常用距离计算汇总

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，甚至关系到分类的正确与否。

本文的目的就是对常用的相似性度量作一个总结。

本文目录：

1. 欧氏距离(Euclidean Distance)

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。

(1)二维平面上两点 a(x1,y1)与 b(x2,y2)间的欧氏距离：

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

(2)三维空间两点 a(x1,y1,z1)与 b(x2,y2,z2)间的欧氏距离：

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

(3)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：

$$d_{12} = \sqrt{(a-b)(a-b)^T}$$

(4)Matlab 计算欧氏距离

Matlab 计算距离主要使用 pdist 函数。若 X 是一个 M×N 的矩阵，则 pdist(X)将 X 矩阵 M 行的每一行作为一个 N 维向量，然后计算这 M 个向量两两间的距离。

例子：计算向量(0,0)、(1,0)、(0,2)两两间的欧式距离

```
X = [0 0 ; 1 0 ; 0 2]
D = pdist(X,'euclidean')
```

结果：

```
D =
    1.0000    2.0000    2.2361
```

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

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

(3) Matlab 计算曼哈顿距离

例子：计算向量(0,0)、(1,0)、(0,2)两两间的曼哈顿距离

```
X = [0 0 ; 1 0 ; 0 2]
D = pdist(X, 'cityblock')
```

结果：

D =

1 2 3

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

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

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。

(3)Matlab 计算切比雪夫距离

例子：计算向量(0,0)、(1,0)、(0,2)两两间的切比雪夫距离

X = [0 0 ; 1 0 ; 0 2]

D = pdist(X, 'chebychev')

结果：

D =

1 2 2

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。

(3)Matlab 计算闵氏距离

例子：计算向量(0,0)、(1,0)、(0,2)两两间的闵氏距离（以变参数为 2 的欧氏距离为例）

```
X = [0 0 ; 1 0 ; 0 2]
D = pdist(X,'minkowski',2)
```

结果：

```
D =
    1.0000    2.0000    2.2361
```

5. 标准化欧氏距离 (Standardized Euclidean distance)

(1)标准欧氏距离的定义

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)**。

(2)Matlab 计算标准化欧氏距离

例子：计算向量(0,0)、(1,0)、(0,2)两两间的标准化欧氏距离 (假设两个分量的标准差分别为 0.5 和 1)

```
X = [0 0 ; 1 0 ; 0 2]

D = pdist(X, 'seuclidean',[0.5,1])
```

结果：

```
D =

    2.0000    2.0000    2.8284
```

6. 马氏距离(Mahalanobis Distance)

（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：量纲无关，排除变量之间的相关性的干扰。

(3) Matlab 计算(1 2)，(1 3)，(2 2)，(3 1)两两之间的马氏距离

```
X = [1 2; 1 3; 2 2; 3 1]
Y = pdist(X,'mahalanobis')
```

结果：

```
Y =
    2.3452    2.0000    2.3452    1.2247    2.4495    1.2247
```

7. 夹角余弦(Cosine)

有没有搞错，又不是学几何，怎么扯到夹角余弦了？各位看官稍安勿躁。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。

夹角余弦的具体应用可以参阅参考文献[1]。

(3)Matlab 计算夹角余弦

例子：计算(1,0)、(1,1.732)、(-1,0)两两间的夹角余弦

```
X = [1 0 ; 1 1.732 ; -1 0]
D = 1- pdist(X, 'cosine')  % Matlab 中的 pdist(X, 'cosine')得到的是 1 减夹角余弦的值
```

结果：

```
D =
    0.5000   -1.0000   -0.5000
```

8. 汉明距离(Hamming distance)

(1)汉明距离的定义

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）。

(2)Matlab 计算汉明距离

Matlab 中 2 个向量之间的汉明距离的定义为 2 个向量不同的分量所占的百分比。

例子：计算向量(0,0)、(1,0)、(0,2)两两间的汉明距离

```
X = [0 0 ; 1 0 ; 0 2];
D = PDIST(X, 'hamming')
```

结果：

```
D =
    0.5000    0.5000    1.0000
```

9. 杰卡德相似系数(Jaccard similarity coefficient)

(1) 杰卡德相似系数

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。

(2) 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。

(3) 杰卡德相似系数与杰卡德距离的应用

可将杰卡德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。

而样本 A 与 B 的杰卡德距离表示为：

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

(4)Matlab 计算杰卡德距离

Matlab 的 pdist 函数定义的杰卡德距离跟我这里的定义有一些差别，Matlab 中将其定义为不同的维度的个数占"非全零维度"的比例。

例子：计算(1,1,0)、(1,-1,0)、(-1,1,0)两两之间的杰卡德距离

X = [1 1 0; 1 -1 0; -1 1 0]

D = pdist(X , 'jaccard')

结果

D =

0.5000 0.5000 1.0000

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

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

(3)Matlab 计算(1, 2 ,3 ,4)与(3 ,8 ,7 ,6)之间的相关系数与相关距离

X = [1 2 3 4 ; 3 8 7 6]

C = corrcoef(X') %将返回相关系数矩阵

D = pdist(X , 'correlation')

结果：

C =

  1.0000 0.4781

  0.4781 1.0000

D =

0.5219

其中 0.4781 就是相关系数，0.5219 是相关距离。

11. 信息熵(Information Entropy)

信息熵并不属于一种相似性度量。那为什么放在这篇文章中啊？这个。。。我也不知道。 (ノ▽ヽ)

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

参考资料：

[1]吴军. 数学之美 系列 12 - 余弦定理和新闻的分类.

http://www.google.com.hk/ggblog/googlechinablog/2006/07/12_4010.html

[2] Wikipedia. Jaccard index.

http://en.wikipedia.org/wiki/Jaccard_index

[3] Wikipedia. Hamming distance

http://en.wikipedia.org/wiki/Hamming_distance

[4] 求马氏距离（Mahalanobis distance ）matlab 版

http://junjun0595.blog.163.com/blog/static/969561420100633351210/

[5] Pearson product-moment correlation coefficient

http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient