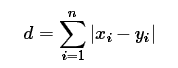
# [机器学习中常用的距离公式](https://www.iteblog.com/archives/2317.html)

## ****曼哈顿距离****

曼哈顿距离又称计程车几何距离或方格线距离，是由十九世纪的赫尔曼·闵可夫斯基所创词汇 ，为欧几里得几何度量空间的几何学之用语，用以标明两个点上在标准坐标系上的绝对轴距之总和。曼哈顿距离的正式意义为L1-距离或城市区块距离，也就是在欧几里得空间的固定直角坐标系上两点所形成的线段对轴产生的投影的距离总和。例如在平面上，坐标(x1,y1)的点P1与坐标(x2,y2)的点P2的曼哈顿距离为：



等价于下面的汇总式：

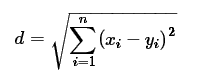


## 欧几里得距离

****欧几里得度量（euclidean metric）也称欧氏距离：**** 在数学中，欧几里得距离或欧几里得度量是欧几里得空间中两点间“普通”（即直线）距离。在欧几里得空间中，点x=(x1,x2,...,xn)和 y=(y1,y2,...,yn)之间的欧氏距离为:



等价于下面的汇总式：



## ****切比雪夫距离****

数学上，切比雪夫距离（Chebyshev distance）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为其各座标数值差的最大值。以p(x1,y1)和q(x2,y2)二点为例，其切比雪夫距离为

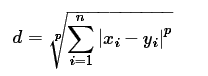


一般形式为：



## ****闵可夫斯基距离****

****闵可夫斯基距离或闵氏距离（Minkowski Distance）：****以俄罗斯数学家闵可夫斯基命名的距离；是欧式距离的推广，闵氏距离不是一种距离，而是一组距离的定义。其定义如下：



从上面公式可以看出：

* 当p=1时，就是曼哈顿距离
* 当p=2时，就是欧氏距离
* 当p→∞时，就是切比雪夫距离

## ****马氏距离****

****马氏距离(Mahalanobis distance)：**** 由印度统计学家马哈拉诺比斯(P. C. Mahalanobis)提出，表示数据的协方差距离。它是一种有效的计算两个未知样本集的相似度的方法。与欧氏距离不同的是它考虑到各种特性之间的联系（例如：一条关于身高的信息会带来一条关于体重的信息，因为两者是有关联的）并且是尺度无关的(scale-invariant)，即独立于测量尺度，如果协方差矩阵为单位矩阵，马氏距离就简化为欧式距离，如果协方差矩阵为对角阵，其也可称为正规化的马氏距离。 计算公式如下：  
对于一个均值为μ=(μ1,μ2,μ3,...μp)T ,协方差矩阵为Σ，其马氏距离为：



马氏距离也可以定义为两个服从同一分布并且其协方差矩阵为Σ 的随机变量 x⃗与y⃗ 的差异程度：



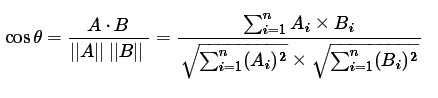
## ****汉明距离****

在信息论中，两个等长字符串之间的汉明距离（英语：Hamming distance）是两个字符串对应位置的不同字符的个数。换句话说，它就是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。例如：

* 1011101与1001001之间的汉明距离是2
* 2143896与2233796之间的汉明距离是3
* toned 与 roses 之间的汉明距离是3

## ****余弦相似度****

余弦相似性通过测量两个向量的夹角的余弦值来度量它们之间的相似性。0度角的余弦值是1，而其他任何角度的余弦值都不大于1；并且其最小值是-1。从而两个向量之间的角度的余弦值确定两个向量是否大致指向相同的方向。两个向量有相同的指向时，余弦相似度的值为1；两个向量夹角为90°时，余弦相似度的值为0；两个向量指向完全相反的方向时，余弦相似度的值为-1。这结果是与向量的长度无关的，仅仅与向量的指向方向相关。余弦相似度通常用于正空间，因此给出的值为0到1之间。给定两个属性向量， A和B，其余弦相似性θ由点积和向量长度给出，如下所示：

  
这里的Ai和Bi分别代表向量A和B的各分量

## ****杰卡德距离****

****杰卡德距离(Jaccard Distance) ：****它是杰卡德相似系数的补集，被定义为1减去Jaccard相似系数。而杰卡德相似系数(Jaccard similarity coefficient)，也称杰卡德指数(Jaccard Index)，是用来衡量两个集合相似度的一种指标。  
****Jaccard相似指数****用来度量两个集合之间的相似性，它被定义为两个集合交集的元素个数除以并集的元素个数。



****杰卡德距离****如下：

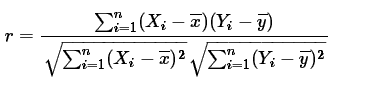


****性质：****  
1) 若A、B两个集合都为空，则J(A,B)=1

2) 0≤J(A,B)≤1

## ****皮尔森相关系数****

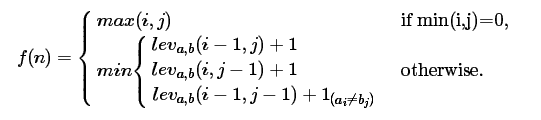
****皮尔森相关系数（Pearson correlation coefficient）：****也称皮尔森积矩相关系数(Pearson product-moment correlation coefficient) ，是一种线性相关系数。皮尔森相关系数是用来反映两个变量线性相关程度的统计量。相关系数用r表示，其中n为样本量，分别为两个变量的观测值和均值。r描述的是两个变量间线性相关强弱的程度。r的绝对值越大表明相关性越强。  
****计算公式：****



分子是两个集合的交集大小，分母是两个集合大小的几何平均值。是余弦相似性的一种形式

## ****编辑距离****

****编辑距离（Edit Distance）:****又称Levenshtein距离，是指两个字串之间，由一个转成另一个所需的最少编辑操作次数。许可的编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除一个字符。一般来说，编辑距离越小，两个串的相似度越大。俄罗斯科学家Vladimir Levenshtein在1965年提出这个概念。编辑距离越小的两个字符串越相似，当编辑距离为0时，两字符串相等。  
****计算公式：****



## ****K-L散度****

****K-L散度（Kullback-Leibler Divergence）：****即相对熵；是衡量两个分布(P、Q)之间的距离；越小越相似。  
****计算公式：****

