Pattern Classification and Recognition
ECE 681

Spring 2016
Course Project

Deliverable Due Dates

Project Proposal ..... 4:30PM Thursday, February 18
Project 25% Status Report ..... 4:30PM Thursday, March 10
Project 50% Status Report ..... 4:30PM Thursday, April 7
Project Final Report ..... 12PM NOON Monday, May 2

Overview

Many would argue that the best way to learn about and understand pattern classification and recognition algorithms is to implement and apply the algorithms and techniques yourself. The purpose of our course project is to give you an opportunity to explore elements of pattern classification and recognition, including data pre-processing, feature generation/extraction, and feature selection in addition to classifier design, within the context of an application area that is interesting to you. (Building a working system is also a lot more fun than solving pencil-and-paper homework problems or taking exams, not to mention the opportunity it gives you to flex your creative muscles!)

Course projects will be completed *individually*, however, proposals from groups of at most 2 students will be considered if the proposed project represents a *significantly more challenging undertaking*. That is, a group of two students will be expected to complete a project that requires approximately twice the effort of a project completed by a student individually. In these cases, both students in the group will be required to complete a team assessment form and individually document (about 1 page) his/her individual contributions to the project. This documentation is in addition to the jointly authored project final report. The information in the team assessment forms and individual documentations of contributions to the project may be used to adjust individual team member scores for the project (*i.e.* individual team member scores for the project may not be identical to the team score for the project).

Your course project should address an interesting and meaningful problem. Your inspiration for your project may come from your future career goals (*i.e.*, your target industry or employer – this is an opportunity to complete a project you can showcase to prospective employers), research you are conducting with a faculty advisor, a personal "pet-project" or hobby you have, or some other past or current life experience. While most of the information here is focused on projects that build a classifier for a classification problem of interest to you, projects addressing more theoretical questions or comparative studies or that are based on simulated data are also perfectly acceptable. Be creative about the classification problem you want to address or the comparative or theoretical study you want to perform – the idea is for this project to be something that engages you!

The publicly available databases listed on the last page contain numerous datasets. If you are having a difficult time developing a project proposal, look through those databases to find a topic that interests you there and propose a project for the dataset of your choice. Some examples of projects using these datasets are: developing a full classification system that requires feature generation/extraction, feature selection, and classifier optimization; comparing classifiers; comparing feature selection methods; or comparing feature generation (dimensionality reduction) techniques.

Project Expectations

The final report is expected to be complete, as would be a report delivered to a customer who hired you to complete this project. There are no page requirements or limits – the final report should be parsimonious (as long as it needs to be to fully describe what you did, but no longer than necessary).

The content and emphasis of your final project report will be highly individualized to your project, as such it may emphasize or de-emphasize particular areas listed below, or even include areas not mentioned in this list. Reports for most projects will be expected to fully address the Problem Description, Data Description, Classifier Performance, and Conclusions. In general, most final project reports should address:

- **Problem Description**
  Provide background and context for the problem you have selected, and motivation for considering it.

  Some questions to think about are: Why is this an interesting or important problem? What is its significance? Why might others be interested in this problem?

- **Data Description**
  Provide a description of the data you used, where it came from, and why it is a good choice for the problem you have selected.

  Some questions to think about are: What data did you use? Why did you select this data? If you collected your own data, what was your data collection procedure? If you generated your own simulated data, what was your data generation procedure? If you leveraged a publicly available dataset, how was the data collected? How does this data support addressing the problem you have selected?

- **Data Pre-Processing, Feature Extraction/Generation, and Feature Selection**
  Provide a description of and motivation for any pre-processing you did, including data normalization, and/or how you generated and/or selected features. Someone else should be able to replicate your process from your description.

  Some questions to think about are: What did you do? Why did you do it? How did it improve the classification system, or benefit your study?

- **Classification Algorithm(s) Design**
  Provide a description of how you approached developing the classifier(s) you built, including any design trade-off decisions you made.

  Some questions to think about are: Why did you choose this classifier(or these classifiers)? How did you choose classifier parameters? How did you train the classifier(s)?

- **Classification Performance Results**
  Provide a quantitative description of how well the system performs (*i.e.*, include ROC curves, AUC as a function of a classifier parameter, etc.), and in the text of the report interpret the results.

  Some questions to think about are: Overall, how does the classifier perform? For what cases does the classifier perform well? For what cases does the classifier perform poorly?

- **Conclusions**
  Provide your overall assessment of the pattern classification system you built, including what you see as its strengths and weaknesses.

  Some questions to think about: How well does the system address the problem you set out to investigate? What are great things about the system? What would you do to improve the system? What would you do differently next time? What worked really well, and you would do the same way next time?

Deliverables

Proposal* ..... *Due 4:30*PM *Thursday, February 18

The project proposal should include a description of the problem you would like to address, including options for data sources, and how you plan to address the problem.

25% Status Report* ..... *Due 4:30*PM*, Thursday, March 10

The project 25% status report should include first steps toward the problem you have proposed, including preliminary data pre-processing and feature extraction/generation, or preliminary theoretical developments.

50% Status Report* ..... *Due 4:30*PM*, Thursday, April 7

The project 50% status report should include significant steps toward your proposed problem, including preliminary classifier performance results, or preliminary theoretical results.

Final Report* ..... *Due 12*PM NOON *Monday, May 2

The project final report should be a complete report, and include all the relevant elements outlined above in the project expectations.

Publicly Available Datasets

Some publicly available datasets that you may wish to leverage for your project include:

UCI Machine Learning Repository

`http://archive.ics.uci.edu/ml/`

The Center for Machine Learning and Intelligent Systems at UC Irvine maintains a repository of datasets (and data generators) that is available to the machine learning community at-large for empirical algorithm analysis. There are many, many datasets available here – 2335 datasets, 243 of which were originally intended for classification, as of January 12, 2016!

MNIST Database of Handwritten Digits

`http://yann.lecun.com/exdb/mnist/index.html`

The MNIST database of handwritten digits is a subset of the larger NIST database. The digits in the MNIST database have already been size-normalized and centered in the stored image, and so have already been formatted for pattern classification problems. (This does not mean that you should not consider further processing for feature generatation/extraction if you choose to use the MNIST database!)

Vanderbilt Biostatistics Datasets

`http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets`

The Department of Biostatistics at Vanderbilt University maintains a list of biomedical datasets covering a wide range of medical applications. (This page also contains an extensive list of links to other datasets at the bottom of the page.)

Kaggle

`http://www.kaggle.com/competitions`

Kaggle has an extensive database of real-world data from a wide variety of problem domains. Some of the data are associated with active competitions which offer cash prizes!