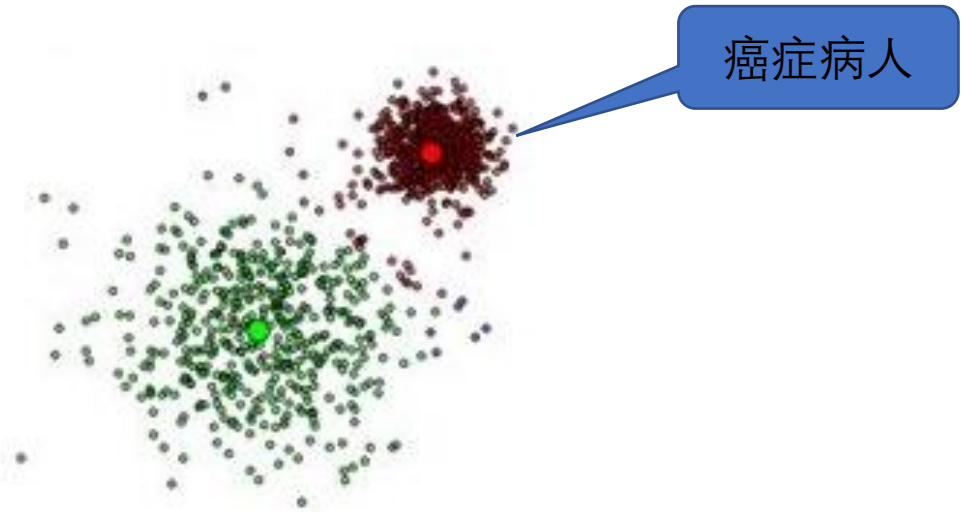
第5章 聚类

本章目标

- 基本概念;
- 了解K-means算法的基本原理、关键技术和应用。

引例

- 基因检测可用于发现癌症。
- 每个样本中测量了114个基因特征，无需任何医学知识就可以使用聚类方法找到患有癌症的病人。



5.1 聚类分析技术

- 将多个无明显分类特征的对象，按照某种相似性分成多个簇（Cluster）的分析过程。
- 与分类不同，聚类分析是在没有给定划分类别的情况下，根据数据相似度进行样本分组的一种方法。
- 与分类模型需要使用有类标记样本构成的训练数据不同，聚类模型可以建立在无类标记的数据上，是一种非监督的学习算法。

5.1 聚类分析技术

- 在正式讨论聚类前，我们要先弄清楚一个问题：如何定量计算两个可比较元素间的相异度。
- 用通俗的话说，相异度就是两个东西差别有多大，例如人类与章鱼的相异度明显大于人类与黑猩猩的相异度，这是能我们直观感受到的。
- 但是，计算机没有这种直观感受能力，我们必须对相异度在数学上进行定量定义。

5.1 聚类分析技术

- 设

$$X = \{x_1, x_2, \ldots, x_n\}, Y = \{y_1, y_2, \ldots, y_n\}$$

  - 其中X，Y是两个元素项，各自具有n个可度量特征属性，那么X和Y的相异度定义为：

  - 其中R为实数域。也就是说相异度是两个元素对实数域的一个映射，所映射的实数定量表示两个元素的相异 $d(X, Y) = f(X, Y) \rightarrow R$

