CSCE 970: Pattern Recognition **Spring Semester, 1999**

Homework Assignment 2

March 26, 1999 *Due Date: April 23*

On this homework you may work with a partner if you wish. **This is optional!** If you do work with a partner, submit a single homework with both your names on it. You and your partner will receive the same grade. Note that if you work with a partner, then my expectations during grading will be higher.

Consider doing some of the optional items on this homework. Five bonus points on this homework (the minimum amount for full credit of each optional item) is worth approximately 0.91% to your final grade. Of course, first focus on the core parts of this homework (especially your report), and do not neglect your project, as these are worth much more to your final grade.

1. (45 pts) Implement or find on the web an implementation of a nonlinear classifier such as a multi-layer neural network or a decision tree algorithm. You may also use the tools from the PRW software package if you wish (some of you have already installed this; come see me if you wish to try it). If you search the web for a tool, then the following URLs might be useful. Note that I have not seen any of the following myself, so I can neither endorse them nor comment on their ease of use.

   URLs for neural networks (thanks to Jamil Saquer, Gautam Jindal, Menghua Fu, and Luyin Cao):

   `http://www.ph.tn.tudelft.nl/~sprlib/`
   `http://www.ph.tn.tudelft.nl/PRInfo/PRInfo/software.html`
   `http://www.bgif.no/neureka/products.html`
   `http://www.emsl.pnl.gov:2080/proj/neuron/neural/systems/shareware.html`
   `http://www.itl.atr.co.jp/comp.speech/Section6/Recognition/nico.html`
   `http://www.isbiel.ch/I/Projects/janet/jaNet-1.0beta1.html`
   `http://www.suhep.phy.syr.edu/courses/modules/MM/Neural/Neur_train.html`
   `http://www.cs.bgu.ac.il/~omri/Perceptron [this might only be a perceptron]`
   `http://www.trajan-software.demon.co.uk/sharewar.htm`

   URL for decision trees (a source for C4.5, a famous decision tree algorithm).

   `http://www.cse.unsw.edu.au/~quinlan/`

   The above code and a description are also available in the book *C4.5: Programs for Machine Learning* by J. Ross Quinlan. If you check this book out at the library, please be courteous to your fellow students who also wish to implement a decision tree and do not keep the book too long (or better yet, share it with the class).

   If you download a package from the web or use PRW, state exactly where you got your tools and in your report give an overview of how to use them. If you write your own tool, describe in your report how to use it and e-mail me your code.

   (a) Run your nonlinear classifier on the same data sets from homework 1, varying parameters whenever possible. Then contrast your results with those of the classifiers from homework 1 with respect to error *and* training and testing time. In your report, you do not need to re-report all your performance numbers from homework 1, but you should summarize some of the numbers to indicate interesting performance differences between the old classifiers and your new one.

   (b) Then run your new classifier on new data sets that I will make available on the web page and on cse (I'll send you e-mail when they are ready). These data will have more than two classes, so keep that in mind when you select your nonlinear classifier implementation.

   Optional Item (worth 5 extra pts): Run your classifiers from homework 1 on the new (multiclass) data and contrast their performance with that of your new classifier.

   Optional Item (worth 5 extra pts): When you measure the error rates of your new classifier and contrast them with the rates from the old classifiers, use hypothesis testing techniques to give confidence intervals on the error rates and/or a confidence level that one classifier's error rate is statistically significantly less than another's.

Optional Item (worth 5–10 extra pts): Implement a simple feature selection scheme from Chapter 5 and apply it to the old and new data before running your classifiers. To receive the full 10 points, you must also report on how the feature selection process affected the time complexity and error rates of your classifiers, meaning that you must still run the classifiers with the full-featured (original) data in order to contrast performance.

Optional Item (worth 5–10 extra pts): Implement a boosting algorithm to run with the classifiers (old and new) and report on how boosting affects error on the training and testing sets, including discussion on whether or not overtraining occurred. For the 5 points, you need only run boosting with 2 or 3 classifiers (e.g. Perceptron and Winnow) and thoroughly report the results. To be eligible for the full 10, you must run boosting on all classifiers, old and new.

Optional Item (worth 5–10 extra pts): In your experiments, use only the training sets for training and testing. Here you need to use some of the holdout and resampling techniques described in Lecture 6. Doing this for only the new classifier is worth up to 5 points, also performing it on the homework 1 classifiers is worth up to another 5. To receive full credit, you should vary the total size of the data set and vary the amount of holdout (i.e. the value of k in leave-k-out). Compare your generalization performance estimates with those based on the full test set. How similar are they?

As with homework 1, much weight will be given to your presentation. If you do any of the above optional items, you must report on them well in order to receive full credit. Also as with homework 1, innovative ideas, approaches, and conclusions will be rewarded.

2. (45 pts) Implement (or load from the web) *one* of the following clustering algorithms: MBSAS (Section 12.4), TTSAS (Section 12.5), the complete link algorithm or a related algorithm (Section 13.2.2), or the Isodata algorithm (Section 14.5.1). Run the algorithm on the data sets that I will make available on the web page and on cse (I'll send you e-mail when they are ready). You may assume these data form compact clusters, so use point representatives in your algorithms if representatives are required in your choice of proximity measure. The data will be two-dimensional, so you may visually inspect the results of the clusterings.

   You may choose any appropriate proximity measure for your algorithm, and are only required to use one. If you run MBSAS or TTSAS, then you should vary the thresholds Θ and q and the order to which the vectors are presented to the algorithm. For each combination, graphically display the resultant clustering and comment on how "good" you think it is based on visual inspection. If your run the complete link (or a closely related) algorithm, then you should graphically display the clusterings you get by cutting the dendogram at several different levels (displaying the actual dendograms is optional), and then comment on the visual "goodness" of each clustering. If you run Isodata, then run the algorithm with different choices of initial $\boldsymbol{\theta_j}$'s, and display the resultant clusterings, assessing their visual "goodness". No matter which algorithm you use, you should also report timing information. Given the amount of time taken and "goodness" of the best cluster, do you feel that your algorithm was a good choice for the given data sets? Describe all your results in a detailed report, separate from your report for problem 1.

   Optional Item (worth 5 extra pts): Run your clustering algorithm with other proximity measures (i.e. other functions $d(\cdot,\cdot)$ or $s(\cdot,\cdot)$). Contrast the results you get (time and goodness of clustering) with your other clustering results.

   Optional Item (worth 5 extra pts): Use the procedures in Sections 12.3.1 and 13.5 and Chapter 16 to determine the "best" number of clusters in the data (your choice of procedure will be determined by the clustering algorithm you use).

   Optional Item (worth 5 extra pts): Run another clustering algorithm on the data, possibly one not covered in class, such as a fuzzy clustering algorithm. (You might try a refinement procedure from Section 12.6.) Contrast your new results (time and goodness of clustering) with your other clustering results.

   Optional Item (worth 10 extra pts): Run some of the cluster validity techniques from Chapter 16 on your clustering results.

3. (5 pts) Describe at least two things covered in lecture and/or in the textbook that you do not fully understand, excluding questions you asked on homework 1.

4. (5 pts) State the approximate amount of time you spent on each problem.