Markus Dabell
CS 542
Nikhil Krishnaswamy

Perception Differences among NLP Models

The goal of my project will be to determine skews and differences among classifier models trained with different subsets of the same dataset.

(scientific/empirical question I'm seeking to answer) I'd like to know to what extent the training data impacts the classifications a model produces.

(hypothesis) More specifically I think the skew will depend on the amount that the input data is skewed, and I think it will be logarithmic in nature. (If the input data is extremely skewed, or exponentially skewed among different models, I hypothesize the results will be skewed, but they won't be exponentially skewed).

(common approaches in the field to your problem) I've seen a few research papers about how to correct bias or how to identify bias, and what the biased results are, but from a semi-cursory search I couldn't find any mathematically rigorous examples that mathematically linked the input data to the results nor compared different models trained with different subsets of the same data.

My idea in a nutshell:

model 1 is trained on 100 movie reviews ranging from 1 star to 5 stars:

25 1-star, 20 2-star, 20 3-star, 20 4-star, 15 5-star

model 2 is trained on 100 movie reviews ranging from 1 star to 5 stars:

15 1-star, 20 2-star, 20 3-star, 20 4-star, 25 5-star

model 3 is trained on 100 movie reviews ranging from 1 star to 5 stars:

20 1-star, 20 2-star, 20 3-star, 20 4-star, 20 5-star

I want to do this on a bigger scale than the above toy example. I'd like to use the "Distributional Semantics and word2vec" lecture code and data. My plan is to use DyNet on the data from the lecture, but to adjust the data and train 5+ models on different portions of data.

In short:
-get dataset from lecture

-get code to run with the dataset from lecture

-write code to split the dataset in a few different ways to present skews in the data

-train multiple models on each biased split of the dataset

-use mathematical methods to analyze the results (differences between models, determine any correlations with input data skew, create graphs showing the differences between models and the amounts of skew related to the input data subsets)