原 协方差矩阵和散布矩阵（散度矩阵）的意义

置顶 2017年03月31日 19:27:42 pan_jinquan 阅读数：12173 更多

协方差矩阵和散布矩阵的意义

【尊重原创，转载请注明出处】http://blog.csdn.net/guyuealian/article/details/68922981

在机器学习模式识别中，经常需要应用到协方差矩阵C和散布矩阵S。如在PCA主成分分析中，需要计算样本的散度矩阵，有的论文是计算协方差矩阵。实质···意义和阵（散度矩阵）前乘以系数1/(n-1)就可以得到协方差矩阵了。

在模式识别的教程中，散布矩阵也称为散度矩阵，有的也称为类内离散度矩阵或者类内离差阵，用一个等式关系可表示为：

关系： 散度矩阵=类内离散度矩阵=类内离差阵=协方差矩阵×（**n-1**）

样本的协方差矩阵乘以*n-1*倍即为散布矩阵，*n*表示样本的个数，散布矩阵的大小由特征维数*d*决定，是一个为*d*×*d* 的半正定矩阵。

一、协方差矩阵的基础

对于二维随机变量（X,Y）之间的相互关系的数字特征，我们用协方差来描述，记为Cov(X,Y)：

$$Cov(X,Y)=E\{[X-E(X)][Y-E(Y)]\}=E(XY)-E(X)E(Y)$$

$$Cov(X,Y)=\frac{1}{n-1}\sum_{i=1}^{n}(X-\bar{X})(Y-\bar{Y})$$

1。通常来说特征值没有一个上限，为了方便量化比较，归一化到 0-1 范围，可以将单个特征值di 用 di/（d1+d2+...+dn）来代替。另外，在数理统计上，协方差矩阵一定是半正定的，半正定矩阵特征值一定不小于0。2。协方差矩阵的特征值度量了三维局部结构沿三个正交方向的延展程度，但这一个正交方向不一定是坐标系XYZ，而是数据分布方差最大的三个方向。也就是说，协方差矩阵特征值不受物体旋转/平移/镜像等刚性变换（Rigid Transformation）的影响。协方差矩阵特征值从小到大排列对应的特征向量指向数据分布的方差从大到小的方向。这些方向（以物体XYZ坐标系为参考）显然受物体旋转等刚性变换的影响。理解这一点，对于3D物体分类识别特征提取特别重要。3。回顾PCA的计算，协方差矩阵奇异值分解（可以理解为特征值分解，只不过特征值分解只适用于方阵）后，奇异值（就是特征值）从大到小排列，对应特征向量重要性（即数据分布方差的大小）从大到小排列。将特征向量矩阵取前k列，与原矩阵相乘，这样的几何意义是将原矩阵投影到k个特征向量上，因为矩阵乘法的意义就是一个变换矩阵作用于另一个矩阵X。

那么二维随机变量(X,Y)的协方差矩阵，为：

$$C_{2\times 2}=\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}=\begin{pmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{pmatrix}$$

对于三维随机变量***X***=（***X₁***, ***X₂***, ***x₃***）的协方差矩阵可表示为：

$$\begin{pmatrix} cov(X,X) & cov(X,Y) & cov(X,Z) \\ cov(Y,X) & cov(Y,Y) & cov(Y,Z) \\ cov(Z,X) & cov(Z,Y) & cov(Z,Z) \end{pmatrix}$$

$$Cov(X,Y)=\frac{1}{n-1}\sum_{i=1}^{n}(X-\bar{X})(Y-\bar{Y})$$

*n*表示样本个数

对于*n*维***X***=（X_1, X_2....X_n）协方差矩阵：

$$C=E\{(X-\mu)(X-\mu)^T\}$$

注意：在二维**协方差**计算公式中，变量 ***X*** 和 ***Y*** 都是**特征矢量**（准确来说应表示为**列向量**的形式：$\vec{X}$ 和 $\vec{Y}$），在统计学中称为**随机变量**。而 *n* 维**协方差矩阵**计算公式中，只含有一个变量 ***X***，此变量 ***X*** 是样本特征构成的**特征矩阵**，类似于二维随机变量 ***X***=[$\vec{X}$，$\vec{Y}$]，其中矩阵 ***X*** 中的每个列分量，实质就是一个特征矢量（随机变量）。

