# 常用距离

## 1、欧式径向基函数

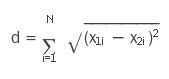
Radial basis function（径向基函数）

径向基函数是一个取值仅仅依赖于离原点距离的实值函数，也就是Φ（x）=Φ(‖x‖),或者还可以是到任意一点c的距离，c点称为中心点，也就是Φ（x，c）=Φ(‖x-c‖)。任意一个满足Φ（x）=Φ(‖x‖)特性的函数Φ都叫做径向基函数，标准的一般使用欧氏距离（也叫做欧式径向基函数），尽管其他距离函数也是可以的。在神经网络结构中，可以作为全连接层和ReLU层的主要函数。

那么核函数值约等于0。由于这个函数类似于高斯分布，因此称为高斯核函数，也叫做径向基函数(Radial Basis Function 简称RBF)。它能够把原始特征映射到无穷维。

## 2、欧式距离

欧式距离源自N维欧氏空间中两点x1,x2间的距离公式：



## 3、标准化欧式距离

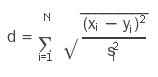
引入标准化欧式距离的原因是一个数据xi的各个维度之间的尺度不一样。

【对于尺度无关的解释】如果向量中第一维元素的数量级是100，第二维的数量级是10，比如v1=(100,10,30),v2 = (500,40)，则计算欧式距离

2018-03-02 10-38-54 的屏幕截图

可见欧式距离会给与第一维度100权重，这会影响第二维度的权重。对所有维度分别进行处理，使得各个维度分别满足标准正态分布。

还有一种对欧式距离的处理是均值化但没有归一化（Normalized），即



其中s2i是第i维度的方差（此处虽然举例只有X,Y两个点，但整个数据集中会有无数个点，根据数据集得到方差分布）

## 4、马氏距离

标准化欧式距离即是将集合X={xi}先进行归一化，映射到正太分布N(0,1)的区间：

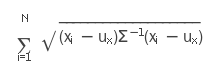
xipost=(xi−u)/s

其中u为均值，s2为方差。

其实，这就是马氏距离要做的事。马氏距离的结果也是将数据投影到N(0,1)区间并求其欧式距离。

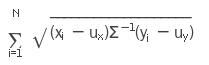
马氏距离又称为数据的协方差距离。它是一种有效的计算两个未知样本集的相似度的方法。与标准化欧氏距离不同的是它考虑到各种特性之间的联系。

假设ux为向量X={x1,x2,...,xN}的均值，uy为Y={y1,y2,...yN}的均值，Σ 是协方差矩阵，则x的马氏距离：



这个公式有问题，应该只有一个（x-u）

点X与Y的马氏距离：



其中Σ是X与Y的协方差矩阵

可见：

如果Σ是单位矩阵，则马氏距离退化成欧式距离；

如果Σ是对角矩阵，则称为归一化后的欧式距离。

首先抛出马氏距离的特点：

A）尺度无关

B）考虑进数据之间的联系

马氏距离可以通过协方差自动生成相应的权重，而使用逆则抵消掉这些权重。

最典型的就是根据距离作判别问题，即假设有n个总体，计算某个样品X归属于哪一类的问题。此时虽然样品X离某个总体的欧氏距离最近，但是未必归属它，比如该总体的方差很小，说明需要非常近才能归为该类。对于这种情况，马氏距离比欧氏距离更适合作判别。比如下图表示男女两类身高和体重的采样数据，假如男和女先验分布已知，可以用马氏距离判别新的数据。

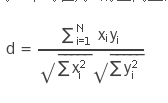
## 5、夹脚余弦距离（余弦相似性）

严格来讲余弦距离不是距离，而只是相似性。其他距离直接测量两个高维空间上的点的距离，如果距离为0则两个点“相同”，但余弦的结果为1，只能确定两者高度相似。

从几何上看，n维向量空间的一条线段作为底边和原点组成的三角形，其顶角大小是不确定的。对于两条空间向量，即使两点距离一定，他们的夹角余弦值也可以随意变化。

假设两用户只对两件商品评分，向量分别为(3,3)和(5,5)，这两位用户的认知其实是一样的，但是欧式距离给出的解显然没有余弦值合理。

相对于标准化后的欧式距离，余弦距离少了将数据投影到一个均值为0的区间里这一步骤。对于点X和点Y，其余弦距离：



余弦距离在给文本分类的词袋模型中使用，例如给一篇文章一共出现过6000个词；因此用一个6000维度的向量X表示这篇文章，每个维度代表字出现数目。另外一篇文章也恰好只出现了这6000字并用向量Y表示该文章，则这两篇文章相似度可以用余弦距离来测量。

**优点**：余弦距离根据向量方向重合度来判断向量相似度

**缺点**：没有对每个维度进行0均值化处理，如在图片分类中一张图片受光照影 响后，前后的余弦距离可能就非常大了。

## 6、汉明距离

汉明距离是两个等长字符串之间的汉明距离是两个字符串对应位置的不同字符的个数。比如：

1011101 与 1001001 为 2

2143896 与 2233796 是 3

可以把它看做将一个字符串变换成另外一个字符串所需要替换的字符个数。

此外，汉明重量是字符串相对于同样长度的零字符串的汉明距离，如：

11101 的汉明重量是 4。

所以两者间的汉明距离等于它们汉明重量的差a-b