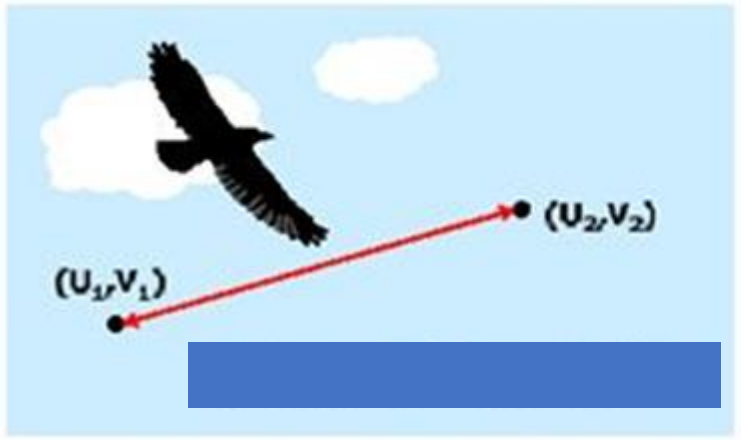
- 下面介绍不同类型变量相异度计算方法。

5.1 聚类分析技术

- 1、标量
- 标量也就是无方向意义的数字，也叫标度变量。现在先考虑元素的所有特征属性都是标量的情况。
- 例如，计算 X={2,1,102} 和 Y={1,3,2} 的相异度。
- 一种很自然的想法是用两者的距离来作为相异度。

5.1 聚类分析技术

- 欧式距离
  - 最易于理解的距离计算方法，源自欧氏空间中两点间距离公式。



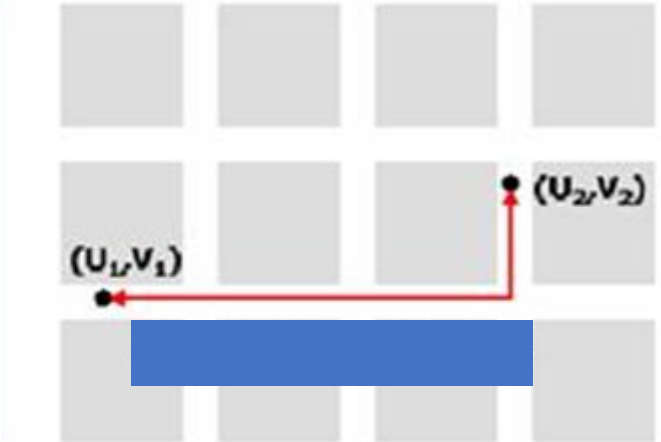
欧氏距离

计算 X={2,1,102} 和 Y={1,3,2} 的相异度

$$d(X,Y)=\sqrt{(2-1)^2+(1-3)^2+(102-2)^2}=100.025$$

5.1 聚类分析技术

- 曼哈顿距离
  - 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)。



曼哈顿距离

5.1 聚类分析技术

- 切比雪夫距离
  - 国际象棋的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?
  - 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

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

5.1 聚类分析技术

- 闵可夫斯基距离

$$d(X,Y)=\sqrt[p]{|x_1-y_1|^p+|x_2-y_2|^p+\cdots+|x_n-y_n|^p}$$

- 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

5.1 聚类分析技术

- 用p个属性来表示n个样本的数据矩阵如下，现在计算第i个样本和第j个样本之间的距离

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

曼哈顿距离：

欧几里得距离：

5.1 聚类分析技术

- 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。
- 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。

5.1 聚类分析技术

- 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，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。
- 导致的问题：取值范围大的属性对距离的影响高于取值范围小的属性 。例如计算 X={2,1,102} 和 Y={1,3,2} 的相异度，第三个属性的取值跨度远大于前两个，这样不利于真实反映真实的相异度，为了解决这个问题，一般要对属性值进行规格化。

5.1 聚类分析技术

- 标量的规格化问题
- 所谓规格化就是将各个属性值按比例映射到相同的取值区间，这样是为了平衡各个属性对距离的影响。通常将各个属性均映射到 [0,1] 区间，映射公式为：

$$a_i' = \frac{a_i - min(a_i)}{max(a_i) - min(a_i)}$$

- 其中 max(ai) 和 min(a 为最大值和最小值。
- 将上例中的元素规格化到 [0,1] 区间后，变成了 X'={1,0,1}，Y'={0,1,0}，重新计算欧氏距离约为 1.732 。

5.1 聚类分析技术

- 标准化欧氏距离
- 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，那先将各个分量都“标准化”到均值、方差相等吧。
- 假设样本集X的均值为m，标准差为s，那么X的“标准化变量”表示为：

