机器学习中的相似性度量

在做分类时常常需要估算不同样本之间的相似性度量(Similarity Measurement)，这时通常采用的方法就是计算样本间的“距离”(Distance)。采用什么样的方法计算距离是很讲究。

1. 欧氏距离(Euclidean Distance)

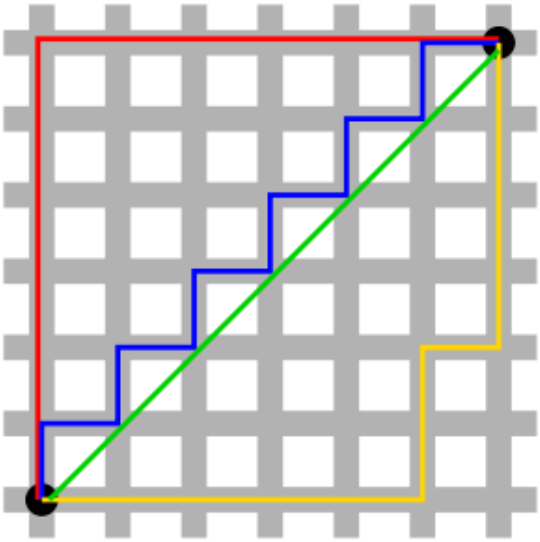
主要方式就是考虑点与点之间的距离，无论是几维的，但是我们可以两维的情况来估计 N 多维的情况。欧氏距离是最易于理解的一种距离计算方法，源自欧氏空间中两点间的距离公式。

两个 n 维向量 a(x11,x12,…,x1n)与 b(x21,x22,…,x2n)间的欧氏距离：

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

也可以用表示成向量运算的形式：$d_{12} = \sqrt{(a-b)(a-b)^T}$

2. 曼哈顿距离(Manhattan Distance)



从名字就可以猜出这种距离的计算方法了。想象你在曼哈顿要从一个十字路口开车到另外一个十字路口，驾驶距离是两点间的直线距离吗？显然不是，除非

你能穿越大楼。实际驾驶距离就是这个“曼哈顿距离”。而这也是曼哈顿距离名称的来源， 曼哈顿距离也称为**城市街区距离(City Block distance)**。

(1)二维平面两点 a(x1,y1)与 b(x2,y2)间的曼哈顿距离

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

(2)两个 n 维向量 a(x11,x12,…,x1n)与 b(x21,x22,…,x2n)间的曼哈顿距离

$$\mathrm{d}_{12} = \sum_{k=1}^{n} |\mathrm{x}_{1k} - x_{2k}|$$

3. 切比雪夫距离 (Chebyshev Distance)

国际象棋玩过么？国王走一步能够移动到相邻的 8 个方格中的任意一个。那么国王从格子(x1,y1)走到格子(x2,y2)最少需要多少步？自己走走试试。你会发现最少步数总是 max(| x2-x1 | , | y2-y1 |) 步 。有一种类似的一种距离度量方法叫切比雪夫距离。

(1)二维平面两点 a(x1,y1)与 b(x2,y2)间的切比雪夫距离

$$d_{12} = \max(|x_1 - x_2|, |y_1 - y_2|)$$

(2)两个 n 维向量 a(x11,x12,…,x1n)与 b(x21,x22,…,x2n)间的切比雪夫距离

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

这个公式的另一种等价形式是

$$\mathrm{d}_{12} = \lim_{k\to\infty}\left(\sum_{i=1}^{n}|x_{1i} - x_{2i}|^k\right)^{1/k}$$

看不出两个公式是等价的？提示一下：试试用放缩法和夹逼法则来证明。

4. 闵可夫斯基距离(Minkowski Distance)

闵氏距离不是一种距离，而是一组距离的定义。

(1) 闵氏距离的定义

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：

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

其中 p 是一个变参数。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。根据变参数的不同，闵氏距离可以表示一类的距离。

(2)闵氏距离的缺点

闵氏距离，包括曼哈顿距离、欧氏距离和切比雪夫距离都存在明显的缺点。

举个例子：二维样本(身高,体重)，其中身高范围是 150~190，体重范围是 50~60，有三个样本：a(180,50)，b(190,50)，c(180,60)。那么 a 与 b 之间的闵氏距离（无论是曼哈顿距离、欧氏距离或切比雪夫距离）等于 a 与 c 之间的闵氏距离，但是身高的 10cm 真的等价于体重的 10kg 么？因此用闵氏距离来衡量这些样本间的相似度很有问题。

