## KNN算法

优点：精确度高，对离群的数据不敏感；no assumptions about data？

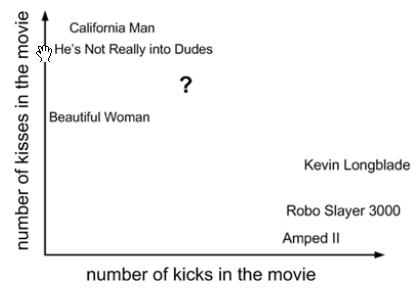
缺点：计算量大；占用内存多。

从哪里反映出缺点和优点的呢？

KNN原理。当前有一批TrainingSet，然后给出一个目标数据，采用下面的方式对目标数据进行分类：从Training Set中找出前K个数据，它们与目标数据最相近，或者说是相似，然后数一数这K个数据中有m个属于类型C1，有n个属于类型C2，并且m>n，最终认为目标数据是属于类型C1.

有以下问题。这个K取多大最好呢，取1个还是100个呢？怎么知道目标数据与训练数据集中的数据的相似性呢，最终如何反映到数值上呢？

现在已经知道电影有两个分类：动作电影和爱情电影。通过数一数电影中kick和kiss的数量来确定这个电影是动作电影还是爱情电影，如果kick多的话就认为这个电影是动作电影，如果是Kiss次数多的话就认为是爱情电影。可见，影响电影分类的有两个变量kick数量和kiss数量，这两个变量可以在平面坐标轴上反映出来，如下图：



问号处的电影在图上能看出kiss数量更多，它与其他的爱情电影在距离上更近，所以认为它是爱情电影。

扩展一下。如果有多个变量影响分类，就比如上面的变量从kick和kiss的基础上又增加了一个—男女在同一帧上时的帧数，如何比较距离？同样，这种计算距离的方式具有普遍性，如果有三个变量的话，可以比较在三维空间中的距离大小，如果是4个变量的话，就比较4维空间中的距离，n个变量的话可以比较n维空间中的距离大小。在计算距离时使用几何中计算两点距离的方式，即平方和再开方，这样就把问题引入到数值上来了，就可以“算一算”两个电影的相似程度了。再深入详细一些：目标数据的在空间中的位置与其他已知数据的位置间的距离分别为a1,a2,a3…an，按从小到大的次序排序，然后就可以得到前k个最相似的分类了，再从k个分类中看哪个分类占据的量大，就认为目标数据是哪个分类。

随着Training Set变大，需要计算的数据也会变多了，随着影响分类的变量增多，计算量也会增大，所以计算量是很大的。

对于KNN算法，前期需要搜集大量的Training Set，这个工作量应该很大。现实中，如何搜集这种数据呢，肯定是需要拿到一个电影后从头到尾数一下kick和kiss的数量吧，其实往往收集到的数据并不能直接用于算法的计算，也不能直接当训练数据。

对于KNN算法，像其他分类方法那样，先找出最终分类的几种结果，然后再找影响结果的变量。

从第2.2.2.节可以看出前期的分析有很大作用。有时某个数据对其他数据的掩盖作用太大，比如某种变量的数据值的范围是从1000到2000，另一个变量的取值范围是1到2，这样的话，第一个变量对结果的影响太大，这会导致得到的结果只是由第一个变量影响的。为了避免这样的情况，需要对各个变量的数据进行加权，再进行计算，才会得到比较符合现实的结果。

接下来有一个问题，用什么方法对数值进行归一化呢，变量1的取值的量级是103，变量2的取值的量级是10-1，这时看起来第变量1的数值除以1000并对变量2的数值乘以1000就行了。其实这样做默认认为变量1和变量2对结果的影响力都占据着50%。最好的方法是不是加权呢？怎么知道权重呢？使用TrainingSet预测权重的过程就是使用KNN前的分析。其实在规格化数据时有一个统一的公式：newValue = (oldValue-min)/(max-min)。

使用已知数据的90%用于训练分类器，再使用余下的10%用于测试。这是一个使用数据的好办法。

识别手写数字的例子。如何把问题转换成可用数学方式解决的形式，这是最重要的，就像识别手写数字的问题，把图片上数字转换为像素的序列，看手写图片和训练图片的像素序列的吻合度来判断手写图片上的数字是哪个数字。对于这个问题，先要找分类结果，这里的分类结果是1到9 ，这9个数字；然后再找影响结果的变量，变量只有一个，就是像素序列，这佧序列是由0和1组成的。在书上的这个例子中，已经做了很多的假设和限制，比如手写图片和作为训练的图片都是32\*32像素的，所图片上像素的颜色不是白色就是黑色，或者说不是0就是1，这两个前提条件一约束，对于识别就简单的多了。考虑一下，如果图片的大小不一样怎么办，如果图片上有噪音点（彩色点）怎么办？图片上的数字不只是一个怎么办呢？

KNN算法在计算距离时是否会用得到余弦夹角的定理呢？欧式距离：一个变量认为是一维空间的距离度量，2个变量时就是2维，n个变量时是n维空间中距离的度量。余弦夹角：向量的余弦值。这两个方式哪个好呢？当初怎么没有想到这两个求距离的方式呢？

收集数据和建立计算距离的数学模型这两点比较重要。收集数据可以从网络上使用抓虫的方式收集原始数据，但是要对原始数据进行处理，处理数据是收集数据中比较困难的一环。建立计算用的数学模型也是很困难的，可能需要我有深厚的数学基础。

距离可以分多种：欧式距离，曼哈顿距离，余弦距离、汉明距离、编辑距离、还有切比雪夫距离。

K的选择。K值使用哪个数值呢，1还是10或者100？有人说是使用训练数据量n的开方值，这个值看起来是小于n的最大素数呢？理论依据是什么？

### 相关网址

<http://blog.csdn.net/jmydream/article/details/8644004>

<http://coolshell.cn/articles/8052.html>

<https://my.oschina.net/u/1412321/blog/194174>

<http://blog.csdn.net/suipingsp/article/details/41964713>

<http://taop.marchtea.com/07.01.html> 对各种距离解析

<https://core.ac.uk/download/pdf/41458752.pdf>

## decision tree