Samuel Valentim-Gervais October 20th, 2019

Deliverable 2

**Problem Statement**

Given a dataset of blog posts (n = approx. 600K), labeled with the age and gender of the poster, I’m trying to generate two models to predict age and gender based on text.

**The Dataset**

Still the same as originally, haven’t changed anything about that. As for preprocessing, I lowercased all the blog posts and deleted anything that wasn’t a letter of the alphabet or a space, and removed any instances of stopwords that I found on this list: <https://www.ranks.nl/stopwords>, which I loaded into a text file. I also applied a stemming algorithm from the nltk library, which reduces words into their root. For example, the word “going” becomes “go”, the word “apples” becomes “apple”. By itself, that got rid of about 10K useless features. Then, I delete words that are present in less than 0.05% of blog posts. Those words are probably typos, and are present in a small enough quantity that they shouldn’t affect the model too much. Eliminating them brings me down to about 10K unique words. I use the bag-of-words representation to vectorize my data like in assignment 2, but instead of using a numpy matrix, I use a sparse matrix from scipy which stores zeros implicitly and speeds up matrix computations involving a lot of zero values immensely, which is perfectly suited to my data! This speedup allows me to apply dimension reduction (SVD in this case) and bring my dimensions into the hundreds.

**The Model**

Originally, I was going to use Naïve Bayes, as I had a lot of features, but since I managed to apply dimension reduction, the sky is the limit !For gender classification, I decided to try k-NN (k=10), since I have a crapton of data and way fewer dimensions to fit them all. I’m really not sure how many dimensions is optimal, and training it took a while, so finding the optimal dimensionality can’t really be done programmatically. I’m also considering trying pretty much everything and seeing what sticks, but I haven’t had the time to try much else yet; processing 500K datapoints involves a LOT of waiting. I haven’t even touched the age regression yet.

**Results**

At 100 dimensions, Naïve Bayes gives a rather disappointing 58% training and validation accuracy. (Important to note is that male-female ratio in the dataset is 50-50) 10-NN fares slightly better, with 66% accuracy, but the validation accuracy is at about 60%, so it’s starting to trail off. I don’t have any other results at this point.

**Next steps**

I’m definitely going to want to play with hyperparameters a bit, especially the number of dimensions on the SVD. I’m also going to try different models and hopefully hit at least 70% validation accuracy (Random forest? Something else? I really don’t know.) As for the age, some form gradient descent might be good. Don’t know about polynomial regression, since I have binary data (by that I mean my X matrix is all 0s and 1s). Would it still work as well? I’ll try L1 regularization, and try to eliminate features that aren’t really predictive. Maybe I can later refine my model based on that.