FALL 2009

THEORY & APPLICATIONS OF PATTERN RECOGNITION

ADVANCED TOPICS IN PATTERN RECOGNITION

Midterm Exam II

Due December 1, 2009.

INSTRUCTIONS

Since this is an exam, you are expected to work alone. Given that you will take this exam home and have about 24-48 hours to work on it, it is not easily possible for me to enforce / police that you will work on your own. Therefore, I ask that you honor this policy (and sign the below disclaimer before returning your work) so that I can fairly evaluate everyone. All answers should be your individual work!

Academic dishonesty, in any form, will not be tolerated and will be dealt with according to university rules and regulations, though you can safely expect, at a minimum, a zero from this exam and possibly failing this class altogether.

I, ______________________________, certify that I have worked alone on all parts of this exam. I have not received any help from any other student or other person, nor any unauthorized help from online sources, nor have I presented any work created / done by other people as mine, nor have I provided any help to anyone of any nature. I understand that doing any of the above mentioned acts constitutes academic dishonesty, and accept the consequences of such an action.

Sign: X_________________________________
THEORY & APPLICATIONS OF PATTERN RECOGNITION

ECE 09.555
ADVANCED TOPICS IN PATTERN RECOGNITION

MIDTERM II

This is an open ended design problem. Please find the dataset on class webpage. Download the zip file and unzip the two files in it. You will need a password, which is the same as other class documents.

The two datasets include various features obtained from MRI images of two cohorts (groups) of people. The nature of the cohort is irrelevant for the purposes of this exercise, so let’s called them Group 1 (class 1) and Group 2 (class 2). The two datasets actually come from overlapping people, that is, some of the same people had both types of MRI images obtained, and hence appear in both files. Obviously, these people have the same label.

The datasets are Matlab structs, including the fields of info, labels and rawdata. The info field has two subfields, IDs and features. You will be able to determine which subjects are common to both files from the IDs, and the name of attributes from the features. Read the feature names carefully, do a background search for using MRI data on neurological disorders and what type of features are used (generally, you will find that the features are often brain volumes of different parts of the brain, referred to as regions of interest). Study the features, the numbers, their nature, etc. very carefully before proceeding to determine what – if any – preprocessing you want to do. (Hint: For example, Feature #29 in dataset 1, may be used to normalize the first 28 features).

You goal is to use whatever preprocessing, data analysis, classification algorithm you wish, or think appropriate, to get the best classification performance from these two datasets. Ordinarily, you will need to
1) determine some preprocessing (normalizing, computing additional features, PCA, etc.) approach;
2) a classification algorithm (or a combination of them used as an ensemble of classifiers); and
3) an evaluation process to make sure that your estimate is an appropriate one.

It is up to you what algorithm you want to use. There is no single correct answer here. You can use an algorithm we have seen in class (naïve Bayes, FLD, kNN, MLP, SVM, RBF, decision trees, any of which can be put in an ensemble system using bagging, AdaBoost.M1, AdaBoost,M2, stacked generalization, mixture of experts, etc.), or you can pick something else. Your evaluation process must be defensible, that is some form of cross validation (3-fold, 5-fold, 10-fold, leave-one-out, etc. you need to explain which one you choose, and why, and how you implemented it).

Undergraduate students: Pick three algorithms of your choice, implement, test, evaluate and report your findings on each dataset separately. You may need to try different algorithms before you pick your three. You may choose to do the graduate student work for an addition 15% credit.

Graduate students: In addition to analyzing each dataset separately (as specified for undergraduate students), investigate whether an improved performance can be obtained by combining the datasets using an ensemble or feature combination method (on the subset of subjects that have data in both databases).
What to submit:
1) Your report explaining your reasoning, the outcome of your lit review on MRI processing, how did you preprocess the features and why (if any, it may or may not be necessary), which classifiers did you pick and why, etc., along with your results and their interpretation;

2) Fully documented code. Your code should be organized such that one Main file should be run from the command line with no arguments. You must determine and fix the appropriate parameters of the algorithm you have chosen (e.g., if you use SVM, which kernel and what C parameter –and why!). Assuming the data is in the same directory, your code should run just by calling it (and of course, the main function may call other sub functions which should be placed either in the main function, or as separate functions placed in the same directory).

Grading: The person who gets the best – verifiable and defendable - performance in each dataset will get 15 bonus points. Other than that, the exact performance will not determine the grade, so long as you have a reasonable and defendable approach, and that the code runs. Much of your grade will come from your reasoning for choice of the subcomponents, correct implementation of these components and defendable interpretation of your results.