Assignment 7

Assigned: Thursday Apr 19, 2001
Due: Friday May 4, 2001

For this assignment, you have two choices: you can create a mini-project of your own choosing, or you can enter the classifier competition.

Mini-project

For the mini project, I want you to pick a tractable yet interesting supervised learning problem and construct a solution using either the software you've developed during the semester or by writing software to implement other ML algorithms. Your work should include the following steps, which you should describe in a brief write up:

1. Obtain a data set
2. Divide into training and test data
3. Determine a representation of the input and the output that is suitable to the problem. Use any knowledge you have of the domain to derive a representation in which relevant information about the domain is encoded explicitly, and will thus be easier for the ML system to exploit.
4. Determine an ML technique appropriate for the problem, and specialize it if possible based on your knowledge of the domain. By "specialize," I mean to design some type of bias into the algorithm that is an appropriate bias for the domain.
5. Perform model selection to determine the best ML model based on your training data.
6. Evaluate performance of the ML system on the test data.

Classifier competition

For the classifier competition, I will provide you with preprocessed speech data and your task is to classify an utterance into a response class indexed from 1 to 10. The training data consists of 4163 samples, and the test data consists of 5730 samples, with roughly equal numbers in each class. Your task is to develop the best classifier based on the training data, and then to send me predictions for the test set (which is not labeled). I will report to you the accuracy your classifier obtains on the test set. You can use the training data however you like to obtain the best classifier. The steps you take should be similar to those for the mini-project.

Each training example will look something like this:

```
train4163 10
023 045 070 084 078 061 047 006 008 030
118 087 085 079 085 072 068 108 112 070
053 034 038 028 049 039 053 055 060 055
055 084 074 060 079 085 097 092 075 108
079 115 113 112 112 117 114 090 087 110
063 114 115 122 118 118 092 054 037 033
021 034 050 057 055 042 033 031 036 048
039 028 013 005 012 023 041 043 035 035
068 097 107 109 107 099 097 080 082 083
44
```

The label "train4163" is an identifier for the pattern. The next number, 10, is the class label (1-10) for the example. The data then consists of 9 rows of 10 numbers, followed by a single number. Each number is between 0 and 127. The columns in this array correspond to time slices: columns on the left are the beginning of an utterance and columns on the right are the end of the utterance. The rows in this array corre-

spond to different features extracted from the speech signal. The final number is another feature of utterance, but doesn't depend on time. This is all I am going to tell you about the patterns. The only domain knowledge I have provided you with is that adjacent columns correspond to adjacent points in time. You may be able to use this to bias your classifier.

Each test example looks the same as the training examples, except that the class label is replaced by "0". You should submit to `mozer@cs.colorado.edu` a list of class labels for the examples, in the order that they appear in the test set, with one label per line. I will report back to you the accuracy of your classifier, and your ranking among the other submissions, including one of my own.

You can obtain the data at:

```
ftp://ftp.cs.colorado.edu/users/mozer/5622/competition_train.dat.gz
ftp://ftp.cs.colorado.edu/users/mozer/5622/competition_test.dat.gz
```