Data Mining (CS 484) – Project guidelines

The project gives you an opportunity to explore in depth a particular topic/area of the course that interests you. The topic of the project, of course, should be related to the material covered in class, but otherwise you are free to select the specific topic. Possible types of projects include:

1. **An application research project**. The project demonstrates the application of some techniques discussed in class in an application domain (e.g., text mining, bioinformatics, computer vision, image processing, artificial intelligence etc.). Properties, drawbacks, advantages of the used techniques are analyzed within the context of the explored application domain.
2. **A theoretical or methodological research project**. A study of different classes of models and approaches; proving experimentally (and possibly theoretically) properties of known algorithms; designing a new approach.

Follow these steps (deadlines are on the class webpage):

- Form a team of two students.
- Identify a problem and a dataset (or more than one) to work on. The dataset(s) should have a reasonable size. Datasets with 100 instances and 10 dimensions are not interesting for this class.
- Write a 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, its size (number of instances, number of features) 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 8-10 page project report, describing the approach, the results, and the related work. The overall form of the paper depends on the nature of the project. The report should address:
  - **Problem description**: Give a brief but precise 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 and number of features, type of features etc.)
  * **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**.

To give you some idea what other students have done in the past, here are some topics and datasets they collected/used. Note: Some datasets take time to collect, like twitter data, so if you're interested in those kinds of data, make sure you start early.

- A music recommender system using the KDD-Cup 2011 data, available at http://webscope.sandbox.yahoo.com/catalog.php?datatype=c
- Clustering of images into advertisement or not. Data available at: http://archive.ics.uci.edu/ml/datasets/Internet+Advertisements
- An association mining approach for detecting malicious behavior in software systems
- Automatic detection of North Atlantic Right whales. Data available at: http://www.kaggle.com/c/whale-detection-challenge
- Automated clustering for web results.