MATLAB求马氏距离(Mahalanobis distance)

作者：凯鲁嘎吉 - 博客园 http://www.cnblogs.com/kailugaji/

1.马氏距离计算公式

$$d^2(x_i, x_j)=(x_i-x_j)^T S^{-1}(x_i-x_j)$$

其中，S是总体的协方差矩阵，而不是样本的协方差矩阵。

2.matlab中现有的函数

```
>> x=[155 66;180 71;190 73;160 60;190 68;150 58;170 75]

x =

   155    66
   180    71
   190    73
   160    60
   190    68
   150    58
   170    75

>> Y = pdist(x,'mahal')

Y =

  Columns 1 through 5

   1.572816369474562   2.201942917264386   1.635800793960578   2.695107559788053   1.478413355546874

  Columns 6 through 10

   1.404831102709996   0.629126547789825   1.713111078598705   1.391260434780810   2.103238561272744

  Columns 11 through 15
```

```
   1.590313263839551   2.103238561272744   1.090736759616727   2.589223001191582   2.033867095735081

  Columns 16 through 20

   1.825496244926879   0.629126547789825   2.743712945526665   2.441925172889290   2.980237487501595

  Column 21

   2.793761753017197
```

其中，X每一行代表一个样例，X是个二维的。Y的第一个数表示x_1与x_2之间的马氏距离。

3.求x_1与x_2之间的马氏距离

```
>> x=[155 66;180 71;190 73;160 60;190 68;150 58;170 75]

x =

   155    66
   180    71
   190    73
   160    60
   190    68
   150    58
   170    75

>> cov=cov(x)

cov =

   1.0e+02 *

   2.702380952380953   0.739285714285714
   0.739285714285714   0.412380952380952

>> s=inv(cov)

s =

   0.007261927639280  -0.013018640484967
  -0.013018640484967   0.047588267151168

>> a=[-25 -5]*s*[-25;-5]
```

```
a =

   2.473751332087140

>> sqrt(a)

ans =

   1.572816369474561
```

4.注意

计算两两马氏距离时，中间的协方差矩阵永远是总体的，而不是这两个的。所以，马氏距离很容易受总体的影响，总体一变化，两个样例之间的马氏距离就会变化。

以下叙述来自：欧氏距离 vs 马氏距离 - bluenight专栏 - CSDN博客 *https://blog.csdn.net/chl033/article/details/5526337*

1）马氏距离的计算是建立在总体样本的基础上的，这一点可以从上述协方差矩阵的解释中可以得出，也就是说，如果拿同样的两个样本，放入两个不同的总体中，最后计算得出的两个样本间的马氏距离通常是不相同的，除非这两个总体的协方差矩阵碰巧相同；

2）在计算马氏距离过程中，要求总体样本数大于样本的维数，否则得到的总体样本协方差矩阵逆矩阵不存在，这种情况下，用欧式距离来代替马氏距离，也可以理解为，如果样本数小于样本的维数，这种情况下求其中两个样本的距离，采用欧式距离计算即可。

3）还有一种情况，满足了条件总体样本数大于样本的维数，但是协方差矩阵的逆矩阵仍然不存在，比如A（3，4），B（5，6）；C（7，8），这种情况是因为这三个样本在其所处的二维空间平面内共线（如果是大于二维的话，比较复杂？？？）。这种情况下，也采用欧式距离计算。

4）在实际应用中“总体样本数大于样本的维数”这个条件是很容易满足的，而所有样本点出现3）中所描述的情况是很少出现的，所以在绝大多数情况下，马氏距离是可以顺利计算的，但是马氏距离的计算是不稳定的，不稳定的来源是协方差矩阵，这也是马氏距离与欧式距离的最大差异之处。

我们熟悉的欧氏距离虽然很有用，但也有明显的缺点。它将样品的不同属性（即各指标或各变量）之间的差别等同看待，这一点有时不能满足实际要求。马氏距离有很多优点。它不受量纲的影响，两点之间的马氏距离与原始数据的测量单位无关；由标准化数据和中心化数据(即原始数据与均值之差）计算出的二点之间的马氏距离相同。马氏距离还可以排除变量之间的相关性的干扰。它的缺点是夸大了变化微小的变量的作用。

5. MATLAB求两个矩阵之间的马氏距离，使用pdist2()函数

```
>> x=rand(4,3)

x =

   0.792207329559554   0.849129305868777   0.743132468124916
```

```
   0.959492426392903   0.933993247757551   0.392227019534168
   0.655740699156587   0.678735154857773   0.655477890177557
   0.035711678574190   0.757740130578333   0.171186687811562

>> y=rand(2,3)

y =

   0.706046088019609   0.276922984960890   0.097131781235848
   0.031832846377421   0.046171390631154   0.823457828327293

>> z=pdist2(x, y, 'mahal')

z =

  11.881392154588022   8.912492295829436
  10.377530870286948   8.703763775002274
   9.513297701500704   6.612259802538707
  10.858334218503852   8.268677052674791
```

其中，数据X是n*d的，数据Y是m*d的，则马氏距离是n*m的矩阵，代表数据X的第i个样例与数据Y的第j个样例之间的马氏距离。