简单说来，闵氏距离的缺点主要有两个：

(1)将各个分量的量纲(scale)，也就是“单位”当作相同的看待了。

(2)没有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 (Standardized Euclidean distance)

标准化欧氏距离是针对简单欧氏距离的缺点而作的一种改进方案。标准欧氏距离的思路：既然数据各维分量的分布不一样，好吧！那我先将各个分量都“标准化”到均值、方差相等吧。均值和方差标准化到多少呢？这里先复习点统计学知识吧，假设样本集 X 的均值(mean)为 m，标准差(standard deviation)为 s，那么 X 的“标准化变量”表示为：

而且标准化变量的数学期望为 0，方差为 1。因此样本集的标准化过程(standardization)用公式描述就是：

$$X^* = \frac{X - m}{s}$$

标准化后的值 = (标准化前的值 － 分量的均值) /分量的标准差

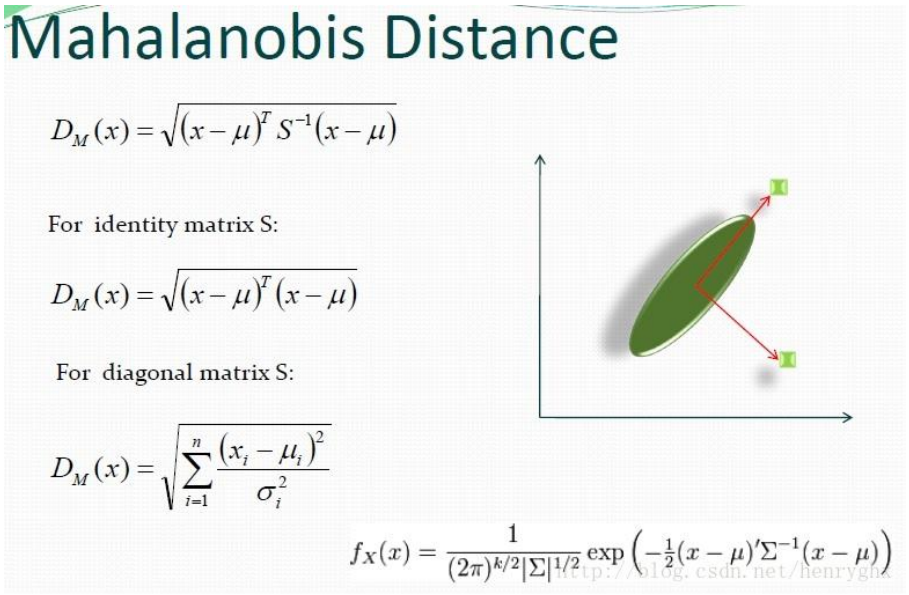
经过简单的推导就可以得到两个 n 维向量 a(x11,x12,…,x1n)与 b(x21,x22,…,x2n)间的标准化欧氏距离的公式：

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

如果将方差的倒数看成是一个权重，这个公式可以看成是一种**加权欧氏距离(Weighted Euclidean distance)**。

6. 马氏距离(Mahalanobis Distance)

马氏距离(Mahalanobis distance)是由印度统计学家 P. C. Mahalanobis(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。



（1）马氏距离定义

有 M 个样本向量 X1~Xm，协方差矩阵记为 S，均值记为向量 μ，则其中样本向量 X 到 u 的马氏距离表示为：

$$D(X) = \sqrt{(X-\mu)^T S^{-1}(X-\mu)}$$

而其中向量 Xi 与 Xj 之间的马氏距离定义为：

$$D(X_i,X_j)=\sqrt{(X_i-X_j)^T S^{-1}(X_i-X_j)}$$

若协方差矩阵是单位矩阵（各个样本向量之间独立同分布）,则公式就成了：

$$D(X_i,X_j)=\sqrt{(X_i-X_j)^T (X_i-X_j)}$$

也就是欧氏距离了。若协方差矩阵是对角矩阵，公式变成了标准化欧氏距离。

(2)马氏距离的优缺点：

优点：量纲无关，排除变量之间的相关性的干扰。

缺点：不同的特征不能差别对待，可能夸大弱特征。

(3)python 代码：

```
from numpy import *
import numpy
x = numpy.array([[3,4],[5,6],[2,2],[8,4]])
xT=x.T
D=numpy.cov(xT)
invD=numpy.linalg.inv(D)
tp=x[0]-x[1]
print numpy.sqrt(dot(dot(tp,invD), tp.T))       #1.24316312102
```

