1.欧式距离

2.曼哈顿距离

3.切比雪夫距离

4.闵科夫斯基距离

5.马氏距离

6.汉明距离

7.余弦相似度

8.编辑距离

9.K-L散度

1.欧式距离

欧氏距离，最常见的两点之间或多点之间的距离表示法，又称之为欧几里得度量，它定义于欧几里得空间中，如点 x = (x1,...,xn) 和 y = (y1,...,yn) 之间的距离为：

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

缺点：将样本的不同属性（各指标或各变量量纲）之间的差别等同看待，有时不能满足实际要求。欧式距离适用于向量各分量的度量标准统一的情况。

2.曼哈顿距离

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。

坐标（x1, y1）的点P1与坐标（x2, y2）的点P2的曼哈顿距离为：$|x_1-x_2|+|y_1-y_2|$,.

曼哈顿距离依赖坐标系统的转度，而非系统在坐标轴上的平移或映射，当坐标轴不同时，点间的距离就会不同。

$$d=\sum_{i=1}^{n}|x_i-y_i|$$

3.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$p(x_1,y_1)$和$q(x_2,y_2)$二点为例，其切比雪夫距离为

$$D_{Chebyshev}(p,q)=max(|x_2-x_1|,|y_2-y_1|)$$

一般形式为：

$$D_{Chebyshev}(p,q)=\max_i(|p_i-q_i|)=\lim_{k\to\infty}(\sum_{i=1}^{n}|p_i-q_i|^k)^{1/k}$$

4.闵科夫斯基距离

欧式距离的推广，是一组距离的定义。

$$d=\sqrt[p]{\sum_{i=1}^{n}|x_i-y_i|^p}$$

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$p\to\infty$时，就是切比雪夫距离

5.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$\mu=(\mu_1,\mu_2,\mu_3,\dots\mu_p)^T$,协方差矩阵为Σ ，其马氏距离为：

$$D_M(x)=\sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

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 $\vec{x}$与$\vec{y}$的差异程度：

$$d(\vec{x},\vec{y})=\sqrt{(\vec{x}-\vec{y})^T\Sigma^{-1}(\vec{x}-\vec{y})}$$

6.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

7.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：

$$\cos\theta=\frac{A\cdot B}{||A||\,||B||}=\frac{\sum_{i=1}^{n}A_i\times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

这里的 A_i和 B_i分别代表向量 A和 B的各分量

8.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：

$$f(n) = \begin{cases} max(i,j) & \text{if min(i,j)=0,} \\ min\begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

9.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：

$$D(P||Q) = \sum_{i=1}^{n} P(i) log \frac{P(i)}{Q(i)}$$