CS567 –Machine Learning (Spring 2003)

Instructor: Dr. Sethu Vijayakumar **TA:** Aaron D'Souza

FINAL PROJECT (Instructions & Guidelines)

As a part of the assigned work for this course, you are required to complete a project of your own choosing that is based on the material of this course. The premise of the project must be closely related to some aspect of the course material but may explore an avenue that was left un-addressed in class.

Credits: 40% of total grade
Due: April 29, 2003
Groups: You are required to form groups of 2/3/4 for each project. If it is a 4 people project, the effort must be commensurate to justify it.
Presentation: The last 2 classes will be dedicated to project presentations. Please submit a ***short*** project report about why you think the problem you solved (or tried to solve!!) is important and a summary of the insights you gained from it. The details of the project itself should be presented during the class presentation. The final report should outline the role of each team member in the project.

Project type and policies

There are various types of projects you can consider:

1. The project may be ***very practical*** in terms of applying techniques you have learned in the course to a real problem such as classification of email messages.
2. The project may involve ***designing or adapting existing algorithms*** to a novel class of problems. For example, how might we solve multiple related classification tasks? How can we improve document clustering by designing a new clustering metric?
3. The project may consist of a ***theoretical analysis of a method*** we have discussed. For example, this may be in terms of complexity, convergence, etc.
4. The project can be a ***theoretical or more applied survey of a branch of machine learning*** that we didn't go through in detail. For example, you may write about the use of machine learning in natural language processing, time series prediction or review sample complexity of machine learning algorithms.

The project can be related to your research area (if you have one). Do not submit anything you have completed prior to attending the course.

Project proposal

In order to judge the suitability of your choice of a project, you are strongly encouraged to submit a brief project proposal (at most one or two paragraphs) that describes your idea for the project, the work you intend to perform, and all the people involved in the project. You should submit the proposal via email to the instructor (sethu@usc.edu) anytime ***after*** March 1. Of course, you are encouraged to think about the details starting today!!!

Project size, presentation and the report

We expect that the ``size" of your project should be equal to about the amount of work required for one and a half homework assignments. The project, however, should be in some sense ``complete". By this, we mean that you should not ignore relevant machine learning issues. In the final report you shouldn't just say what you did but also why it was a reasonable thing to do given the course material. You should be able to articulate your project (justify it's importance) in front of the class during the presentation.

You shouldn't worry about getting ``great" results. The idea and your understanding of the machine learning issues involved are much more important than getting ``great" results.

Some examples

There are many avenues that you may pursue for this project and we encourage you to be creative even if you don't think you'll necessarily get ``great" results.
Here are some ideas:

1. **Comparison of algorithms**: Throughout the course, we've been discussing various algorithms and their properties, but only on occasion have we dealt with these algorithms with real sets of data. Often times, algorithms don't work like expected and algorithms may need to be adapted or modified to better fit the assumptions inherent in the data. What work needs to be done to adapt a model to an interesting set of data that you've found? How do various algorithms perform on the same set of data? What are the properties of the various algorithms that exhibit such performance?
2. **Missing information**: Various real world classification problems involve missing components in the input vectors. How can you deal with such missing information? Do you expect your method to degrade rapidly if more information is missing?
3. **Clustering metric**: How do we cluster various types of examples such as sequences? Can you devise a clustering metric or a clustering algorithm that is appropriate in such cases? What if we know that the examples can be transformed

in various ways (e.g., translation of images) without changing their ``essence''? How can we incorporate such prior knowledge into a clustering algorithm?

4. **The choice of the kernel function in SVMs**: The kernel function in SVMs defines how examples are to be compared. How do we choose the kernel function? How could we adjust the kernel function if we thought it should have a particular form? Can you adapt/design a kernel function to a specific problem we are interested in solving?

5. **Finding out the intrinsic dimensionality of natural movement data:** With the various dimensionality reduction techniques discussed in class, one can think of looking at techniques of finding the intrinsic dimensionality of natural movement data (for example). What method is most suitable for doing this and why?

Other project ideas include:

Simple language modeling using Markov models/HMMs
Improve classification by using EM with unlabeled data
Selecting the number of mixture components based on data
Clustering input attributes, designing clustering metrics
Image/email/biosequence classification
Detecting abnormal/novel examples in a stream of data
Recognition of acoustic features (towards speech recognition)
Creating backgammon/go/chess/etc. player

Important Note: Neither the instructor nor the TA will be able to constantly guide you through projects or provide additional help (from a simple time spent * number of projects point of view). The best we can do is clear some basic+short doubts.
In other words, the project is totally ***your BABY***!!!

Project Diversity

In the interest of diversity in projects, we will maintain a ***link on the course web page*** starting March 1, 2002 ***displaying a list of people*** who have already (tentatively) decided on a ***project topic***. In order that we all learn maximum from the course, we should try to avoid duplicate projects.