CS 324: Data Mining **Dave Musicant**
Final Project **Winter 2015**

Description

The class project is due at the end of finals. You can choose to do essentially anything inside the area of data mining, so long as I approve it in advance. You can build on material we do in class, or you can use it as an opportunity to learn about subject matter we don't have time to cover. That said, here are two specific categories of projects of which you might consider one:

Category 1: Study and implement an algorithm or approach

Perhaps you'd like to find an algorithm described in one of our textbooks, some other book, or in a research paper that we haven't talked about, and implement it. This might be for a data mining problem we've already considered, or it might be a completely different kind of problem. If you go this route, you should ultimately turn in a short description of what you implemented, citation(s) for where you learned about your approach, and your working software with sample data to run it on.

Category 2: Use existing data mining software to study a dataset

Perhaps you can find a dataset of interest, and use either your own code or R, Weka, or some other tool to heavily analyze it and learn what you can. There are a number of data mining contests out there: perhaps you want to pick a current one and try it. If you go the route of analyzing a dataset, you should turn in a paper describing what you've learned: tell me what patterns you've found in the dataset, and interpret the results (tie it back to reality). If you go this approach, do not underestimate the time it may take to learn how to use the software that you intend to use.

Here are some potential sources of data:

1. UCI KDD Archive
2. UCI Machine Learning Repository
3. Kaggle. Kaggle is often running a variety of contests you may be interested in trying.

Most online data comes with some kind of license. Make sure that you can appropriately use the data as you intend to.

Proposal

Submit a one page proposal, including the names of both people involved (if it is a team project), describing what it is you want to do for your project.

Deadline

Your deadline is the end of the last final exam. That will be on Monday, March 16, at 9:30 pm. This is not a midnight deadline: the end of the last exam is 9:30 pm. If you miss this deadline, I am by forbidden by college policy to extended it, and I cannot grade it. You will need to contact your class dean for permission to have late work graded.

Submitting your data

If you are using a large dataset, you shouldn't submit it via Courses. There are submission limits, and the disk capacity on Courses is likely limited. If your dataset is large, make sure to work with me before the deadline so you can get me the data via some other means.