Subject: K-Mean clustering
Posted by Fabzi on Wed, 20 Feb 2013 17:38:01 GMT
View Forum Message <> Reply to Message

Dear IDLers,

I am doing clustering using CLUST_WTS and CLUSTER I have a doc-related problem. The doc for CLUST_WTS says:

"The CLUST_WTS function computes the weights (the cluster centers) of an n-column, m-row array, where n is the number of variables and m is the number of observations or samples. CLUST_WTS uses k-means clustering. With this technique, CLUST_WTS starts with k random clusters and then iteratively moves items between clusters, minimizing variability within each cluster and maximizing variability between clusters"

From what I know from my textbook (Wilks, "Statistical Methods in the Atmospheric Sciences") k-mean clustering does work iteratively and computes the centroids, and the centroids are computed as the mean of the clustered population. Now, this "mean" value can be defined according to any arbitrary distance measure, which in the case of CLUST_WTS, is not clearly defined. So I assume it to be the euclidian distance, and therefore I would assume that the cluster centers computed by CLUST_WTS are also the mean values of my cluster populations, but this is not *exactly* the case, no matter how many iterations I use. They are close enough, sure, but not equal.

What am I missing?

Thanks,

Fab

For those interested, here is a programm (using cg). It is enough to run it a few times to see what I mean:

```
pro test_cluster

  ; Construct 3 separate clusters in a 3D space:
  n = 50
  c1 = RANDOMN(seed, 2, n)
  c1[0, *] -= 3
  c2 = RANDOMN(seed, 2, n)
  c2[0, *] += 3

  array = [[c1], [c2]]
```

```
   ; Compute cluster weights, using three clusters:
   weights = CLUST_WTS(array, N_CLUSTERS = 2, N_ITERATIONS=100)
   ; Compute the classification of each sample:
   result = CLUSTER(array, weights, N_CLUSTERS = 2)

   pc1 = where(result eq 0)
   pc2 = where(result eq 1)

   cgWindow
   cgPlot, array[0,*], array[1,*], /ADDCMD, /NODATA
   cgPlotS, c1[0,*], c1[1,*], /ADDCMD, COLOR='red', PSYM=1
   cgPlotS, c2[0,*], c2[1,*], /ADDCMD, COLOR='blue', PSYM=1

   cgPlots, weights[0, 0], weights[1, 0], PSYM=16, COLOR='black',
/ADDCMD, SYMSIZE=2
   cgPlots, weights[0, 1], weights[1, 1], PSYM=16, COLOR='black',
/ADDCMD, SYMSIZE=2
   cgPlots, mean(array[0,pc1]), mean(array[1,pc1]), PSYM=17,
COLOR='black', /ADDCMD, SYMSIZE=2
   cgPlots, mean(array[0,pc2]), mean(array[1,pc2]), PSYM=17,
COLOR='black', /ADDCMD, SYMSIZE=2

end
```