Michael Mason
ECE/ME 532 Course Project Update #1
Due: 11/17/2020

Progress (all codes have git links provided below)

Pre-processing: My work began with creation of a .mat file from the raw csv data that can be easily read in my python scripts (via a simple MATLAB script I wrote) in the form of a **X feature matrix with y vector of resulting page shares**. As it is not pre-processed initially, my python scripts normalize the raw columns of X so that all columns have the same 2-norm. I did this on the python-side to keep all data-processing self-contained within the python language. Early work with the dataset showed that a set of features pertaining to keywords in the articles is poor as the values appear to be dependent on time of publication (which is otherwise *not* a dependency), with earlier articles having values of zero and the values steadily increasing in the dataset, despite this not making intuitive sense, hence this subset of the characteristics is first removed, **leaving 48 features for examination**.

Least squares: I have implemented and run cross-validation using multiple **least squares** approaches: **pure least squares, least squares with polynomial values of the features added in new columns, ridge-regression, and LASSO**. **I did not have anything beyond pure least squares in my proposal, but wanted to look at more techniques learned in class after the writing on the initial proposal.** For the polynomial addition cases, resulting singular matrices made least squares solutions difficult, and as this was a side step in the project and wasn't in the proposal, I decided not to progress further. Similarly, with ridge regression, after using the optimization setup learned in class (for reducing the inverted matrix sizes), memory limits still were hit, and again as this wasn't in my proposal, I moved on. **For my successful implementations of pure least squares and LASSO, I ran cross-validation across 11 subsets of the data, with the training X matrix containing 9 of the sets, lambda for LASSO determined from one of the hold-out sets, and both methods then tested on the second holdout set**. This was repeated for all combinations. **In addition, this was run over both the full set of features as well as 11 subsets of X breaking the set of 48 features into smaller, directly-connected groupings** such as title features, article contents, subjectivity measures, and others. The least-squares code thus runs over 12 X matrix sets and subsets for 110 cross validations each, completing **1320 total test cycles in approximately 2 hours**.

LS Results: The results from this subset have shown poor performance for both pure least squares and LASSO, with LASSO performing much better than pure least squares. **The error metric used is = sum($|y_{actual}-y_{computed}|/y_{actual}$).** Overall, some of the subset cases have performed best, but none of the methods or subsets performed better than 75% error.

Initial k-means: I am currently writing the code for processing k-means clustering to determine optimal low-rank approximation of X and the subsets, to project estimates of y for new data. I will be using an iteration approach similar to finding the optimal lambda in our LASSO and ridge regression work to find the optimal value for k. This is planned to be finished by November 23.

Next-steps

In my proposal, I had planned on also having k-means finished by this point, but with the additional methods for least squares attempted/employed, I now plan to have it finished by the end of next weekend (Nov 23). As we're learning about neural networks the following week per the schedule, I will implement that part of the project after the k-means work, and will be on track to finish the project otherwise on the original schedule. The new schedule (from Nov 17 onwards) is provided below.

Additional work from proposal, and clarification from comments on Project Proposal

As mentioned in my progress section, I employed multiple least-squares based approaches beyond a basic least squares implementation as learning about the other methods, in particular sparse solutions, led me to want to test more of them. **I also have written code for (included in the script) but not yet implemented an alternate method where large and small outliers are removed from the testing data, as I suspect some of the negative performance is due to a small number of articles having ~10 times more shares than most (presumably articles that went viral).**

From the project proposal instructor comments and per a brief meeting in office hours, I want to add that the k-means work will have an optimized k-value determined from a holdout test set similar to the lambda-optimization we have used for LASSO and ridge regression. Additionally, the goal of this project is to predict as well as possible the number of shares an article will generate, and so the problem being considered is regression.

Git

For my first work, I was not uploading/modifying my git resources as I went along, but will be doing so for all future work with this project. I apologize for not keeping it up during the first work period, I am still new to the git workflow.
Pre-configuring in MATLAB:
https://github.com/mamason2/ECE532Project_Mason/blob/master/PreProcRawData.m

Least-squares cross-validation (complete):
https://github.com/mamason2/ECE532Project_Mason/blob/master/ProjectStartUp1_LeastSquares.ipynb
(PDF of run):
https://github.com/mamason2/ECE532Project_Mason/blob/master/ProjectStartUp1_LeastSquares.pdf
Work so far on k-means:
https://github.com/mamason2/ECE532Project_Mason/blob/master/ProjectStartUp2_KMeans.ipynb

Timeline (updated)
November 17: Implementation and validation of least squares (pure and LASSO, as well as early attempts with other approaches) complete, initial work for k-means approach set up, ***update due***
November 23: Implementation and validation of k-means complete.
November 27: Neural networking complete.
December 1: Neural networking complete, validation under way, ***update due***
December 7: Neural networking validation complete, analysis of results complete, report written (*complete early to potentially allow all-encompassing method/model to be attempted, as well as a buffer for unforeseen delays in project*)
December 12: ***Project submitted***