CS-482/CS-682 MACHINE LEARNING
FINAL PROJECT GUIDELINES

Your class project is an opportunity for you to explore an interesting machine learning problem of your choice in the context of a real-world data set. You can have three people in a project group for CS-482 and at most two per group of CS-682. Below, you will find some project ideas, but the best idea would be to combine machine learning with problems in your own area of interest.

CS-682 students are expected to work on this project alone.

Your project will be worth 20% of your final class grade and will replace your final exam. It will have 4 deliverables:

- **Week 6: Proposal and Proposal Consultation with Instructor** 1 page (12 points font) (10%)
- **Week 8: Midway Report** (3-4 pages) (12 points font) (20%)
- **Week 10 and Week 11: Project Presentation to class** (20%) **(every member must be present)**
- **Week 11: Final Report** 6-8 pages (12 point font) (50%)

It will be graded on following rubric:

1. The novelty of the project ideas and applications. The groups are encouraged to come up with original ideas and novel applications for the projects. A project with new ideas, and interesting applications of existing algorithms is scored higher than a project without much new idea/application.
2. The extensiveness of the study and experiments. A project that produces a more intelligent system by combining several ML techniques together, or a project that involves well-designed experiments and thorough analysis of the experimental results, or a project that nicely incorporates various real world applications, are scored higher.
3. Projects that have very clear presentation, visuals and graphs will be scored higher, compared those that simply print text output.
4. The project that uses multiple ML techniques for feature engineering, preprocessing, evaluation will be scored higher than those that do minimal work.
5. The writing style and the clarity of the written final report.

PROPOSAL (WEEK 6 DUE DATE)

A list of suggested projects and data sets are posted below. Read the list carefully. You are encouraged to use one of the suggested data sets, because it is known that they have been successfully used for machine learning in the past. If you prefer to use a different data set, I will consider your proposal, but you must have access to this data already, and present a clear

proposal for what you would do with it. *You must choose a dataset in which there is a test set that is NOT used for training and you must report performance metrics on this test set.*

Page limit: Proposals should be one page maximum.

Include the following information in the proposal

- Project title
- Authors Names
- Data set Description
- Project idea. This should be approximately two paragraphs. What questions are you trying to answer? Why do you believe it is important to answer this question.
- Methodology

    1. Data Preprocessing (list the techniques you will use.. see the list of techniques in the Data Preprocessing chapter)
    2. Model Development Techniques (list the techniques- parameter tuning and grid search and cross validation etc)
    3. Evaluation Metris – (list metrics for classification and regression including, precision, recall, accuracy, ROC curve, f-statistic, histogram, heatmap, RMSE, R^2 and anything else
    4. Feature Engineering Techniques
    5. List of models you will compare.

- Teammate (if any) and work division. We expect projects done in a group to be more substantial than projects done individually.
- Which member of the will be in charge of turning in documents on blackboard. Only one member of team should submit all of the documents.
- Midway Report milestone: What will you complete by Week 8?

MIDWAY REPORT (Week 8 Due Date)

The midway report should be a 4-5 pages short report in the form of a paper and it serves as a check-point. It should consist of the same sections as your final report (Abstract , Introduction, Data Description and task description, Preprocessing, Feature Engineering, Model development, Model Evaluation, and, Conclusion.), with a few sections `under construction'. Specifically, the introduction and Data Description must be in the final form; the section on the proposed method should be almost finished; the sections on the experiments and conclusions will have whatever results you have obtained, as well as `place-holders' for additional results you plan/hope to obtain.

Page limit: Midway report should be 4-5 pages (5 pages maximum).

Grading scheme for the project report will be based on:

- proposed method (should be almost finished) and experiments done so far

- the design of upcoming experiments
- plan of activities

PROJECT PRESENTATION (Week 10 Due Date)

You will make a presentation to the class. At least 20 slides are expected in the presentation. Every member must present their part. The presentation will include Introduction, Data Description, Data Preprocessing, Model training, Model Evaluation and Conclusions.

FINAL REPORT (Week 11 Due Date)

Sample final report is attached which is 8 pages long: Sections must include Abstract , Introduction, Data Description and task description, Preprocessing, Feature Engineering, Model development, Model Evaluation, and, Conclusion. There are three final reports attached to the assignment for you to review to give you an idea of how to write final report.

CS-682 REQUIREMENT:

In addition to the algorithms and techniques taught in class, the graduate student MUST include a new algorithm not covered in class. For example natural language processing, time series algorithms etc.

OR

The graduate student should come up with new theoretical enhancement to one of the algorithms taught in class. Then prove that this new theory enhances existing theory and demonstrate experiments with the same.

