Data Biography for Dataset IRIS

The dataset *iris.xls* comes from the Matlab kmeans-cluster toolbox. ( *iris-with-guesses.xls* is an experiment to see what possible combination of features could nail down the three species of iris flowers.)

It contains 150 measurements of the length and width of sepals and petals of a variety of iris flower species.

See <https://en.wikipedia.org/wiki/Iris_flower_data_set> for the origin of the data set

Matlab Code to load and cluster using kmeans:

[NUM,TXT]=xlsread('iris.xlsx');

IDX = kmeans(NUM, 3);silhouette(NUM(:,1:3),IDX,'cityblock');figure(1)

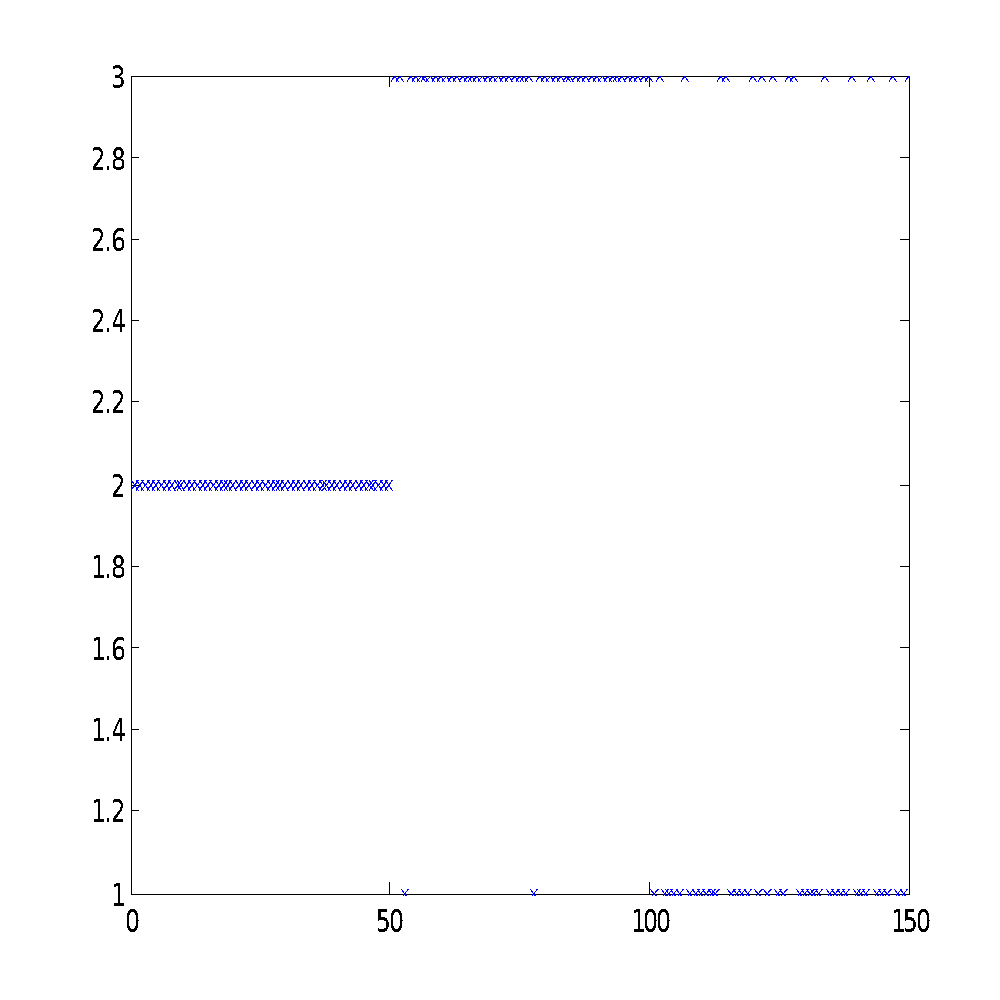
Here is the resulting clustering after executing the kmeans command several time (each time produces a different trial, this one looked pretty good:



