补充内容

各种距离定义及含义

主讲：蔡 波

武汉大学网络安全学院

1）欧氏距离

欧式距离是最容易值观理解的距离度量方法。

二维平面上点a(x_1,y_1)与b(x_2,y_2)间的欧氏距离:

$$d_{12}=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$$

三维空间点a(x_1,y_1,z_1)与b(x_2,y_2,z_2)间的欧氏距离:

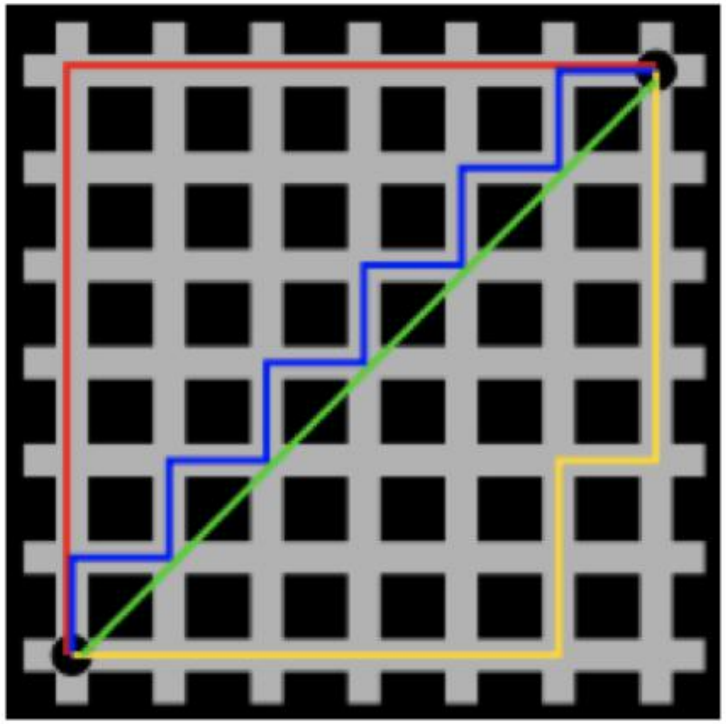
$$d_{12}=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2}$$

n维空间点a(x_{11},x_{12},...,x_{1n})与b(x_{21},x_{22},...,x_{2n})间的欧氏距离（两个n维向量）：

$$d_{12}=\sqrt{\sum_{k=1}^{n}(x_{1k}-x_{2k})^2}$$

2）曼哈顿距离

在曼哈顿街区要从一个十字路口开车到另一个十字路口，驾驶距离显然不是两点之前的直线距离。这个实际的驾驶距离就是"曼哈顿距离"。曼哈顿距离也称“城市街区距离”。



二维平面两点a(x_1,y_1)与b(x_2,y_2)间的曼哈顿距离：

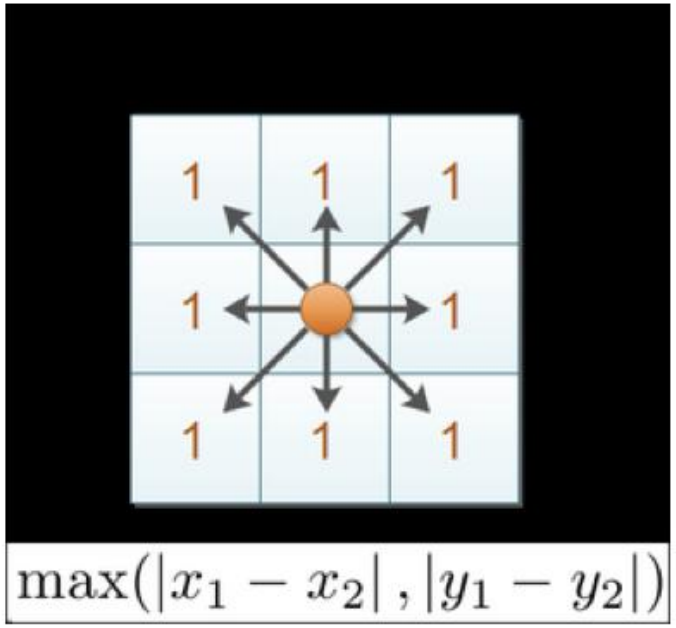
$$d_{12} = |x_1 - x_2| + |y_1 - y_2|$$

n维空间点a(x_{11},x_{12},...,x_{1n})与b(x_{21},x_{22},...,x_{2n})的曼哈顿距离：

$$d_{12} = \sum_{k=1}^{n} |x_{1k} - x_{2k}|$$

3）切比雪夫距离

国际象棋中，国王可以直行、横行、斜行，所以国王走一步可以移动到相邻8个方格中的任意一个。国王从格子(x1,y1)走到格子(x2,y2)最少需要走多少步？这个距离就叫做切比雪夫距离。



