依据及经验，给出一部电影的打斗镜头和接吻镜头的数目，我们就可以判断这部电影的类型。K-邻近算法也是如此，只不过K-邻近算法是依靠已有的样本数据来判断。

距离度量：

当有新的数据时，我们如何比较它与样本数据的相似性。我们可以用空间距离来表示。以上述例子，有两个特征属性，即计算新的数据和样本数据的二维距离。



当新的数据为（101,20）时：

通过计算，我们可以得到如下结果：

(101,20)->动作片(108,5)的距离约为16.55

(101,20)->动作片(115,8)的距离约为18.44

(101,20)->爱情片(5,89)的距离约为118.22

(101,20)->爱情片(1,101)的距离约为128.69

计算可知，距离动作片（108,5）距离最近。如果以此为依据判断分类，那么就是最邻近算法而非K-邻近算法。

K-邻近算法描述如下：

1.计算已知类别数据集中的点与当前点之间的距离；

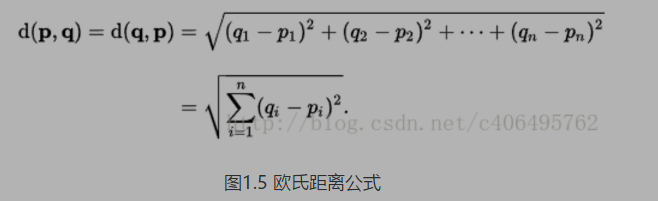
2.按照距离递增次序排序；

3.选取与当前点距离最小的k个点；

4.确定前k个点所在类别的出现频率；

5.返回前k个点所出现频率最高的类别作为当前点的预测分类。

另外地，如果样本特征属性超过两个，那么就用欧氏距离来计算距离



参考资料：

1. 机器学习实战.Petter Harrinon 著
2. CSDN博主Jack-Cui http://my.csdn.net/c406495762