DATASETS

1. UC Irvine has a repository that could be MOST useful for you project. Any dataset here is well tested and there are results available in a research paper that you might guide you in the right direction. http://www.ics.uci.edu/~mlearn/MLRepository.html.

2. **30-40 Datasets for Education Data Mining**

[Dataset Webpage:] Go to the webpage https://pslcdatashop.web.cmu.edu/ (You may need log in through WebISO) and click "Public Datasets".

There are about 30-40 public datasets available from the webpage (Only select the datasets whose status is labeled "complete"). If you clicking each datasets, you will find a general description and the related publications. To look at the datasets, click "Export" link.

Project Idea 1:
For each dataset, you can compare various machine learning techniques (at least five to seven different ML methods) on predicting "Correct First Attempt values" (Generally listed in the column "Outcome"). Please report the Root Mean Squared Error (RMSE).

Project Idea 2:
Across datasets, you can compare several machine learning techniques (at least two to three different ML methods) on predicting "Correct First Attempt values"(Generally listed in the column "Outcome"). Please report the Root Mean Squared Error (RMSE) on the test data. The hypothesis here is that there may not be an absolute winner, different machine learning techniques may be effective on different task domains. For example, you can split the datasets into science (physics & math) vs. second language learning (Chinese, French).

3. **NBA statistics data**
This download contains 2004-2005 NBA and ABA stats for:

-Player regular season stats
-Player regular season career totals
-Player playoff stats
-Player playoff career totals
-Player all-star game stats
-Team regular season stats
-Complete draft history
-coaches_season.txt - nba coaching records by season
-coaches_career.txt - nba career coaching records
Currently all of the regular season

Project idea:
* outlier detection on the players; find out who are the outstanding players.
* predict the game outcome.

4. **Precipitation data**
   This dataset has includes 45 years of daily precipitation data from the Northwest of the US:

   Download Dataset

   Project ideas:
   Weather prediction: Learn a probabilistic model to predict rain levels.
   Sensor selection: Where should you place sensor to best predict rain.

5. **WebKB**
   This dataset contains webpages from 4 universities, labeled with whether they are professor, student, project, or other pages.
   Download Dataset.

   Project ideas:
   * Learning classifiers to predict the type of webpage from the text.
   * Can you improve accuracy by exploiting correlations between pages that point to each other using graphical models?

   Papers:
   * http://www-2.cs.cmu.edu/~webkb/.

6. **Email Annotation**
   The datasets provided below are sets of emails. The goal is to identify which parts of the email refer to a person name. This task is an example of the general problem area of Information Extraction.

   Download Dataset

   Project Ideas:
   * Model the task as a Sequential Labeling problem, where each email is a sequence of tokens, and each token can have either a label of "person-name" or "not-a-person-name".

   Papers: http://www.cs.cmu.edu/~einat/email.pdf

7. **Netflix Prize Dataset**
   The Netflix Prize data set gives 100 million records of the form "user X rated movie Y a 4.0 on 2/12/05". The data is available here: Netflix Prize.

   Project idea:
   Can you predict the rating a user will give on a movie from the movies that user has rated in the past, as well as the ratings similar users have given similar movies?

Can you discover clusters of similar movies or users?
Can you predict which users rated which movies in 2006? In other words, your task is to predict the probability that each pair was rated in 2006. Note that the actual rating is irrelevant, and we just want whether the movie was rated by that user sometime in 2006. The date in 2006 when the rating was given is also irrelevant. The test data can be found at this website.

8. **Physiological Data Modeling (body media)**
Physiological data offers many challenges to the machine learning community including dealing with large amounts of data, sequential data, issues of sensor fusion, and a rich domain complete with noise, hidden variables, and significant effects of context.

1. Which sensors correspond to each column?
characteristic1 age
characteristic2 handedness
sensor1 gsr_low_average
sensor2 heat_flux_high_average
sensor3 near_body_temp_average
sensor4 pedometer
sensor5 skin_temp_average
sensor6 longitudinal_accelerometer_SAD
sensor7 longitudinal_accelerometer_average
sensor8 transverse_accelerometer_SAD
sensor9 transverse_accelerometer_average

2. What are the activities behind each annotation?
The annotations for the contest were:
5102 = sleep
3104 = watching TV

Datasets can be downloaded from here.

Project idea:
* behavior classification; to classify the person based on the sensor measurements.

9. **Enron E-mail Dataset**
The Enron E-mail data set contains about 500,000 e-mails from about 150 users.

The data set is available here

Project ideas:
* Can you classify the text of an e-mail message to decide who sent it?

10. **More data**

    Sam Roweis also has a link to several datasets out there: http://www.cs.toronto.edu/~roweis/data.html.

    Dr. Jan Wiebe MPQA opinion annotated corpus: http://www.cs.pitt.edu/mpqa/