- 标准化后的值 =（标准化前的值 － 分量的均值）/分量的标准差 。

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

5.1 聚类分析技术

- 经过推导可以得到两个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。

5.1 聚类分析技术

- 马氏距离(Mahalanobis Distance)
- 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：

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

5.1 聚类分析技术

- 马氏距离(Mahalanobis Distance)
- 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:

- 也就是欧氏距离了。 $D(X_i,X_j)=\sqrt{(X_i-X_j)^T(X_i-X_j)}$
- 若协方差矩阵是对角矩阵，公式变成标准化欧氏距离。
- 马氏距离的优缺点：量纲无关，排除变量之间的相关性的干扰。

5.1 聚类分析技术

- 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）。

5.1 聚类分析技术

- 杰卡德相似系数(Jaccard similarity coefficient)

- 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。

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

- 杰卡德相似系数是衡量两个集合的相似度一种指标。

5.1 聚类分析技术

- 杰卡德距离
- 　　与杰卡德相似系数相反的概念是杰卡德距离(Jaccard distance)。杰卡德距离可用如下公式表示:

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

- 　　杰卡德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。

5.1 聚类分析技术

- 杰卡德相似系数衡量样本的相似度
  - 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的杰卡德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。

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

5.1 聚类分析技术

- 相关系数与相关距离
  - 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）。

- 相关距离的定义

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

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

5.1 聚类分析技术

- 2、二元变量
- 所谓二元变量是只能取 0 和 1 两种值变量，有点类似布尔值，通常用来标识是或不是这种二值属性。
- 对于二元变量，上一节提到的距离不能很好标识其相异度，我们需要一种更适合的标识。
- 一种常用的方法是用元素相同序位同值属性的比例来标识其相异度。

5.1 聚类分析技术

- 2、二元变量
- 设有 X={1,0,0,0,1,0,1,1} ， Y={0,0,0,1,1,1,1,1}，可以看到，两个元素第 2 、 3 、 5 、 7 和 8 个属性取值相同，而第 1 、 4 和 6 个取值不同，那么相异度可以标识为
- 3/8=0.375 。
- 一般的，对于二元变量，相异度可用“ 取值不同的同位属性数/单个元素的属性位数 ”标识。

5.1 聚类分析技术

- 2、二元变量
- 上面所说的相异度应该叫做对称二元相异度。
- 现实中还有一种情况，就是我们只关心两者都取 1 的情况，而认为两者都取 0 的属性并不意味着两者更相似。
- 例如在根据病情对病人聚类时，如果两个人都患有肺癌，我们认为两个人增强了相似度，但如果两个人都没患肺癌，并不觉得这加强了两人的相似性。

5.1 聚类分析技术

- 2、二元变量
- 在这种情况下，改用“ 取值不同的同位属性数/(单个元素的属性位数-同取0的位数) ” 来标识相异度，这叫做非对称二元相异度。
- 设有 X={1,0,0,0,1,0,1,1} ， Y={0,0,0,1,1,1,1,1} ， 非对称二元相异度可以标识为
- 3/6=0.5 。
- 如果用 1 减去非对称二元相异度，则得到非对称二元相似度，也叫 Jaccard 系数。

5.1 聚类分析技术

- 3、分类变量
- 分类变量是二元变量的推广，类似于程序中的枚举变量，但各个值没有数字或序数意义。
- 如颜色、民族等等，对于分类变量，用“ 取值不同的同位属性数/单个元素的全部属性数 ”来标识其相异度。

5.1 聚类分析技术

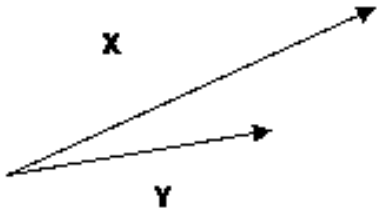
- 4、序数变量
- 序数变量是具有序数意义的分类变量，通常可以按照一定顺序意义排列，如冠军、亚军和季军。
- 对于序数变量，一般为每个值分配一个数，叫做这个值的秩，然后以秩代替原值当做标量属性计算相异度。

