在数据分析和数据挖掘的过程中，我们经常需要知道个体间差异的大小，进而评价个体的相似性和类别。最常见的是数据分析中的相关分析，数据挖掘中的分类和聚类算法，如K最近邻（KNN）和K均值（K-Means）。当然衡量个体差异的方法有很多，最近查阅了相关的资料，这里整理罗列下。

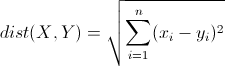
　　为了方便下面的解释和举例，先设定我们要比较X个体和Y个体间的差异，它们都包含了N个维的特征，即X=（x1, x2, x3, … xn），Y=（y1, y2, y3, … yn）。下面来看看主要可以用哪些方法来衡量两者的差异，主要分为距离度量和相似度度量。

**距离度量**

　　距离度量（Distance）用于衡量个体在空间上存在的距离，距离越远说明个体间的差异越大。

**欧几里得距离(Euclidean Distance)**

　　欧氏距离是最常见的距离度量，衡量的是多维空间中各个点之间的绝对距离。公式如下：



　　因为计算是基于各维度特征的绝对数值，所以欧氏度量需要保证各维度指标在相同的刻度级别，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。

**明可夫斯基距离(Minkowski Distance)**

　　明氏距离是欧氏距离的推广，是对多个距离度量公式的概括性的表述。公式如下：

Minkowski Distance

　　这里的p值是一个变量，当p=2的时候就得到了上面的欧氏距离。

**曼哈顿距离(Manhattan Distance)**

　　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，即当上面的明氏距离中p=1时得到的距离度量公式，如下：

Manhattan Distance

**切比雪夫距离(Chebyshev Distance)**

　　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：

Chebyshev Distance

　　其实上面的曼哈顿距离、欧氏距离和切比雪夫距离都是明可夫斯基距离在特殊条件下的应用。

**马哈拉诺比斯距离(Mahalanobis Distance)**

　　既然欧几里得距离无法忽略指标度量的差异，所以在使用欧氏距离之前需要对底层指标进行[数据的标准化](http://webdataanalysis.net/data-analysis-method/data-normalization/" \t "_blank)，而基于各指标维度进行标准化后再使用欧氏距离就衍生出来另外一个距离度量——马哈拉诺比斯距离（Mahalanobis Distance），简称马氏距离。

**相似度度量**

　　相似度度量（Similarity），即计算个体间的相似程度，与距离度量相反，相似度度量的值越小，说明个体间相似度越小，差异越大。

**向量空间余弦相似度(Cosine Similarity)**

　　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。公式如下：

Cosine Similarity

**皮尔森相关系数(Pearson Correlation Coefficient)**

　　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：

Pearson Correlation Coefficient

**Jaccard相似系数(Jaccard Coefficient)**

　　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 Coefficient

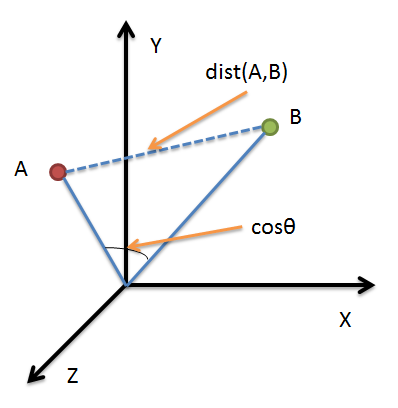
**调整余弦相似度(Adjusted Cosine Similarity)**

　　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。

**欧氏距离与余弦相似度**

　　欧氏距离是最常见的距离度量，而余弦相似度则是最常见的相似度度量，很多的距离度量和相似度度量都是基于这两者的变形和衍生，所以下面重点比较下两者在衡量个体差异时实现方式和应用环境上的区别。

　　借助三维坐标系来看下欧氏距离和余弦相似度的区别：

[](http://webdataanalysis.net/wp-content/uploads/2011/10/distance-and-similarity.png)

　　从图上可以看出距离度量衡量的是空间各点间的绝对距离，跟各个点所在的位置坐标（即个体特征维度的数值）直接相关；而余弦相似度衡量的是空间向量的夹角，更加的是体现在方向上的差异，而不是位置。如果保持A点的位置不变，B点朝原方向远离坐标轴原点，那么这个时候余弦相似度cosθ是保持不变的，因为夹角不变，而A、B两点的距离显然在发生改变，这就是欧氏距离和余弦相似度的不同之处。

　　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）。

　　上面都是对距离度量和相似度度量的一些整理和汇总，在现实的使用中选择合适的距离度量或相似度度量可以完成很多的数据分析和数据挖掘的建模，后续会有相关的介绍。

*来源：*[*网站数据分析*](http://webdataanalysis.net/) *»* [*《距离和相似度度量》*](http://webdataanalysis.net/reference-and-source/distance-and-similarity/)