Soft Computing
Kai Goebel/Bill Cheetham
Homework 6: Clustering exercise (10 points)
due: Oct. 23th at the beginning of class (no later than 6:10pm)

The botanist believes that besides petal length and petal width the features sepal length and sepal width would also help to distinguish the species iris. He wants you to determine if this is true or not. Since setting up a rule-base manually becomes somewhat cumbersome for larger number of features, you are asked to implement a clustering tool that finds clusters in data. The botanist has again provided you with the knowledge about what kind of flower each data sample is (there are 3 different classes). You can download the data file from the web site http://www.cs.rpi.edu/courses/fall01/soft-computing/hw/iris4D.dat. The features are in the following order: sepal length (x_1), sepal width (x_2), petal length (x_3), petal width (x_4), correct class assignment (in column 5). Do not use the last column for your clustering task. It is there to evaluate the error only.

In order to cluster the data (in 4-D) use the fuzzy c-means algorithm (c=3) with two different values for parameter "m" (1.001, 2). Show the evaluation for both values of m. Evaluate the error (number of misclassifications) for all cases. Use matlab to plot the cluster assignment color coded in the x_1-x_2, x_1-x_3, x_1-x_4, and x_2-x_3 planes. Discuss the results. Express the classification using linguistic terms.

Note that the assignment of the classes (1,2,3) in the data sets is arbitrary. Therefore, a switched class name (which will likely happen in the clustering approach) is not considered an error. Automated matching of the class names is not required.

What to hand in:

- Hard copy of your 8 graphs (4 plots for each value of m using the format of homework 3) and the associated classifications and misclassifications for the value of "m" using the tabular enumeration similar to homework 3.
- Linguistic interpretation of the clusters (spell out some rules)
- Discussion

The data file looks as follows (without the header):

sepal length	sepal width	petal length	petal width	class
5.1000000e+00	3.5000000e+00	1.4000000e+00	2.0000000e-01	1.0000000e+00
4.9000000e+00	3.0000000e+00	1.4000000e+00	2.0000000e-01	1.0000000e+00
4.7000000e+00	3.2000000e+00	1.3000000e+00	2.0000000e-01	1.0000000e+00
4.6000000e+00	3.1000000e+00	1.5000000e+00	2.0000000e-01	1.0000000e+00
5.0000000e+00	3.6000000e+00	1.4000000e+00	2.0000000e-01	1.0000000e+00
5.4000000e+00	3.9000000e+00	1.7000000e+00	4.0000000e-01	1.0000000e+00
…				

General submission guidelines:

Please submit your homework solutions (excluding source code) including all graphs, numerical output, and supporting information as a hard copy. Submit source code electronically to the email of the instructors (goebel@cs.rpi.edu, cheetham@cs.rpi.edu). Zip up the source code files and use your last name as the archive name. Identify the homework in the subject line.

Part 2: Soft Computing Homework 7: Starting Your Project (3 points)
Due: Thursday October 30th

Describe the scope of your project. Include specific issues that you will and will not address in your project.

What is the input to your project?
Include a few examples

What is the output of your project?
Include an example for each input above

What are the data sources you are using?
Include a few examples and a description so we can understand the data

How does your project work?
What elements of soft computing are you using
Create a top level diagram describing your project (like on Jang figure 22.6)

How are you going to test your project?
Include a few sample tests