$$\max(|x_1 - x_2|, |y_1 - y_2|)$$

二维平面两点a(x_1,y_1)与b(x_2,y_2)间的切比雪夫距离：$d_{12} = \max(|x_1 - x_2|, |y_1 - y_2|)$

n维空间点a(x_{11},x_{12},…,x_{1n})与b(x_{21},x_{22},…,x_{2n})的切比雪夫距离：

$$d_{12} = \max(|x_{1i} - x_{2i}|)$$

棋盘上所有位置距f6位置的切比雪夫距离

	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1	♔	1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	

4）闵可夫斯基距离

闵氏距离不是一种距离，而是一组距离的定义，是对多个距离度量公式的概括性的表示。

两个n维变量a(x11,x12,...x1n)与b(x21,x22,...,x2n)间的闵可夫斯基距离定义为：

$$d_{12}=\sqrt[p]{\sum_{k=1}^{n}\left|x_{1k}-x_{2k}\right|^{p}}$$

其中p是一个变参数：

- p=1的时候，就是曼哈顿距离；
- p=2的时候，就是欧式距离；
- p→∞的时候，就是切比雪夫距离。

就是根据参数p的不同，闵氏距离可以表示某一种类/种的距离。

闵氏距离、曼哈顿距离、欧式距离和切比雪夫距离都存在明显的缺点

- 将各个分两的量纲，也就是“单位”相同看待了。
- 未考虑各个分量的分布（期望、方差等）可能是不同的。

5）标准化欧氏距离

标准化欧式距离是针对欧式距离的缺点而做的一种改进

思路：既然数据各维分两的分布不一样，那就先将各个分量都”标准化“到均值、方差等。

Sk表示各个维度的标准差

标准化欧氏距离公式：

$$d_{12} = \sqrt{\sum_{k=1}^{n} \left(\frac{x_{1k} - x_{2k}}{s_k}\right)^2}$$

如果将方差的倒数看成一个权重，也可以称之为加权欧式距离

6）余弦距离

几何中，夹角余弦可用来衡量两个向量方向的差异；机器学习中，借用这一概念来衡量样本向量之间的差异。

夹角余弦取值范围为[-1,1]。余弦越大表示两个向量的夹角越小，余弦越小表示两向量的夹角越大。当两个向量的方向重合时余弦取最大值1，当两个向量的方向完全相反余弦取最小值-1。

结果越趋近于1越正相关，越趋近于-1则越负相关，越趋近于0说越无相关。

相比距离度量，余弦相似度更加注重两个向量在方向上的差异，而非距离或长度上。公式如下：

$$sim(X,Y) = cos\theta = \frac{\vec{x} \cdot \vec{y}}{\|x\| \cdot \|y\|}$$

7）汉明距离

汉明距离是以理查德·卫斯里·汉明的名字命名的。在信息论中，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。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。

