EE 559 **Course Project Assignment** Posted: Thur., 3/26/2015
Jenkins Due: Mon., 5/4/2015

Introduction

- Goal of project

  The goal of this project is to develop a pattern recognition system that operates on real-world data, using tools and techniques we cover in class, to learn more about how they work in a real-world setting, and to develop a system that performs well. Also, to give you an opportunity to apply what is learned in class to a problem domain that is of interest to you.

- Individual or teams

  Each student will do their own project. You may collaborate in a team of 2 students if you would like to; if you do, each of you will submit your own project report that will describe your part of the work, will also include a brief description of the collaboration and any conclusions from the collaborative effort; so please keep this in mind in designing the project and separating out the work.

- Choices and requirements

  You will have significant freedom in choosing your project topic and designing what you will do for your project. There are also a set of required aspects to the project that everyone is expected to do. These are detailed below.

Computer languages and available code

- You may use any coding language(s) you like, including Matlab, Python, C, C++, etc. You may use toolboxes such as Pattern Classification, PRTools, and LibSVM. Be sure to clearly state in your project report what languages and toolboxes you used, and what you coded up yourself. Also, please keep in mind that your project grade will depend on your understanding of the functions/routines you are using, and your interpretation of the results.

Requirements and guidelines

- Your system must perform pattern recognition; it must use a classifier that is trained on data and classifies a test set or unknowns.
- You must use techniques we have covered (or will be covering) in class. In addition, you may optionally use other techniques to compare with (that aren't covered in class).
- You must have at least two "default" systems you compare your results with, (i) and (ii) below:
  - (i) Random assignment to classes (with or without the use of priors, your choice; often it's informative to quote both).
  - (ii) Some baseline system, of your choice.

(iii) You may optionally also compare with other systems, such as state of the art published results. (You aren't expected to match or beat state of the art published results in this class project, but it can give you a good target to aim for.)

– You must use and compare at least 3 different classifiers: at least one different classifier from each category below:

(i) distribution free classification;

(ii) statistical classification;

(iii) support vector machines or neural networks. (We will be covering neural networks near the end of the semester, so if you are unfamiliar with them, SVM might be the better choice.)

– You must assess dimensionality - either using dimension reduction techniques (from the number of features given in the data set), or using a mapping to an expanded feature space, or both. In this part, you will decide on a technique for changing the number of dimensions, and perform some procedure to optimize the number of dimensions.

– You must use some procedure for model selection – for example, choosing values of parameters in a given classifier, or choosing among different classifiers.

– You must take care to use training, test, and optionally validation sets correctly.

– You must assess final performance of your best performing system, and give comparisons of your best performing system to your baseline system, random assignment, and any other system you chose for final comparison.

– You must interpret, to the extent you can, what you observe; for example, why does one classifier perform better than another (or why does it perform the same)? why does the plot of performance vs. dimension have the shape it has? if observed behavior seems unexpected, conjecture a reason (and state clearly it is a conjecture); if sufficiently interesting, consider performing an experiment to assess whether your conjecture is correct.

Special cases

– If your project topic relates to other prior or concurrent work of yours (*e.g.*, research work you have been doing, work for an internship, or other class projects), you must include a statement in your report that clearly outlines what was done in the prior or concurrent work, and what was done as new work specifically for the EE 559 class project. The rest of the report should describe the EE 559 class project work. Your grade will be based on the new work done specifically for the EE 559 class project.

Datasets [also refer to "tips", below]

– You are encouraged to choose dataset(s) that are practical for this course project, and that you would enjoy working on. See the separate handout on datasets; you may choose from the short list of datasets provided, or find other dataset(s) of interest to you. The handout also gives suggested web site resources for datasets.

Tips

- Datasets can have outputs (class labels) that are numeric or categorical. Because the topic is pattern recognition/classification, be sure your dataset has categorical outputs, or else can reasonably be configured to have categorical outputs.
- Note that datasets can include numeric and/or categorical features. Categorical features need to be treated differently than numeric. Discussion Session 10 (on Thursday, 3/26) will discuss this issue.
- Note that some datasets will have missing data. This also requires special consideration, and Discussion Session 10 will cover it as well.
- Depending on your dataset, you might benefit from performing some feature extraction (e.g., for datasets that supply data in pattern space rather than feature space). In some problem domains, this can involve a lot of time and effort; in others, it is simpler.
- You might consider normalizing your data - for example, you could evaluate performance with and without normalization or compare a few kinds of normalization. Some classifiers effectively have some auto-normalization built in, and others do not.
- If you have plenty of data, it would be wise to separate out a final test set (randomly sampled from each class), before trying any classifiers on the data, in order to ensure you will have an unused set for assessing the final performance.
- Keep in mind that computation time can be a limiting factor. You might try a few sample runs of what you have in mind to get an idea of what to expect. You can also adjust your project as needed if computation takes too long (e.g., reducing data set size, adjusting the scope of the search space in model selection, or considering the complexity of the classifiers used).
- If possible, it is be helpful to consider degrees of freedom (d.o.f.), and number of constraints, e.g. using the rule of thumb of 3 to 10 times more constraints (samples) than d.o.f. (this is easier to evaluate for some classifiers than for others; and yet for others, such as SVM, it doesn't directly apply).

Grading criteria

- Grading criteria will include: inclusion of required elements; understanding and interpretation (of approach, algorithms used, and results); technical soundness and final performance; quantity and quality of effort; and report write-up (clarity, conciseness, and completeness)

Final report

- More detailed guidelines for the written report will be posted later.