CS 584 – Data Mining

Project Guidelines[1]

For the class project, follow these procedures:

- Form a team of two students.
- Decide on a problem that your team would like to work on. Provide background and motivation in your proposal (and subsequently, project report). Explain why the problem is interesting.
- *All proposals need to be approved.* Your grade will depend on several factors, including challenge level of the problem, novelty, demonstration of knowledge in course materials, evidence of research effort and depth of thinking, experimental design and evaluation, analysis of results, the report itself and the presentation.
    - To help narrow down the topic that you'd like to work on, you can think of the project in terms of two focuses: application and/or methodology. For application, the focus is on solving a real-world problem (see Kaggle competitions, for example), but there should be some challenges that you try to address. For methodology, the focus is on the development of novel techniques that improve upon existing methods. Of course, your project may very well consist of both components.
- See a list of datasets in the bottom of this document. Feel free to find your own dataset.
- Write a 1-page project proposal. Your project proposal should be structured into the following sections (it should concisely answer the following questions):
    - **What is the problem your team is solving?** Give a brief but precise description or definition of the problem.
    - **What data will you use?** Briefly describe the data, the sizes (number of records, file size) and where will you get the data.
    - **How will you solve the problem?** Describe your approach: what method, algorithm, or technique do you plan to develop or use? *Be as specific as you can!*
    - **How will you evaluate your method?** Describe how you will measure performance or success of your method. Against what baseline methods will you compare your algorithm or how do you plan to obtain ground-truth labeled data so that you can then measure accuracy, precision, recall or some other metric that will tell me how well is your method really performing.
- Write a project report that is well-formatted, using the ACM template: http://www.acm.org/sigs/publications/proceedings-templates. The report should be at least 5 pages long. Describe the problem, your approach, the results, and the related work.

[1] Adapted from CS341 at Stanford, Project in Mining Massive Datasets

The report should have the following sections:

- **Abstract**: Summary of the report.
- **Introduction:** Talk about motivation of the problem; provide a description or definition of the problem or hypothesis you set to evaluate.
- **Related work:** How does this problem and the method relate to problems/methods others have developed in the past.
- **Solution:** How did you solve the problem? Describe the technical approach. Tell us what method/algorithm did you use, develop or extend and how did you implement it.
- **Experiments:**
    - **Data:** Briefly describe the data and its size (number of records, file size)
    - **Experimental setup:** Describe how did you setup your experiments, how the training/testing data was prepared, what performance metrics are you considering, what baseline methods for comparison are you using.
    - **Experimental results:** Describe your experimental results. Structure your experiments around particular aspects of your method. For example, you could structure the experiments as follows: (1) a table showing results of your method using different types of features; (2) table comparing the performance of your method to the baselines; (3) a graph plotting the size of the training dataset vs. the time it takes to train the model; (4) Investigation of the learned model (what are the important features, etc.).
- **Brief conclusion**
- **At the end of the paper, also describe the contribution of each team member. Does each member contribute equally? You can email me privately if you don't feel comfortable putting this in the report.**

- The project will be 30% of your overall grade. Here are the percentage breakdowns:
    - Proposal: 5%
    - Presentation: 5%
    - Report (including code): 20%

Resources (software/datasets/ideas):

- Kaggle Competitions: http://www.kaggle.com/competitions
- http://www.stanford.edu/class/cs341/data.html
- http://www-users.cs.umn.edu/~kumar/dmbook/resources.htm
- UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/

Important Dates:

Proposal due (revised): 3/29

Presentation: TBA
Project report due: TBA