5.1 聚类分析技术

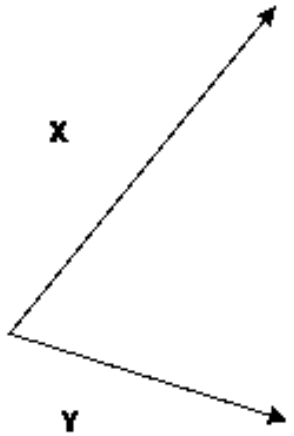
- 5、向量
- 对于向量，由于它不仅有大小而且有方向，所以闵可夫斯基距离不是度量其相异度的好办法，一种流行的做法是用两个向量的余弦度量。
- 余弦距离，也称为余弦相似度，是用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小的度量。

5.1 聚类分析技术

- 5、向量
- 向量，是多维空间中有方向的线段，如果两个向量的方向一致，即夹角接近零，那么这两个向量就相近。而要确定两个向量方向是否一致，这就要用到余弦定理计算向量的夹角。



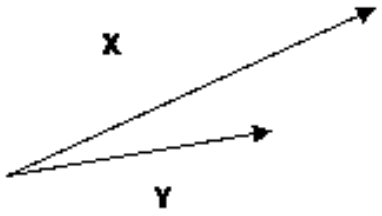
两条新闻相似



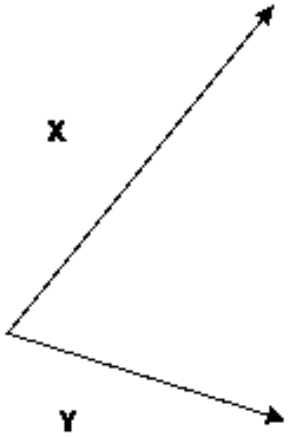
两条新闻无关

5.1 聚类分析技术

- 5、向量
- 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。



两条新闻相似



两条新闻无关

5.1 聚类分析技术

- 余弦定理描述了三角形中任何一个夹角和三个边的关系。给定三角形的三条边，可以使用余弦定理求出三角形各个角的角度。
- 假定三角形的三条边为a，b和c，对应的三个角为A，B和C，那么角A的余弦为：

$$\cos A = \frac{b^2 + c^2 - a^2}{2bc}$$

5.1 聚类分析技术

- 如果将三角形的两边b和c看成是两个向量，则上述公式等价于：

$$\cos A = \frac{< b, c >}{|b \| c|}, \quad sim(X,Y) = cos\theta = \frac{\vec{x} \cdot \vec{y}}{\|x\| \cdot \|y\|}$$

- 其中分母表示两个向量b和c的长度，分子表示两个向量的点积（或内积）。

5.1 聚类分析技术

- 举例，文本相似度算法
- 在目前这个信息过载的星球上，文本相似度计算的应用前景是非常广泛的。
- 文本相似度计算的应用场景
  - 搜索引擎上过滤相似度很高的新闻，或者网页去重
  - 考试防作弊系统
  - 论文查重检查

5.1 聚类分析技术

- 举例，文本相似度算法
- 算法步骤：
  - 预处理
  - 文本特征项选择
  - 加权
  - 生成向量空间模型后计算余弦。

5.1 聚类分析技术

- 预处理主要是进行中文分词和去停用词。
- 文本特征项选择
  - 过滤掉常用副词、助词等频度高的词之后，根据剩下词的频度确定若干关键词。频度计算参照TF公式。
  - Term Frequency即关键词词频，是指一篇文章中关键词出现的频率，比如在一篇M个词的文章中有N个该关键词，则
  - TF=N/M
  - 为该关键词在这篇文章中的词频。

5.1 聚类分析技术

- 加权
  - 加权是针对每个关键词对文本特征的体现效果大小不同而设置的机制，权值计算参照IDF公式。
  - Inverse Document Frequency指逆向文本频率，是用于衡量关键词权重的指数，由公式
  - IDF=log(D/Dw)

