Term Project guidelines

Each project team will be of size from **three to five** students (assigned randomly by instructor), depending on the size of the class. The goal will be to analyze and develop a machine learning solution to a prediction problem of your choosing.

Grading: This project is a critical part of the course, and a significant factor in determining your grade. By default, all team members will receive the same score for their project. To ensure that this is fair, each member needs to submit a "peer assessment" form along with the final report. If these forms suggest HIGHLY imbalanced contributions, then I will need to look at the issue in more detail and most likely have a meeting with all the members together to mediate and come up with a fair score distribution.

Dates:

1. **Project outline due Oct 24**. 2-3 pages describing the problem, data available, some possible approaches you will consider to address the problem, and a short list of references. (need not be fully flushed out, more for a sanity check).
2. **In-class presentation** of project results, Late Nov/early Dec, approx 15-20 mins per group.
3. **Written Report due via Canvas by midnight, Dec 8th**, via Canvas. One submission per group. You are also asked to submit supplementary materials (code, referenced papers) via a pointer to the appropriate URL/dropbox/github/.. location(s).

Project presentation schedule

Project groups, title and schedule will be available on Canvas when ready. Guidelines for your in-class presentation and for the content of the report and the criteria for its evaluation are uploaded into Modules à Projects

Project topics

The project should be centered around some problem with associated data sets that you can mine to provide useful and actionable answers. At the least, this should be an exercise in analyzing a reasonably large dataset. In the process, if you invent new techniques/algorithms or processes, or make inferences that are useful and not done before, of course that is an added bonus, though this is not common. Two types of projects are suggested below.

Type I: Based on a Competition or other Real-World Large Datasets

Data Mining Competitions

There have been several data mining competitions such those hosted by Kaggle (www.kaggle.com). For several of these competitions, as well as those from KDD cup (http://www.kdnuggets.com/competitions/kddcup), the data is still available and you can also find papers on how others have fared on these data sets. There are also several other ongoing competitions (e.g. see http://www.kdnuggets.com/competitions).

Warning: these can be quite addictive, but also quite fun and a learning experience, specially if it is an on-going competition.

Other Public Domain Datasets

There is an astonishing amount and variety of public domain datasets on the web. Google now provides *https://toolbox.google.com/datasetsearch* to help looking for datasets.

You could be even selective on the topic, for example, if you google “multilabel classification dataset”, the first hit is a bunch of datasets associated with the software Mulan.

Microsoft has made available a variety of datasets at
http://research.microsoft.com/en-us/projects/data-science-initiative/default.aspx
They periodically hold competitions as well.

The US Government's Open Data policy has also resulted in a treasure trove of data. See (http://www.data.gov)

Type II: Based on Type of Analysis or Application Domain

You can formulate and address a suitable predictive modeling problem based on data from industry or government. It will be your job to acquire and manage the data. **The project should be doable within a couple of months, but also non-trivial: at the very least it should involve a large (say "rows" times "columns" > 1 million) data set.** Remember that your class presentation is public, however your class report is not, and I (and the TA) can sign NDAs if need be in order to work with you on such a project and to evaluate it. You can choose any topic you want. For example, you could look at healthcare data, or data related to recommendation systems. Some pointers to these two example topics are given below: