COSC 174 Project Proposal – Deadline Prediction

Josh Cook, Adi Mahara, Byoungwook Jang

1. What problem do you want to solve? Describe at a high level the application and the learning problem involved.

The problem we are attempting to solve is to develop an algorithm that predicts the number of deadlines students have on a current day. The output of our algorithm will categorize students as having 0, 1, 2, or 3 or more deadlines on a given day. This will be based on the trends of their activities, moods/mental health, gym usage and other factors recorded from days leading up to that point. The data will be used from the Dartmouth Student Life Application running on the students' issued cell phones. These sorts of predictions could be extremely useful for both students and faculty for improving life on campus.

2. What are suitable methods for this problem? Research the machine learning literature and find a set of techniques that may be applied to solve this task. Include references to these methods in your proposal.

After the first two weeks of class and looking over the data set, we were able to deduce that this a supervised, multi-class classification machine learning problem. We looked into the survey paper, to briefly learn about the most prevalent algorithms to learn from data [1]. More specifically, we are going to explore algorithms such as Naive Bayes Classifier [2-3], Support Vector Machine [4], and other classification algorithms [5].

3. What data sets do you plan to use? Include pointers to training data that you will use in your project.

For data set, we propose to use (44 students * 71 days) = 3124 examples for which we will explore feature sets that will include subjects passive and automated sensing data from their phones. These example sets will be approximate divided into training and testing sets. The information will include information regarding student's mental health (eg: depression, loneliness, stress), behavioral trends (gym time, sleep time, dining habits), and sensor data (time the phone was unlocked, conversation time, activity, audio, app use).

4. What do you expect to accomplish by the milestone due date?

By the milestone we expect to:

- Have matrices of feature set completed to be implemented in algorithms such as Naive Bayes and Support Vector Machines
- Identify the features that have the most influence on the classification
- Visualize the identified feature dataset with regards to the deadlines data
- Explore which algorithm works best for our dataset and get a preliminary understanding of what to use

References:

1. X. Wu et. al, "Top 10 algorithms in data mining", Knowl Inf Syst (2008) 14:1–37.
2. David D. Lewis et. al, "Naive (Bayes) at forty: The independence assumption in information retrieval", Machine Learning: ECML-98 Lecture Notes in Computer Science Volume 1398, 1998, pp 4-15.
3. Andrew McCallum el. al, "A Comparison of Event Models for Naive Bayes Text Classification",AAAI-98 workshop on learning for text, 1998.
4. Arun Kumar et al, "Least squares twin support vector machines for pattern classification",
5. Richard Duda et. al, "Pattern Classification" , Second Edition, A Wiley- Interscience Publication, 2001.