  - 计算而得，其中D为文章总数，Dw为关键词出现过的文章数。

5.1 聚类分析技术

- 向量空间模型VSM及余弦计算
  - 向量空间模型的基本思想是把文档简化为以特征项（关键词）的权重为分量的N维向量表示。
  - 这个模型假设词与词间不相关（这个前提造成这个模型无法进行语义相关的判断，向量空间模型的缺点在于关键词之间的线性无关的假说前提），用向量来表示文本，从而简化了文本中的关键词之间的复杂关系，文档用十分简单的向量表示，使得模型具备了可计算性。

5.1 聚类分析技术

- 向量空间模型VSM及余弦计算
  - 在向量空间模型中，文本泛指各种机器可读的记录。
  - 用D（Document）表示文本，特征项（Term，用t表示）指出现在文档D中且能够代表该文档内容的基本语言单位，主要是由词或者短语构成，文本可以用特征项集表示为
  - D（T1，T2，…，Tn），
  - 其中Tk是特征项，要求满足1<=k<=N。

5.1 聚类分析技术

- 向量空间模型VSM及余弦计算
  - 下面是向量空间模型（特指权值向量空间）的解释。
  - 假设一篇文档中有a、b、c、d四个特征项，那么这篇文档就可以表示为
  - D（a，b，c，d）
  - 对于其它要与之比较的文本，也将遵从这个特征项顺序。对含有n个特征项的文本而言，通常会给每个特征项赋予一定的权重表示其重要程度，即
  - D＝D（T1，W1；T2，W2；…，Tn，Wn）
  - 简记为D＝D（W1，W2，…，Wn）
  - 我们把它叫做文本D的权值向量表示，其中Wk是Tk的权重，1<=k<=N。

5.1 聚类分析技术

- 向量空间模型VSM及余弦计算
  - 在上例中，假设a、b、c、d的权重分别为30，20，20，10，那么该文本的向量表示为
  - D（30，20，20，10）
  - 在向量空间模型中，两个文本D1和D2之间的内容相关度Sim（D1，D2）常用向量之间夹角的余弦值表示，公式为：

- 其中，W1k、W2k分 $Sim(D_1, D_2) = \cos\theta = \frac{\sum_{k=1}^{n} W_{1k} \times W_{2k}}{\sqrt{\left(\sum_{k=1}^{n} W_{1k}^2\right)\left(\sum_{k=1}^{n} W_{2k}^2\right)}}$ 1<=k<=N。

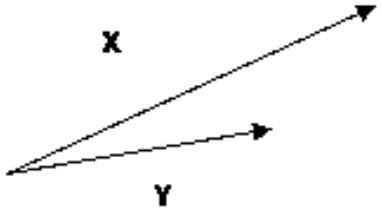
5.1 聚类分析技术

- 举例，计算新闻X和新闻Y的相似度：
  - 新闻X中出现的字（或词）为：Z1c1,Z1c2,Z1c3……Z1cn；它们的个数为：Z1n1,Z1n2,Z1n3……Z1nm；
  - 新闻Y中出现的字（或词）为：Z2c1,Z2c2,Z2c3……Z2cn；它们的个数为：Z2n1,Z2n2,Z2n3……Z2nm；
  - 其中，Z1c1和Z2c1表示两个文本中同一个字（或词），Z1n1和Z2n1是它们分别对应的个数。
- 两者的相似度可以用它们之间夹角的余弦值来表示：

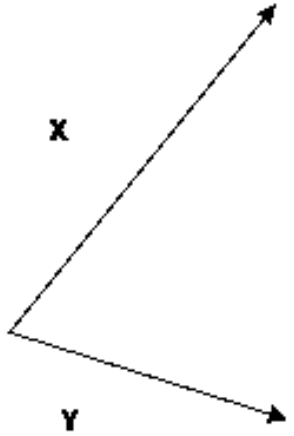
$$\text{SimilaryValue} = \frac{(Z1n1 \times Z2n1) + (Z1n2 \times Z2n2) + (Z1n3 \times Z2n3) \ldots\ldots \times (Z1nn \times Z2nn)}{\sqrt{Z1n1^2 + Z1n2^2 + Z1n3^2 \ldots + Z1nn^2} \times \sqrt{Z2n1^2 + Z2n2^2 + Z2n3^2 \ldots\ldots + Z2nn^2}}$$