7. 夹角余弦(Cosine)

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。

(1)在二维空间中向量 A(x1,y1)与向量 B(x2,y2)的夹角余弦公式：

$$\cos\theta=\frac{x_1x_2+y_1y_2}{\sqrt{x_1^2+y_1^2}\sqrt{x_2^2+y_2^2}}$$

(2) 两个 n 维样本点 a(x11,x12,…,x1n)和 b(x21,x22,…,x2n)的夹角余弦

类似的，对于两个 n 维样本点 a(x11,x12,…,x1n)和 b(x21,x22,…,x2n)，可以使用类似于夹角余弦的概念来衡量它们间的相似程度。

$$\cos(\theta) = \frac{a \cdot b}{|a|\,|b|}$$

即：$$\cos(\theta) = \frac{\sum_{k=1}^{n} x_{1k}\, x_{2k}}{\sqrt{\sum_{k=1}^{n} x_{1k}^{\,2}}\ \sqrt{\sum_{k=1}^{n} x_{2k}^{\,2}}}$$

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

8. 汉明距离(Hamming distance)

两个等长字符串 s1 与 s2 之间的汉明距离定义为将其中一个变为另外一个所需要作的最小替换次数。例如字符串“1111”与“1001”之间的汉明距离为 2。

应用：信息编码（为了增强容错性，应使得编码间的最小汉明距离尽可能大）。

9. Jaccard 相似系数(Jaccard similarity coefficient)/ Tanimote 系数

(1) Jaccard 相似系数

两个集合 A 和 B 的交集元素在 A，B 的并集中所占的比例，称为两个集合的 Jaccard 相似系数，用符号 J(A,B)表示。

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

Jaccard 相似系数是衡量两个集合的相似度一种指标。

(2) Jaccard 距离

与 Jaccard 相似系数相反的概念是 Jaccard 距离。Jaccard 距离可用如下公式表示：

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

Jaccard 距离用两个集合中不同元素占所有元素的比例来衡量两个集合的区分度。

(3) Jaccard 相似系数与 Jaccard 距离的应用

可将 Jaccard 相似系数用在衡量样本的相似度上。

样本 A 与样本 B 是两个 n 维向量，而且所有维度的取值都是 0 或 1。例如：A(0111)和 B(1011)。我们将样本看成是一个集合，1 表示集合包含该元素，0 表示集合不包含该元素。

p ： 样本 A 与 B 都是 1 的维度的个数；q ： 样本 A 是 1，样本 B 是 0 的维度的个数；r ： 样本 A 是 0，样本 B 是 1 的维度的个数；s ： 样本 A 与 B 都是 0 的维度的个数。

那么样本 A 与 B 的 Jaccard 相似系数可以表示为：这里 p+q+r 可理解为 A 与 B 的并集的元素个数，而 p 是 A 与 B 的交集的元素个数。而样本 A 与 B 的 Jaccard 距离表示为：

$$J = \frac{p}{p+q+r}$$

(3)python 代码

```
def Jaccard(a,b):
    c=[v for v in a if v in b]
    return float(len(c))/(len(a)+len(b)-len(c))
```

10. 相关系数 (Correlation coefficient)与相关距离(Correlation distance)

(1) 相关系数的定义

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{E((X-EX)(Y-EY))}{\sqrt{D(X)}\sqrt{D(Y)}}$$

相关系数是衡量随机变量 X 与 Y 相关程度的一种方法，相关系数的取值范围是[-1,1]。相关系数的绝对值越大，则表明 X 与 Y 相关度越高。当 X 与 Y 线性相关时，相关系数取值为 1（正线性相关）或-1（负线性相关）。

(2)相关距离的定义

$$D_{xy} = 1 - \rho_{XY}$$

11. 信息熵(Information Entropy)

信息熵是衡量分布的混乱程度或分散程度的一种度量。分布越分散(或者说分布越平均)，信息熵就越大。分布越有序（或者说分布越集中），信息熵就越小。计算给定的样本集 X 的信息熵的公式：

$$\text{Entropy(X)} = \sum_{i=1}^{n} -p_i \log_2 p_i$$

参数的含义：

n：样本集 X 的分类数

pi：X 中第 i 类元素出现的概率

信息熵越大表明样本集 S 分类越分散，信息熵越小则表明样本集 X 分类越集中。当 S 中 n 个分类出现的概率一样大时（都是 1/n），信息熵取最大值 $\log_2(n)$。当 X 只有一个分类时，信息熵取最小值 0。

Mahout-DistanceMeasure （数据点间的距离计算方法）

在分类聚类算法,推荐系统中,常要用到两个输入变量(通常是特征向量的形式)距离的计算,即相似性度量.不同相似性度量对于算法的结果,有些时候,差异很大.因此,有必要根据输入数据的特征,选择一种合适的相似性度量方法.

