COURSE MEDIAN PREDICTION VIA SYLLABI ANALYSIS

CORALIE PHANORD, GRAESON McMAHON, KELSEY JUSTIS

January 22, 2015

1 Problem Statement

College students often find median grades helpful in the course selection process. Knowledge of a course's median grade provides insight on the difficulty of their courses enabling an individual to set a well-balanced schedule. In this project, supervised machine-learning algorithms are used to predict median grades given course syllabi, or more specifically the identifiable features present in the syllabi.

2 Method

We are considering two regression-based approaches for our problem. Currently, we are leaning towards decision tree algorithms such as Ross Quinlan's C4.5[1], as they excel at handling the mixture of numerical and categorical data which characterizes our feature set. We are also considering artificial neural networks, in part for their 'tolerance to noisy data.'[2]

3 Data

The data used provides training examples in the form of syllabus-median grade pairs; the median grades serve as points to be fitted by algorithms learning features that are present in the corresponding syllabi. Important characteristics of the data are detailed below:

3.1 Sources

Data publicly available via department/college websites for Dartmouth and other colleges and from the Registrar's Office/department databases (requests currently pending).[3][4]

3.2 Training set

Syllabus-median grade pairs from random (partitioning process to be determined) quarters.

3.3 Testing set

Syllabus-median grade pairs from random (partitioning process to be determined) quarters. Although not present before the project due date, we hope to also test on the present term's data.

[1] C4.5 reference: http://en.wikipedia.org/wiki/C4.5_algorithm
[2] Neural network reference: http://www.solver.com/xlminer/help/neural-networks-prediction
[3] Median grades: http://www.dartmouth.edu/ reg/transcript/medians/
[4] Sample syllabi source: https://biology.dartmouth.edu/undergraduate/courses-and-syllabi

3.4 Number of examples

An exact figure is currently being determined; modest estimates (without approved department requests) suggest 500+ examples are obtainable via Dartmouth websites alone.

3.5 Features

The text each syllabus contains offers many data points, because of this, we will test and discover how document formatting/aesthetics, language rhetoric, word connotation, and word frequency factor into a course's median grade. Tentative features include:

- Syllabus length
- Proportion of bolded, italicized, and underlined words
- Course department
- Course number
- Occurrences of assignment words such as 'quiz,' 'exam,' 'midterm,' 'test,' and 'project'
- Presence of labs
- Occurrence of terms suggesting stringency such as: 'late,' 'penalized,' 'mandatory,' and 'prerequisite'
- Frequency of negative words such as: 'no,' 'not,' 'never,' etc.
- Use of x-hours
- Use of teaching assistants
- Time-slot
- Number of '%' symbols

4 Milestone Goal & Project Timeline

4.1 Collect and format data for processing

Accomplished by January 29

Data source and format choice are finalized; data is ready to process.

4.2 Flexible parser to extract features from syllabi

Accomplished by February 12

Code is developed and capable of scanning syllabi for features in raw text and document formatting.

4.3 Initial development of chosen algorithm

Started by February 17 (Milestone due date)

Algorithm used to learn features from syllabi via parser is chosen. Early work is started with emphasis on code structure; outside implementations of chosen algorithm are examined for best practices.