Subject: Re: clustering
Posted by Vince Hradil on Mon, 23 Jul 2007 18:22:42 GMT
View Forum Message <> Reply to Message

On Jul 23, 12:14 pm, little davey <dave-kel...@cox.net> wrote:

```
>  Is it the case that you MUST use standardize() before you call
>  CLUST_WTS()?  The documentation does not say so, but I suspect from
>  the code, and, I actually tried it with the initial poster's data, and
>  got 4 clusters (he wanted 5, but this would appear to be a tough data
>  set to cluster, as variables "2" and "3" are close to each other).
>
>  As I posted an hour ago, part of the source code for CLUST_WTS() is:
>
>    ;Normalized uniformly random cluster weights.
>    ;;Weights = RANDOMU(SEED, N_Variables, N_Clusters) + Zero
>        av1 = average(array[0,*])
>        av2 = average(array[1,*])
>    Weights = RANDOMU(SEED, N_Variables, N_Clusters) + Zero
>    for k = 0L, N_Clusters-1 do Weights[*,k] = $
>      (Weights[*,k] / TOTAL(Weights[*,k])) * Variable_Wts
>
>  However, the variables AV1 and AV2 are NOT USED ANYWHERE IN THE CODE,
>  so I suspect that the data is not "normalized" correctly in
>  CLUST_WTS.  The use of STANDARDIZE() may be necessary for CLUST_WTS to
>  work.
>
>  -- Dave K --
```

No, you just have to be careful about your data. Here's an example:

```
IDL> a=[[3,55],[4,54],[8,55],[9,56]]
IDL> plot, a[0,*], a[1,*], psym=4
```

"obviously", there are two clusters.

```
IDL> wts= clust_wts(a,n_clusters=2)
IDL> print, wts
     1.99307      27.5069
     6.42612      55.0837
IDL> oplot, wts[0,*], wts[1,*], psym=5
(one of the points doesn't even show up, oh, and fix the xrange, too)
IDL> plot, a[0,*], a[1,*], psym=4, xrange=[0,60], yrange=[0,60]
(uh-oh, maybe there is only one cluster)
IDL> oplot, wts[0,*], wts[1,*], psym=5
IDL> print, cluster(a,wts,n_clusters=2)
          1
          1
          1
```

```
           1
```

D'oh!

NOW:

```
IDL> sa = standardize(float(a))
IDL> plot, sa[0,*], sa[1,*], psym=4
IDL> swts= clust_wts(sa,n_clusters=2)
IDL> print, swts
    -0.833454    -0.669169
     0.864960     0.669170
IDL> oplot, swts[0,*], swts[1,*], psym=5
IDL> print, cluster(sa,swts,n_clusters=2)
           0
           0
           1
           1
```

Tada!

The reason the first (unstandardized) doesn't work is that the overall variance is much larger than the variance within each coordinate, so the clusters get "attracted" (if you will) to one of the coordinates and disregards the other.

You have to "stay in touch" with your data - Black boxes are okay, as long as you know what's going on inside.