令 $\boldsymbol{X}=(x_1,x_2,..,x_n)^T$,$\boldsymbol{Y}=(y_1,y_2,...y_n)^T$ 为两个输入向量.

1.欧几里得距离(Euclidean distance)-**EuclideanDistanceMeasure**.

$$dist(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

相当于高维空间内向量说表示的点到点之间的距离。由于特征向量的各分量的量纲不一致，通常需要先对各分量进行标准化，使其与单位无关，比如对身高（cm）和体重（kg）两个单位不同的指标使用欧式距离可能使结果失效。

优点：简单,应用广泛(如果也算一个优点的话)

缺点：没有考虑分量之间的相关性,体现单一特征的多个分量会干扰结果。

2.马氏距离(Mahalanobis distance)-**MahalanobisDistanceMeasure**

$$D(X_i,X_j)=\sqrt{(X_i-X_j)^T S^{-1}(X_i-X_j)}$$

S=E[(Xi-Xi 平均)(Xj-Xj 平均)]为该类输入向量 X 的协方差矩阵.(T 为转置符号，E 取平均时是样本因此为 n-1)

适用场合：

1)度量两个服从同一分布并且其协方差矩阵为 C 的随机变量 Xi 与 Xj 的差异程度

2)度量 Xi 与某一类的均值向量的差异程度,判别样本归属。此时，Xj 为类均值向量.

优点：1）独立于分量量纲；

2）排除了样本之间的相关性影响。

缺点：不同的特征不能差别对待，可能夸大弱特征。

3.闵可夫斯基距离(Minkowsk distance)-**MinkowskiDistanceMeasure**（默认 p=3）

$$dist(X,Y)=(\sum_{i=1}^{n}|x_i-y_i|^p)^{1/p}$$

可看成是欧氏距离的指数推广，还没有见到过很好的应用实例，但通常，推广都是一种进步，特别的，

当 p=1,也成做**曼哈顿距离**，也称绝对距离，曼哈顿距离来源于城市区块距离，是将多个维度上的距离进行求和后的结果。**ManhattanDistanceMeasure**.

$$dist(X,Y) = \sum_{i=1}^{n} |x_i - y_i|$$

当 q=∞时，称为**切比雪夫距离**，**ChebyshevDistanceMeasure**

切比雪夫距离起源于国际象棋中国王的走法，我们知道国际象棋国王每次只能往周围的 8 格中走一步，那么如果要从棋盘中 A 格（x1， y1）走到 B 格（x2， y2）最少需要走几步？扩展到多维空间，其实切比雪夫距离就是当 p 趋向于无穷大时的明氏距离：

$$dist(X,Y) = \lim_{p\to\infty}(\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} = max\,|x_i - y_i|$$

4.汉明距离(Hamming distance)-**Mahout 无**

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。

5.Tanimoto 系数(又称广义 Jaccard 系数)-**TanimotoDistanceMeasure**.

$$d = 1 - \frac{(a_1b_1 + a_2b_2 + \ldots + a_nb_n)}{\sqrt{(a_1^2 + a_2^2 + \ldots + a_n^2)} + \sqrt{(b_1^2 + b_2^2 + \ldots + b_n^2)} - (a_1b_1 + a_2b_2 + \ldots + a_nb_n)}$$

通常应用于 X 为布尔向量，即各分量只取 0 或 1 的时候。此时，表示的是 X,Y 的公共特征的占 X，Y 所占有的特征的比例。

5.Jaccard 系数

Jaccard 系数主要用于计算符号度量或布尔值度量的个体间的相似度，因为个体的特征属性都是由符号度量或者布尔值标识，因此无法衡量差异具体值的大小，只能获得“是否相同”这个结果，所以 Jaccard 系数只关心个体间共同具有的特征是否一致这个问题。如果比较 X 与 Y 的 Jaccard 相似系数，只比较 xn 和 yn 中相同的个数，公式如下：

$$Jaccard(X,Y) = \frac{X \cap Y}{X \cup Y}$$

7.皮尔逊相关系数(Pearson correlation coefficient)-PearsonCorrelationSimilarity

即相关分析中的相关系数 r，分别对 X 和 Y 基于自身总体标准化后计算空间向量的余弦夹角。公式如下：

$$r(X,Y) = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \cdot \sqrt{n\sum y^2 - (\sum y)^2}}$$