5.1 聚类分析技术

- 当两条新闻向量夹角余弦等于1时，这两条新闻完全重复（用这个办法可以删除爬虫所收集网页中的重复网页）；当夹角的余弦值接近于1时，两条新闻相似（可以用作文本分类）；夹角的余弦越小，两条新闻越不相关。



两条新闻相似



两条新闻无关

5.1 聚类分析技术

- 举例，文档自动归类：
  - 一篇文档中有a、b、c、d四个特征项，那么这篇文档就可以表示为D(a，b，c，d)。
  - 对含有n个特征项的文本而言，通常会给每个特征项赋予一定的权重表示其重要程度。即D = D(T1，W1；T2，W2；…，Tn，Wn)，简记为D = D(W1，W2，…，Wn)，我们把它叫做文本D的向量表示。
  - 其中Wk是Tk的权重，1<=k<=N。
  - 在上例子中，假设a、b、c、d的权重分别为30，20，20，10，那么该文本的向量表示为
  - D(30，20，20，10)。

5.1 聚类分析技术

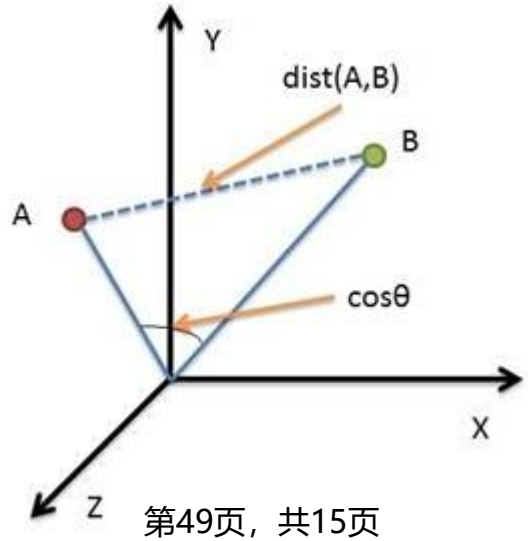
- 举例，文档自动归类：
  - 在向量空间模型中，两个文本D1和D2之间的内容相关度Sim(D1，D2)常用向量之间夹角的余弦值表示。
  - 在自动归类中，可以用类似的方法来计算待归类文档和某类目的相关度。
  - 例如文本D1的特征项为a，b，c，d，权值分别为30，20，20，10，类目C1的特征项为a，c，d，e，权值分别为40，30，20，10，则D1的向量表示为
  - D1（30，20，20，10，0)，C1的向量表示为
  - C1（40，0，30，20，10）。

5.1 聚类分析技术

- 举例，文档自动归类：
  - D1（30，20，20，10，0），C1（40，0，30，20，10）
  - 相似度 = d1*c1/(|d1|*|c1|)
  - 两个向量的点积d1*c1为
  - d1*c1 = 30*40 + 20*0 + 20*30 + 10*20 + 0*10 = 2000
  - 向量d1的模|d1|为
  - |d1| = sqrt(30*30 +20*20 + 20*20 + 10*10 + 0*0) = sqrt(1800)
  - 向量c1的模|c1|为
  - |c1| = sqrt(40*40 + 0*0 + 30*30 + 20*20 + 10*10) = sqrt(3000)
  - 相似度 = d1*c1/(|d1|*|c1|)= 2000/sqrt(1800*3000)= 0.86066

