CS4786/5786: Machine Learning for Data Science, Spring 2015
3/17/2015: Competition 1: Clustering Challenge (Rough draft)

Prefatory note Its on! The competition has officially begun.

Instructions The due date is April 22nd, 11:59PM on CMS. Submit what you have at least once by an hour before that deadline, even if you haven't quite added all the finishing touches — CMS allows resubmissions up to, but not after, the deadline. If there is an emergency such that you need an extension, contact the professors.

You may work in groups of one up to four[1]. Each group of two or more people must create a group on CMS well before the deadline (there is both an invitation step and an accept process; make sure both sides of the handshake occur), and submits 1 submission per group. **We will not be automatically transferring the groups from previous assignments, this means you should form groups anew for A2.** You may choose different groups for different assignments. Please ensure that each member of the group can individually defend or explain your group's submission equally well.

You will submit both a writeup and two datafiles you create. The writeup has to be longer than 5 pages and shorter than 15 pages. **The first challenge is also posted as a Keggle competition and you guys can compete against each other!**

Keep an eye on the course webpage for any announcements or updates.

Academic integrity policy We distinguish between "merely" violating the rules for a given assignment and violating *academic integrity*. To violate the latter is to commit *fraud* by claiming credit for someone else's work. For this assignment, an example of the former would be getting an answer from person X who is not in your CMS-declared group but stating in your homework that X was the source of that particular answer. You would cross the line into fraud if you did not mention X. The worst-case outcome for the former is a grade penalty; the worst-case scenario in the latter is academic-integrity hearing procedures.

The way to avoid violating academic integrity is to always document any portions of work you submit that are due to or influenced by other sources, even if those sources weren't permitted by the rules.[2]

[1]The choice of the number "four" is intended to reflect the idea of allowing collaboration, but requiring that all group members be able to fit "all together at the whiteboard", and thus all be participating equally at all times. (Admittedly, it will be a tight squeeze around a laptop, but please try.)

[2]We make an exception for sources that can be taken for granted in the instructional setting, namely, the course materials. To minimize documentation effort, we also do not expect you to credit the course staff for ideas you get from them, although it's nice to do so anyway.

*Q1 (***Challenge 1: Two clusters, "For" and "Against").** The first challenge is to cluster the 2740 speeches into two groups. First group consisting of "For" speeches and the second group consisting of the "Against" group. To start you off, you are told that points (rows) $\mathbf{3}, \mathbf{14}, \mathbf{19}, \mathbf{25}$ are examples of speeches that belong to category "For" and speeches $\mathbf{1}, \mathbf{4}, \mathbf{28}, \mathbf{178}$ are examples of speeches that belong to the "against" category. We shall represent "For" group with label 0 and "Against" group with label "1". Your goal is to produce a file consisting of 2740 lines, each line having either a "0" or "1" indicating for or against for each of the 2740 data points. You will name this file "votes.csv". We will evaluate your grouping against the ground truth (which is of course hidden from you). Your goal is to make as few mistakes as possible.

*Q2 (***Challenge 2: 38 clusters, one for each debate).** Here your goal is to cluster the points into 38 groups where each group is meant to represent speeches made in a single debate. The 2740 speeches were made on 38 debates. Your goal is to produce a file consisting of 2740 lines, each line having one of "1" to "38" indicating the debate number of each of the 2740 speeches belong to. You will name this file "debates.csv".

Deliverables and instructions : Part of the competition is on Keggle so do sign up and compete with each other. At the end of the competition you will be required to submit two data files described in the two challenges above. But more importantly you need to also submit a writeup/report ("writeup.pdf") of things you tired why you tried them (irrespective of whether they worked or failed). This report has to be at least 5 pages long and not more than 15 pages. The write up will count to at least as much of the grade as the empirical results of the final cluster assignments you submit. Here are a few other pointers:

1. Include visualizations of both successful and unsuccessful trials.
2. Make a note of all successful and unsuccessful methods you tried. Explain why you made the choices you made and why you expected them to work both for successful and not so successful choices and take a shot at explaining why the less successful ones were in fact not so successful.
3. Organize in sections each section heading representative of the methods tried.
4. You are certainly encouraged to try methods you might have picked up outside those covered in class and maybe even extensions you develop on your own for the problem!. If you use methods other than ones covered in class, **do compare the performance both empirically and conceptually with (reasonable choice of) methods covered in class**.
5. We will definitely reward karma points for work that goes above and beyond.
6. **If you turn in a draft in first week, we can provide feedback!**
7. **The first challenge is also set up on Keggle as a competition**.
8. **Please check for new/edited drafts of this instruction document it will keep getting updated**.