Here is the plot of the index – it should be just 1, 2, 3

figure(2);plot(IDX,'x')

Notice how species 1 and 3 are not well separated:

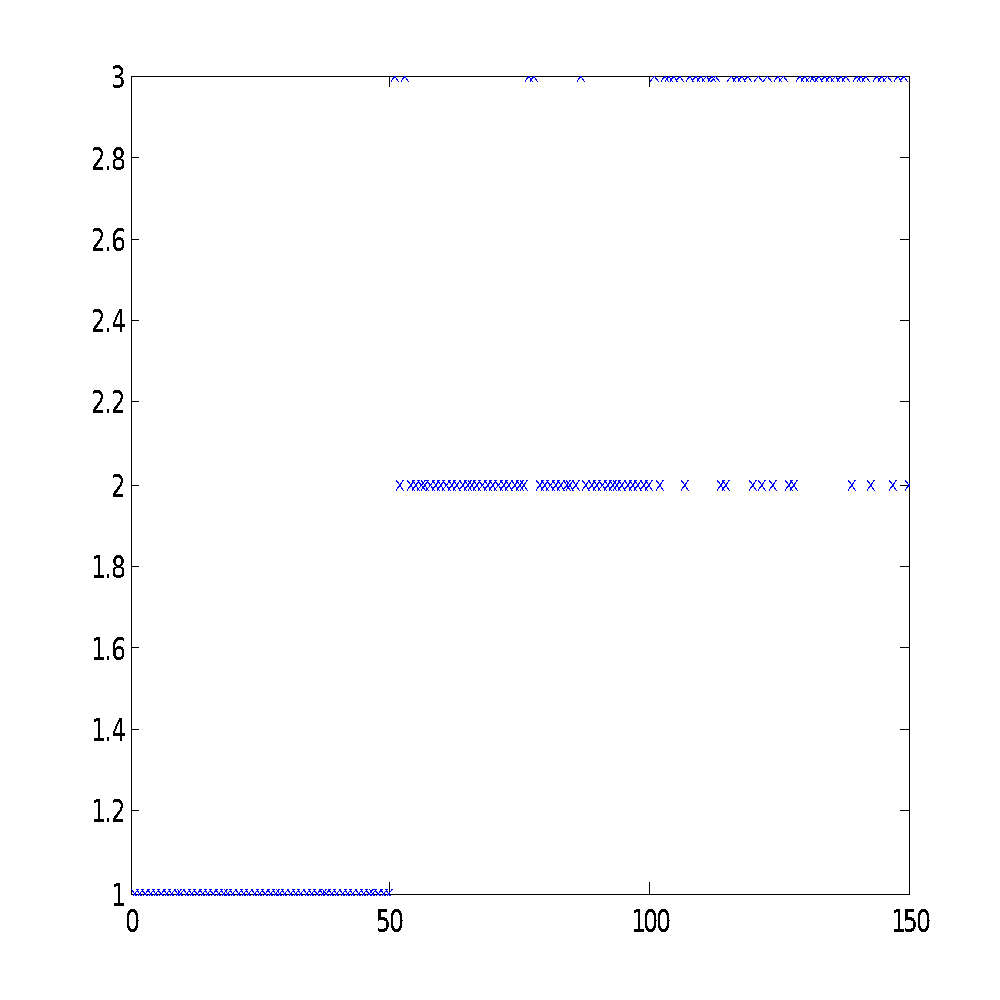


Here is how the SOM Toolbox kmeans clustering did (similar to the Matlab kmeans, had to keep running:

>> [codes,clusters,err] = som\_kmeans('batch', NUM(:,(1:3)), 3, 10);

>> silhouette(NUM(:,1:3),clusters,'cityblock');figure(1)





With the som\_kmeans, you can switch from ‘batch’ to ‘seq’. Also you can increase the number of training epochs from 10, in the example above to 1000000, but it takes a long time. However, it seems to find closer fits… I recall the SPSS kmeans produced the same results every time you ran it, but I don’t recall what all the parameters were back then.