CS4

Hilary Term Assignment 2000

k-Nearest Neighbour Classifier

This assignment should be submitted as a brief report and a code listing. The report should show the application in operation. It may be attempted by groups of two.

Part A. *(60%)*

Develop a *k*-Nearest Neighbour classifier that will work with the Iris data-set and optionally with other datasets that contain examples with numeric feature values. The *k*-NN classifier will read in the data from a file and will read in a target example from the keyboard. It will find the *k* nearest neighbours in the data and predict the category of the target based on these neighbours.

Table 1. Sample data from the Iris Data-set

Feature 1	Feature 2	Feature 3	Feature 4	Category
0.575	0.45	0.125	0.025	0
0.6375	0.4125	0.2125	0.0625	0
0.6	0.425	0.2375	0.025	0
0.625	0.375	0.2	0.025	0
0.875	0.4	0.5875	0.175	1
0.8	0.4	0.5625	0.1875	1
0.7375	0.4	0.6	0.225	1
0.7625	0.35	0.5	0.1625	1
0.8	0.3625	0.5375	0.1625	1
0.825	0.375	0.55	0.175	1
0.7875	0.4125	0.75	0.3125	2
0.725	0.3375	0.6375	0.2375	2
0.8875	0.375	0.7375	0.2625	2
0.7875	0.3625	0.7	0.225	2
0.8625	0.4	0.7125	0.2875	2

To demonstrate this in operation, remove 9 examples (3 from each category) from the file and test the classifier on these.
The Iris data is available from the CS4 web page.

Part B. *(40%)*

Use a basic Condensed Nearest Neighbour technique to reduce the size of the data-set as much as possible. Report the size of the final data-set and its classification accuracy on the removed examples.