注意协方差与协方差矩阵的区别，协方差是一个值，而所有维度的协方差构成的矩阵才是协方差矩阵

大写 C 表示协方差矩阵，小写 c 表示协方差矩阵 C 的某个协方差 $c_{i,j}$。

说明：

（1）协方差矩阵是一个**对称矩阵**，且是**半正定矩阵**，主对角线是各个随机变量 的方差（各个维度上的方差）。

（2）标准差和方差一般是用来描述一维数据的；对于多维情况，而协方差是用于描述任意两维数据之间的关系，一般用协方差矩阵来表示。因此协方差矩阵计算的是协方差，而不是不同样本之间的。

（3）协方差计算过程可简述为：先求各个分量的均值E(Xi)和E(Xj)，然后每个分量减去各自的均值得到两条向量，在进行内积运算，然后求内积后的总和，最后把总和

例子：设有8个样本数据，每个样本有2个特征：(1,2);(3 3);(3 5);(5 4);(5 6);(6 5);(8 7);(9 8)，那么可以看作二维的随机变量（X,Y），即

$X=[1\ 3\ 3\ 5\ 5\ 6\ 8\ 9]$

$Y=[2\ 3\ 5\ 4\ 6\ 5\ 7\ 8]$

Matlab中可以使用cov(X, Y)函数计算样本的协方差矩阵，其中X,Y都是特征向量。当然若用X表示样本的矩阵（X中每一行表示一个样本，……是一么可直接使用cov(X)计算了。


```
clear all
clc
X=[1,2;3 3;3 5;5 4;5 6;6 5;8 7;9 8]%样本矩阵：8个样本，每个样本2个特征
covX= cov(X)%使用cov函数求协方差矩阵
```

运行结果为：

```
covX =

    7.1429    4.8571
    4.8571    4.0000
```

当然，可以按定义计算，Matlab代码如下：

```
clear all
clc
X=[1,2;3 3;3 5;5 4;5 6;6 5;8 7;9 8] %样本矩阵：8个样本，每个样本2个特征
covX= cov(X)                                %使用cov函数求协方差矩阵
%% 按定义求协方差矩阵：（1）使用分量的方法，先求协方差，再组合成协方差矩阵
meanX=mean(X)        %样本均值
varX=var(X)               %样本方差
[Row Col]=size(X);
dimNum=Row;          %s样本个数size(X,1)=8
dim1=X(:,1);              %特征分量1
dim2=X(:,2);              %而在分量2
c11=sum( (dim1-mean(dim1)) .* (dim1-mean(dim1)) ) / ( dimNum-1 );
c21=sum( (dim2-mean(dim2)) .* (dim1-mean(dim1)) ) / ( dimNum-1 );
c12=sum( (dim1-mean(dim1)) .* (dim2-mean(dim2)) ) / ( dimNum-1 );
c22=sum( (dim2-mean(dim2)) .* (dim2-mean(dim2)) ) / ( dimNum-1 );
C22=[c11,c12;c21,c22]%协方差矩阵

%% 或者（2）直接求协方差矩阵：
tempX= repmat(meanX,Row,1);
C22=(X-tempX)'*(X-tempX)/(dimNum-1)
```

运行结果：

```
covX =

    7.1429    4.8571
    4.8571    4.0000

meanX =

     5     5

varX =

    7.1429    4.0000

C22 =

    7.1429    4.8571
    4.8571    4.0000

C22 =
```

```
    7.1429    4.8571
    4.8571    4.0000
```