8）杰卡德距离

杰卡德相似系数：两个集合A和B的交集元素在A和B的并集所占的比例，称为两个集合的杰卡德相似系数，用符号J（A,B）表示：

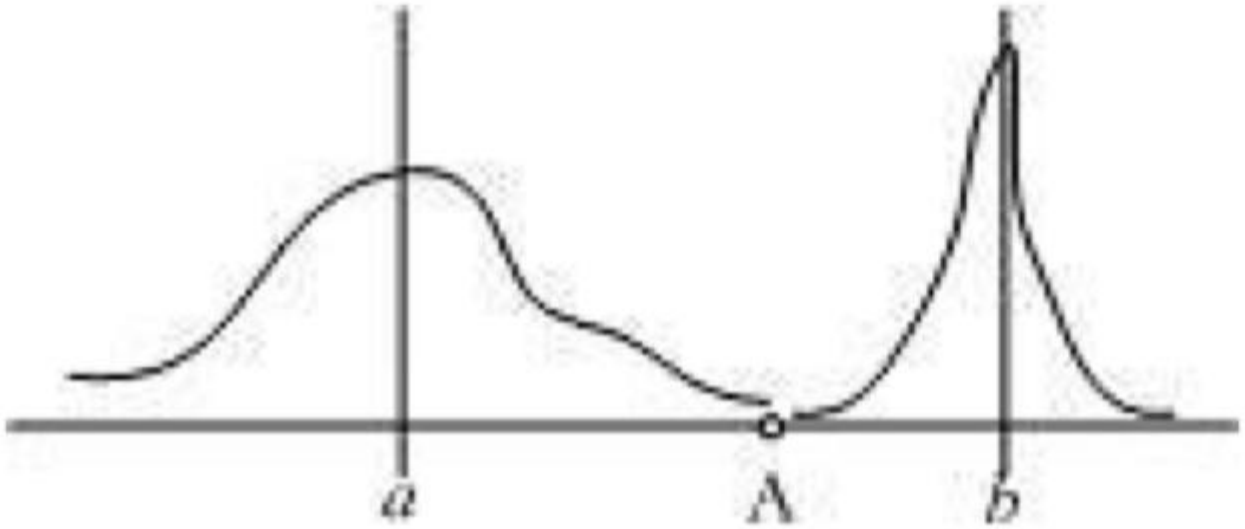
$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}$$

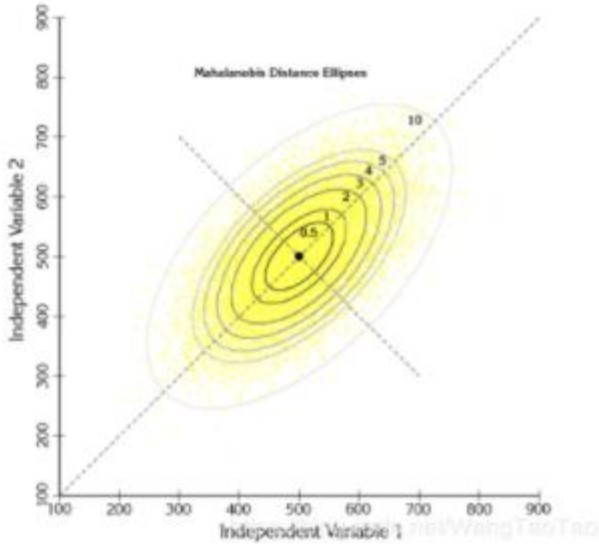
杰卡德距离：与杰卡德相似系数相反，用两个集合中的不同元素占所有元素的比例来衡量两个集合的区分度：

$$J_\delta(A,B)=1-J(A,B)=\frac{|A\cup B|-|A\cap B|}{|A\cup B|}$$

9）马氏距离

下图有两个正态分布图，它们的均值分别为a和b，但方差不一样，则图中的A点离哪个总体更近？或者说A有更大的概率属于谁？显然，A离左边的更近，A属于左边总体的概率更大，尽管A与a的欧式距离远一些。这就是马氏距离的直观解释。





马氏距离是一种基于样本分布的距离

马氏距离是由印度统计学家马哈拉诺比斯提出的，表示数据的协方差距离。它是一种有效的计算两个位置样本集的相似度的方法。

与欧式距离不同的是，它考虑到各种特性之间的联系，即独立于测量尺度。

马氏距离定义：设总体G为m维总体（考察m个指标），均值向量为μ=（μ1，μ2，... ...，μm，）`,协方差阵为∑=（σij），

则样本X=（X1，X2，... ...，Xm，）`与总体G的马氏距离定义为：

$$d^2(X,G)=(X-\mu)'\Sigma^{-1}(X-\mu)$$

$$当m=1时,d^2(x,G)=\frac{(x-\mu)'(x-\mu)}{\sigma^2}=\frac{(x-\mu)^2}{\sigma^2}$$

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∑的随机变量的差异程度：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矩阵，则其也可称为正规化的欧式距离。