CS 484 – Data Mining

Project Guidelines[1]

For the class project, follow these procedures:

- Form a team of two students.
- Choose from one of the 4 tracks.
  1. Application: Your focus is on application and data. Find or create a large dataset and do something interesting with the data.
  2. Survey: Focus on a specific task (e.g. anomaly detection) and compare at least 3 existing techniques. Use at least 5 datasets.
  3. Methodology: Develop a new technique, or improve an existing technique. Compare with at least 2 existing techniques. Use at least 3 datasets.
  4. Other: If there is something you'd like to work on that doesn't belong in any of the tracks above, come talk to me.

  Whichever option you choose, I expect you to implement at least one algorithm. *All proposals need to be approved.* Your grade will depend on several factors, including the level of difficulty of the implemented algorithm(s), so don't just pick the easiest one to implement. Other factors include experimental evaluation and analysis.
- See a list of datasets in the bottom of this document. Feel free to find/create your own dataset, but the data should not be from the UCI Machine Learning Archive if you choose the application track.
- 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 3-5 pages project report, describing the approach, the results, and the related work. Use the ACM template: http://www.acm.org/sigs/publications/proceedings-templates

[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.**

- The project will be 20% of your overall grade. Here are the percentage breakdowns:
    - Proposal: 5%
    - Presentation: 3%
    - Report (including code): 12%

Resources (software/datasets/ideas):

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

Important Dates:

Proposal due: 10/27
Presentation: TBA
Project report due: TBA