说明：从中可以发现，样本的协方差矩阵的对角线即为样本的方差。

二、协方差矩阵的意义

为了更好理解协方差矩阵的几何意义，下面以二维正态分布图为例（假设样本服从二维正态分布）：

（1）均值 u=[0, 0]，协方差矩阵为 $C=\begin{bmatrix}5 & 0\\0 & 1\end{bmatrix}$，则样本分布图的 XOY 平面是椭圆形，主轴方向平行水平 X 轴。


```
clear all;clc
mu=[0,0];          % 均值向量
C=[5 0;0 1]        %样本的协方差矩阵
[V,D] =eigs(C)     %求协方差矩阵的特征值D和特征向量V
 %% 绘制二维正态分布图
[X,Y]=meshgrid(-10:0.3:10,-10:0.3:10);%在XOY面上，产生网格数据
p=mvnpdf([X(:) Y(:)],mu,C);%求取联合概率密度，相当于Z轴
p=reshape(p,size(X));%将Z值对应到相应的坐标上
figure
set(gcf,'Position',get(gcf,'Position').*[1 1 1.3 1])
subplot(2,3,[1 2 4 5])
surf(X,Y,p),axis tight,title('二维正态分布图')
subplot(2,3,3)
surf(X,Y,p),view(2),axis tight,title('在XOY面上的投影')
subplot(2,3,6)
surf(X,Y,p),view([0 0]),axis tight,title('在XOZ面上的投影');
```

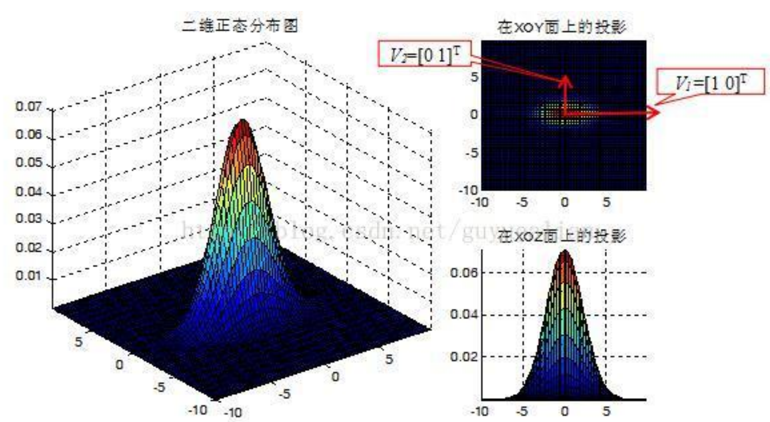
协方差矩阵C的特征值D和特征向量V分别为：

```
V =
     1     0
     0     1

D =
     5     0
     0     1
```

说明：

1)均值[0,0]代表正态分布的中心点，方差代表其分布的形状。

2)协方差矩阵C的最大特征值D对应的特征向量V指向样本分布的主轴方向。例如，最大特征值D1=5对应的特征向量V1=[1 0]^{T}即为样本分布的主轴方向（一般认为是数据次大特征值D2=1对应的特征向量V2=[0 1]^T,即为样本分布的短轴方向。



（2）均值 $u=[0,0]$，协方差矩阵为 $C=\begin{bmatrix}5 & 0\\0 & 5\end{bmatrix}$，则样本分布图的 XOY 平面是一个圆形：

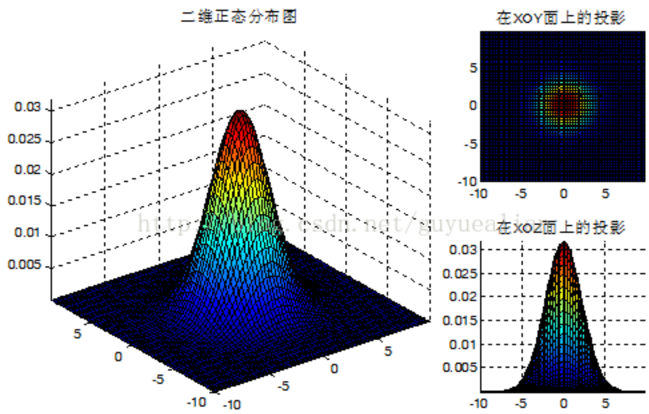
协方差矩阵C的特征值D和特征向量V分别为：

```
V =
     0     1
     1     0

D =

     5     0
     0     5
```