8.余弦相似度(cosine similarity)-CosineDistanceMeasure

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

就是两个向量之间的夹角的余弦值。余弦相似度用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小。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。

优点：不受坐标轴旋转，放大缩小的影响。

9.调整余弦相似度-Adjusted Cosine Similarity

虽然余弦相似度对个体间存在的偏见可以进行一定的修正，但是因为只能分辨个体在维之间的差异，没法衡量每个维数值的差异，会导致这样一个情况：比如用户对内容评分，5 分制，X 和 Y 两个用户对两个内容的评分分别为（1，2）和（4，5），使用余弦相似度得出的结果是 0.98，两者极为相似，但从评分上看 X 似乎不喜欢这 2 个内容，而 Y 比较喜欢，余弦相似度对数值的不敏感导致了结果的误差，需要修正这种不合理性，就出现了调整余弦相似度，即所有维度上的数值都减去一个均值，比如 X 和 Y 的评分均值都是 3，那么调整后为（-2，-1）和（1，2），再用余弦相似度计算，得到-0.8，相似度为负值并且差异不小，但显然更加符合现实。

调整余弦相似度和余弦相似度，皮尔逊相关系数在推荐系统中应用较多。在基于项目的推荐中，GroupLens 有篇论文结果表明调整余弦相似度性能要优于后两者。

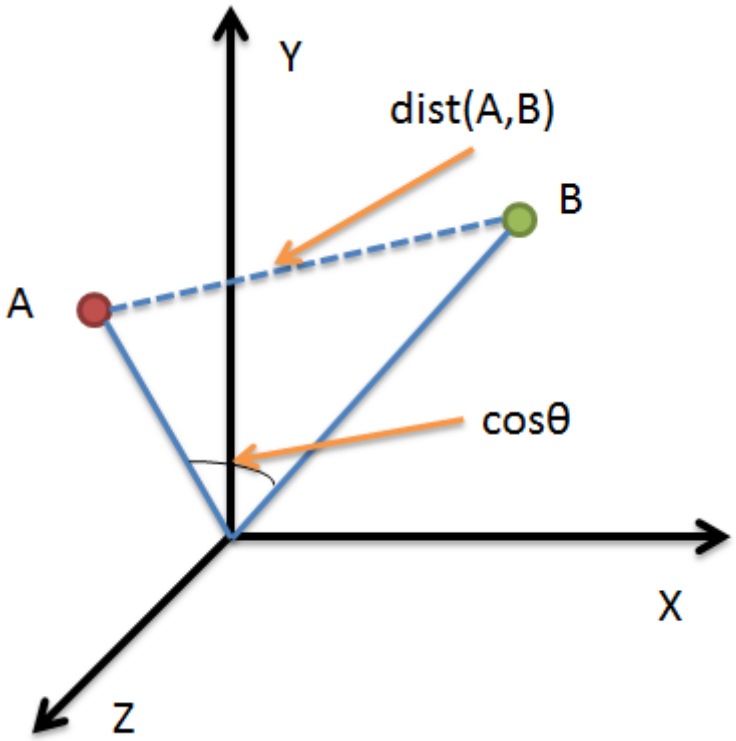
10.基于权重的距离计算方法：

WeightedDistanceMeasure、

WeightedEuclideanDistanceMeasure 、 WeightedManhattanDistanceMeasure

欧氏距离与余弦相似度

借助三维坐标系来看下欧氏距离和余弦相似度的区别：



根据欧氏距离和余弦相似度各自的计算方式和衡量特征，分别适用于不同的数据分析模型：欧氏距离能够体现个体数值特征的绝对差异，所以更多的用于需要从维度的数值大小中体现差异的分析，如使用用户行为指标分析用户价值的相似度或差异；而余弦相似度更多的是从方向上区分差异，而对绝对的数值不敏感，更多的用于使用用户对内容评分来区分用户兴趣的相似度和差异，同时修正了用户间可能存在的度量标准不统一的问题（因为余弦相似度对绝对数值不敏感）。