15CSE481 Machine Learning and Data Mining Lab

Case Study Guidelines

1. Team Selection: You are encouraged to work in groups of four for the term project. This project is a critical part of the course, and a significant factor in determining your grade. Please feel free to discuss your project with faculty in-charge before proposal submission. Also, if you have difficulty in finding a project partner, you could contact faculty in-charge with a list of your interest area(s), so I can try to match you up with another student

2. Project Types: The project can be a practical one (a specific application of data mining), or a theoretical one (propose a new algorithm etc). In-depth survey papers on new or focussed topics are also possible, but it should be an original and unique niche (several broad surveys exist on the web; a cut-and-paste job will not do!). You can choose any topic relating to data mining and use any reasonably large data set. Suggested Areas
   (It is only a suggestion, but the topics are need not be limited to this)
   a. Clustering:
      - clustering with constraints; clustering very high dimensional data; clustering of gene sequences; etc,
      - clustering tools: (a) CLUTO (Karypis): understand this tool; add to its functionality, or (b) Co-clustering: understand this approach; add to its functionality.
   b. Classification:
      - comparing different approaches to solving multi-class problems
      - bagging with strong learners vs. boosting with weak learners
      - dealing with highly different costs; priors
   c. Bioinformatics: See KDD challenge competition
   d. Privacy Preserving Data Mining or Distributed Data Mining: How to do data mining (regression, classification, AR, clustering) over multiple, geographically distributed and possibly quite varying, data sets.
   e. Remote Sensing:
      - predicting forest cover type from satellite images (see UCI).
      - classifying land cover from hyper-spectral data.
      - discovering and modelling heterogenous regions in spatial data
   f. Web Mining: There is a fairly extensive list of projects provided on web mining (along with references, etc) – see KDD Cup 2003, 2005,etc.,
   g. Streaming Data; Change Detection : Analyzing data that you see only once (Johannes Gehrke); Detecting and modeling ”change” in time sequence data, e.g. those gathered from networked computer systems; change in customer segmentation (how do clusters move, get created/die, as the stats of data gathered changes over time?).
   h. Ambient Intelligence: Human Activity Recognition-- Pattern discovery in a smart environment; Ambient Assisted Living.

3. Case Study Rubrics:

It is mandatory that students should exhibit algorithms/techniques beyond the class content to get a score above 50%

Steps	Guidelines	Marks
1. **Frame the problem** 2. **Get the data**	- The quantity & quality of your data dictate how accurate our model is - The outcome of this step is generally a representation of data (Guo simplifies to specifying a table) which we will use for training - Using pre-collected data, by way of datasets from Kaggle, UCI, etc., still fits into this step	5
4. **Explore the data :** EDA and visualization with hypotheses 5. **Prepare the data :** Data transformation	- Wrangle data and prepare it for training - Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.) - Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data - Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis - Split into training and evaluation sets	10
6. **Model the data** 7. **Fine-tune the models**	- Choose the model - Different algorithms are for different tasks; choose the right one - Train the Model ▪ The goal of training is to answer a question or make a prediction correctly as often as possible ▪ Linear regression example: algorithm would need to learn values for *m* (or *W*) and *b* (*x* is input, *y* is output) ▪ Each iteration of process is a training step - Parameter Tuning ▪ This step refers to *hyperparameter* tuning, which is an "artform" as opposed to a science ▪ Tune model parameters for improved performance ▪ Simple model hyperparameters may include: number of training steps, learning rate, initialization values and distribution, etc	15
8. Present the solution	- Evaluate the Model ▪ Uses some metric or combination of metrics to "measure" objective performance of model ▪ Test the model against previously unseen data ▪ This unseen data is meant to be somewhat representative of model performance in the real world, but still helps tune the model (as opposed to test data, which does not)	10

	▪ Good train/eval split? 80/20, 70/30, or similar, depending on domain, data availability, dataset particulars, etc. - Make Predictions ▪ Using further (test set) data which have, until this point, been withheld from the model (and for which class labels are known), are used to test the model; a better approximation of how the model will perform in the real world	

4. Milestone 1: Submission of Project Proposal --
   Due: End of 3rd Week of the course
   First, each group should submit a one/two page proposal summarizing the proposed project including the plan of attack and at least two key references. Your project proposal must be typed and should be approximately 1 page long, single spaced. The purpose of the proposal is to make sure that you are on the right track and to give me enough information so that I can give you useful feedback. In your proposal you should cover the following items:
   - Preliminary title and list of students working on the project (1 or 2 students per project without prior permission from me; 3 or more is possible for very ambitious projects).
   - Abstract: Similar to the abstract that will ultimately appear in your paper. It should be one paragraph long, for now perhaps only 5-15 lines. It should provide a high level summary of your project and outline your main goals.
   - Brief description of what you plan to do.
     - What problem are you trying to solve?
     - How do you formulate the problem as a data mining problem (e.g., is it classification, association rule mining, etc.)? What exactly are you trying to predict (for prediction tasks) and how will you evaluate your results. How will you know if your results are good? What can you compare them to? It is critical that your problem is well-defined.
     - What data sets do you plan to use? If you must do significant work to get the data or convert it into the proper format, then describe the process and approximate effort required. How many examples are in the data set? How many features?
     - What learning tools do you plan to use (e.g., WEKA. Python Scikit) and what algorithms do you plan to use (e.g., decision trees, neural networks, etc.)?
     - If sufficiently valuable, list a few related research papers.
5. Milestone-2: Organization of final case study presentation
   Due : last two labs of the course
   Submission: Paper/Report and Presentation

   1. Introduction

2. Objective
3. Problem statement and Assumption (if any)
4. Architecture
5. Related work and Novelty of the work
6. Data collection and preparation
    a. Feature Selection and/or Feature Extraction
    b. Data Visualization and Hypothesis
    c. Data Cleaning & Pre-Processing

        (Mention why the pre-processing was required)
7. Data Modeling and Inference

        (Mention why the predictive model was selected)
    a. Improvements/Enhancement of models (mandatory)
    b. Performance Evaluation and result discussion
    c. Model Tweaking, Regularization, Hyper Parameter Tuning
8. Conclusion and Future Enhancements
9. References