Subject: Re: Cluster analysis
Posted by Harald Frey on Mon, 06 Dec 2004 22:21:00 GMT
View Forum Message <> Reply to Message

My original data set it too large to be posted but I used an earlier posting about cluster analysis to create a little demonstration. My program follows this posting. It uses a function gauss2 by Craig Markwardt that I include in case you don't have it. In case my program does not make it though the email, it can be accessed at our anonymous ftp site

ftp sprite.ssl.berkeley.edu
cd pub/hfrey/idl/
get cluster_play_10.pro

I create 10 clusters, but IDL find only 8 correctly and puts the remaining 2 clusters close to [0,0] where there are no original data points.

Harald

```
 ;============================================================ ======
;+
; NAME:
;   GAUSS2
;
; AUTHOR:
;   Craig B. Markwardt, NASA/GSFC Code 662, Greenbelt, MD 20770
;   craigm@lheamail.gsfc.nasa.gov
;
; PURPOSE:
;   Compute Gaussian curve given the mean, sigma and area.
;
; MAJOR TOPICS:
;   Curve and Surface Fitting
;
; CALLING SEQUENCE:
;   YVALS = GAUSS2(X, Y, [XCENT, YCENT, SIGMA, PEAK])
;
; DESCRIPTION:
;
;  This routine computes the values of a Gaussian function whose
;  X-values, mean, sigma, and total area are given.  It is meant to be
;  a demonstration for curve-fitting.
;
;  XVALS can be an array of X-values, in which case the returned
;  Y-values are an array as well.  The second parameter to GAUSS1
;  should be an array containing the MEAN, SIGMA, and total AREA, in
```

```
;  that order.
;
; INPUTS:
;   X - 2-dimensional array of "X"-values.
;   Y - 2-dimensional array of "Y"-values.
;
;   XCENT - X-position of gaussian centroid.
;   YCENT - Y-position of gaussian centroid.
;
;   SIGMA - sigma of the curve (X and Y widths are the same).
;
;   PEAK - the peak value of the gaussian function.
;
; RETURNS:
;
;   Returns the array of Y-values.
;
; EXAMPLE:
;
;   p = [2.2D, -0.7D, 1.4D, 3000.D]
;   x = (dindgen(200)*0.1 - 10.) # (dblarr(200) + 1)
;   y = (dblarr(200) + 1) # (dindgen(200)*0.1 - 10.)
;   z = gauss2(x, y, p)
;
;   Computes the values of the Gaussian at equispaced intervals in X
;   and Y (spacing is 0.1).  The gaussian has a centroid position of
;   (2.2, -0.7), standard deviation of 1.4, and peak value of 3000.
;
; REFERENCES:
;
; MODIFICATION HISTORY:
;   Written, 02 Oct 1999, CM
;
;-

function gauss2, x, y, p, _EXTRA=extra

u = ((x-p(0))/p(2))^2 + ((y-p(1))/p(2))^2
mask = u LT 100
f = p(3) * mask * exp(-0.5D * temporary(u) * mask)
mask = 0

return, f
end

 ;============================================================ =======
pro Cluster_play_10
; program to create 10 clusters and try to find them all
```

```
; Harald Frey, December 6, 2004

FORWARD_FUNCTION gauss2

x = findgen(1000)*0.1 - 50. & y = x
xx = x # (y*0 + 1)
yy = (x*0 + 1) # y

; create 10 two-dimensional clusters
z = 30 * gauss2(xx, yy, [ 20D, 33D, .2, 1]) + $
10 * gauss2(xx, yy, [-30D,-13D, .2, 1]) + $
40 * gauss2(xx, yy, [-20D, 21D, .2, 1]) + $
10 * gauss2(xx, yy, [-10D,-11D, .2, 1]) + $
20 * gauss2(xx, yy, [-13D,  1D, .2, 1]) + $
10 * gauss2(xx, yy, [ 23D, 21D, .2, 1]) + $
30 * gauss2(xx, yy, [ 33D,-31D, .2, 1]) + $
10 * gauss2(xx, yy, [  3D,-41D, .2, 1]) + $
50 * gauss2(xx, yy, [ -3D, 11D, .2, 1]) + $
20 * gauss2(xx, yy, [ 10D, -2D, .2D, 1])
zi = floor(z)   ;; Convert to integer

;; Find the positions of significant data points
wh = where(z GT 5, ct)
if ct EQ 0 then message, 'ERROR: no signif points!'
xi = x(wh MOD 1000)
yi = y(floor(wh/1000))
xy = transpose([[xi],[yi]])

; input for cluster analysis
array=xy
; number of clusters
nc=10
; number of iterations
ni=10

; do cluster analysis
weights=clust_wts(array,n_clusters=nc,n_iterations=ni)

; display original distribution
window,0
plot, xi, yi, psym=3,title='Original distribution'

; display distribution with centers of clusters
window,1
plot, xi, yi, psym=3,title='10 clusters'
oplot, weights(0,*), weights(1,*), psym=1, symsize=3
print,'Cluster centers'
print,weights
```

END