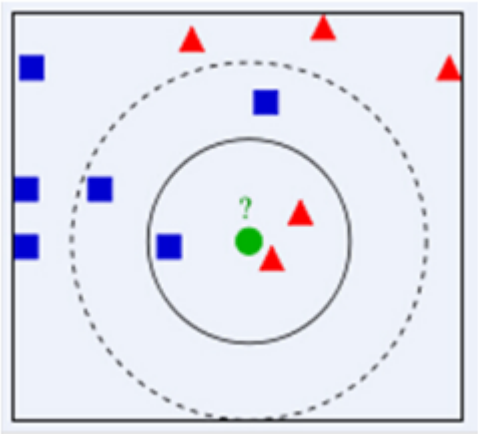
https://www.cnblogs.com/ybjourney/p/4702562.html

KNN是通过测量不同特征值之间的距离进行**分类**。

它的思路是：如果一个样本在特征空间中的k个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则该样本也属于这个类别，其中K通常是不大于20的整数。KNN算法中，所选择的邻居都是已经正确分类的对象。该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。

下面通过一个简单的例子说明一下：如下图，绿色圆要被决定赋予哪个类，是红色三角形还是蓝色四方形？如果K=3，由于红色三角形所占比例为2/3，绿色圆将被赋予红色三角形那个类，如果K=5，由于蓝色四方形比例为3/5，因此绿色圆被赋予蓝色



四方形类。

由此也说明了KNN算法的结果很大程度取决于K的选择。

在KNN中，通过计算对象间距离来作为各个对象之间的非相似性指标，避免了对象之间的匹配问题，在这里距离一般使用欧氏距离或曼哈顿距离：

欧式距离：$d(x,y)=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2}$，　　曼哈顿距离：$d(x,y)=\sqrt{\sum_{k=1}^{n}|x_k-y_k|}$

同时，KNN通过依据k个对象中占优的类别进行决策，而不是单一的对象类别决策。这两点就是KNN算法的优势。

接下来对KNN算法的思想总结一下：就是在训练集中数据和标签已知的情况下，输入测试数据，将测试数据的特征与训练集中对应的特征进行相互比较，找到训练集中与之最为相似的前K个数据，则该测试数据对应的类别就是K个数据中出现次数最多的那个分类，其算法的描述为：

1）计算测试数据与各个训练数据之间的距离；

2）按照距离的递增关系进行排序；

3）选取距离最小的K个点；

4）确定前K个点所在类别的出现频率；

5）返回前K个点中出现频率最高的类别作为测试数据的预测分类。

总结：1.网上查到的欧式距离的定义

计算公式 编辑

二维空间的公式

$\rho = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$,$|X| = \sqrt{x_2^2 + y_2^2}$. 其中，$\rho$ 为点 (x_2, y_2) 与点 (x_1, y_1) 之间的欧氏距离；$|X|$ 为点 (x_2, y_2) 到原点的欧氏距离。

三维空间的公式

$$\rho = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2},$$

$$|X| = \sqrt{x_2^2 + y_2^2 + z_2^2}.$$

n维空间的公式

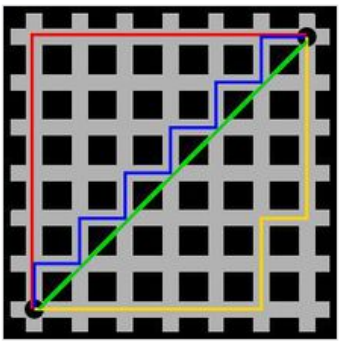
$$d(x, y) := \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

2.曼哈顿距离：

简介

编辑

名词解释



图中红线代表曼哈顿距离，绿色代表欧氏距离，也就是直线距离，而蓝色和黄色代表等价的曼哈顿距离。曼哈顿距离——两点在南北方向上的距离加上在东西方向上的距离，即d（i，j）=|xi-xj|+|yi-yj|。对于一个具有正南正北、正东正西方向规则布局的城镇街道，从一点到达另一点的距离正是在南北方向上旅行的距离加上在东西方向上旅行的距离，因此，曼哈顿距离又称为出租车距离。曼哈顿距离不是距离不变量，当坐标轴变动时，点间的距离就会不同。曼哈顿距离示意图在早期的计算机图形学中，屏幕是由像素构成，是整数，点的坐标也一般是整数，原因是浮点运算很昂贵，很慢而且有误差，如果直接使用AB的欧氏距离(欧几里德距离：在二维和三维空间中的欧氏距离的就是两点之间的距离），则必须要进行浮点运算，如果使用AC和CB，则只要计算加减法即可，这就大大提高了运算速度，而且不管累计运算多少次，都不会有误差。

3。切比雪夫距离

定义

编辑

在数学中，切比雪夫距离或是L∞度量，是向量空间中的一种度量，二个点之间的距离定义是其各坐标数值差绝对值的最大值。以数学的观点来看，切比雪夫距离是由一致范数（uniform norm）（或称为上确界范数）所衍生的度量，也是超凸度量（injective metric space）的一种。