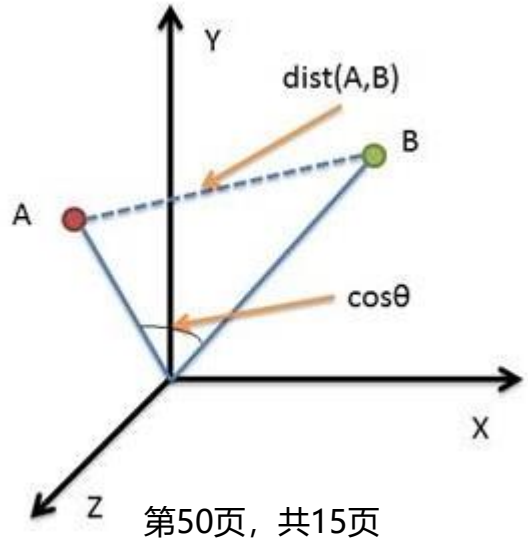
5.1 聚类分析技术

- 余弦距离和欧氏距离的对比
- 余弦距离使用两个向量夹角的余弦值作为衡量两个个体间差异的大小。相比欧氏距离，余弦距离更加注重两个向量在方向上的差异。
- 借助三维坐标系来看下欧氏距离和余弦距离的区别:



5.1 聚类分析技术

- 欧氏距离衡量的是空间各点的绝对距离，跟各个点所在的位置坐标直接相关；而余弦距离衡量的是空间向量的夹角，更加体现在方向上的差异，而不是位置。
- 如果保持A点位置不变，B点朝原方向远离坐标轴原点，那么这个时候余弦距离 是保持不变的（因为夹角没变），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距离之间的不同之处。

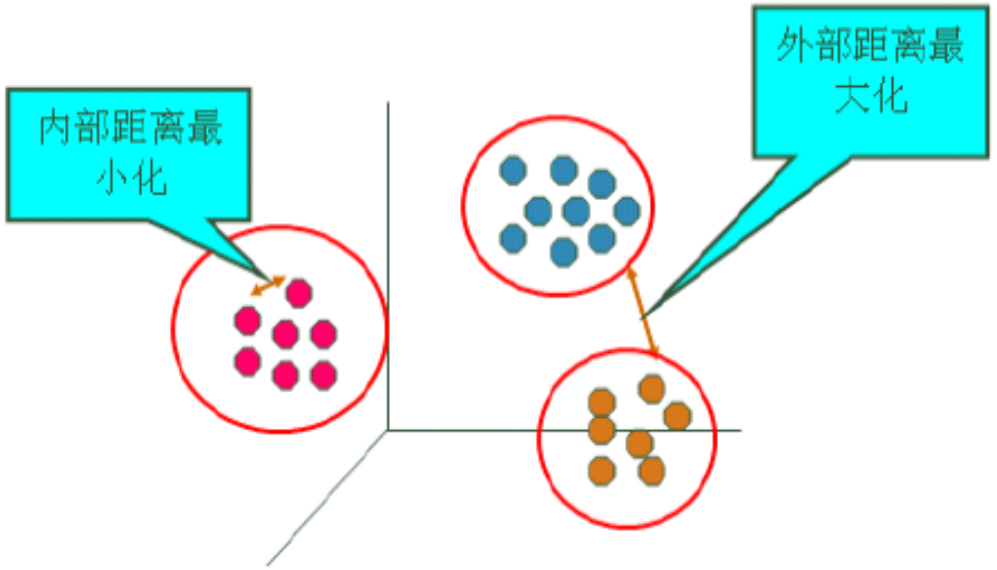


5.1 聚类分析技术

- 欧氏距离和余弦距离各自有不同的计算方式和衡量特征，因此它们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。
- 余弦距离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（因为余弦距离对绝对数值不敏感）。

5.1 聚类分析技术

- 聚类的输入是一组未被标记的样本，聚类根据数据自身的距离或相似度将其划分为若干组。
- 划分的原则是组内样本最小化而组间（外部）距离最大化。



5.1 聚类分析技术

- K-means算法（K-均值算法）
  - 1982年斯图尔特·劳埃德（Stuart Lloyd）提出的聚类技术。
  - 最著名、由于简洁和效率使其成为所有聚类算法中最广泛使用的、聚类效果也很好。

- K-means聚类算法就是基于距离的聚类算法
  - 即采用距离作为相似性度量的评价指标。