说明：

1）由于协方差矩阵C具有两个相同的特征值D1=D2=5,因此样本在V1和V2特征向量方向的分布是等程度的，故样本分布是一样圆形。

2）特征值D1和D2的比值越大，数据分布形状就越扁；当比值等于1时，此时样本数据分布为圆形。



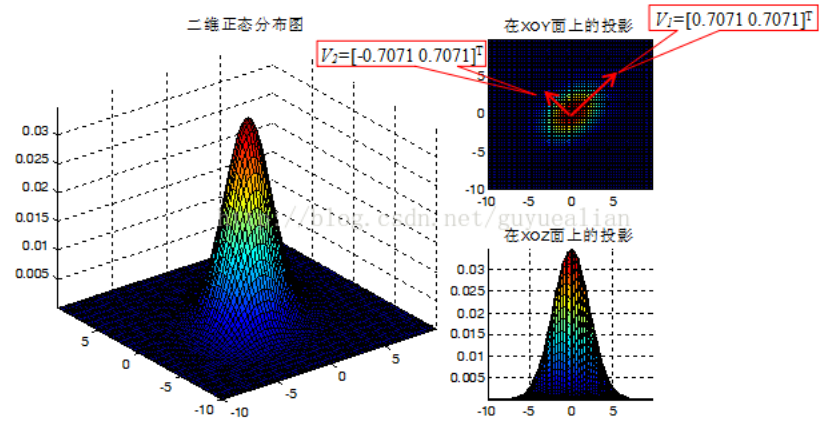
（3）均值 $u=[0,0]$，协方差矩阵为 $C=\begin{bmatrix}5 & 2\\2 & 5\end{bmatrix}$，对角线元素相等，XOY 平面是椭圆形，且向右倾斜 45°。

协方差矩阵C的特征值D和特征向量V分别为：

```
V =
    0.7071   -0.7071
    0.7071    0.7071

D =
     6     0
     0     4
```

说明：

1）特征值的比值D1/D2=6/4=1.5>1，因此样本数据分布形状是扁形，数据传播方向（样本的主轴方向）为V1=[0.7071 0.7071]^T



综合上述，可知：

(1)样本均值决定样本分布中心点的位置。

(2)协方差矩阵决定样本分布的扁圆程度。

是扁还是圆，由协方差矩阵的特征值决定：当特征值D1和D2的比值为1时（D1/D2=1），则样本分布形状为圆形。当特征值的比值不为1时，样本分布为扁形；

偏向方向（数据传播方向）由特征向量决定。最大特征值对应的特征向量，总是指向数据最大方差的方向（椭圆形的主轴方向）。次大特征向量总是正交于最大特征向量方向）。

三、协方差矩阵的应用

协方差矩阵（散布矩阵）在模式识别中应用广泛，最典型的应用是PCA主成分分析了，PCA主要用于降维，其意义就是将样本数据从高维空间投影到低维空间中，并尽间中表示原始数据。这就需要找到一组最合适的投影方向，使得样本数据往低维投影后，能尽可能表征原始的数据。此时就需要样本的协方差矩阵。PCA算法就是求出这堆差矩阵的特征值和特征向量，而协方差矩阵的特征向量的方向就是PCA需要投影的方向。

关于PCA的原理和分析，请见鄙人的博客：

《PCA主成分分析原理分析和Matlab实现方法》：http://blog.csdn.net/guyuealian/article/details/68487833

如果你觉得该帖子帮到你，还望贵人多多支持，鄙人会再接再厉，继续努力的~