CMPS290C Machine Learning

Fall 17 Warmuth

handed out We 11-1-17

Homework Assignment 5

due Th 11-16-17 by midnight

Programming homework!
Work in groups of up to 4.
This time you can work with people you have worked with before.

We set up a Kaggle competition for classifying tweets by Hillary Clinton versus Donald Trump. The base task is not to work with a bag of words model. Instead you are to use an LSTM recurrent Neural Net that takes the sequence of words in each text into account.

We provide a labeled training set. You need to tokenize it and train an LSTM network. We also provide an unlabeled test set. You need to submit your probabilities for predicting each class for all tweets in this test set. Kaggle will rate you by the logistic loss wrt the true (hidden) labels.

Begin by checking out the data format on the competition website.

Find a name for your group. Kaggle will provide a leader board so that you can see how well you are doing compared to other groups in the class.

You are to use tensor flow. Summarize what you did and what you tried in a maximum 4 pages writeup submitted via Canvas.

Again you need to split the task into 2 and 2 of you should work on each subtask: preprocessing and tokenizing the text versus training the LSTM.

Base task: LSTM with a one hot word model.

Extra credit:

- Try a bag of words model as well: Determine what accuracy you can achieve with logistic regression or convolutional neural nets?
- Try word models you get from the internet and see whether they improve the accuracy.

Have fun with this mini project.