NLP William Jiang

Prof. Parikh 12/5/19

Assignment 2 Paper

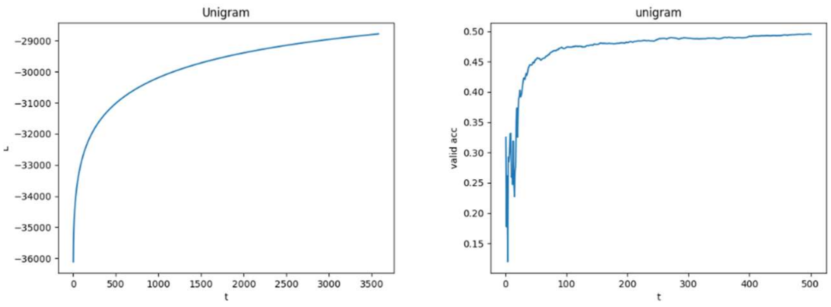
Introduction

In this assignment, we seek to predict the labels of a series of string input names to their respective label class such as drugs, persons, places, movies, and companies. Through supervised learning using a discriminative model, the classifier will learn the weights by maximizing the conditional log likelihood by maximizing the gradient ascent. The model takes in as input a string name, extracts from it features, and uses the set of features as the input matrix. We will explore 3 models with increasing test accuracy: the unigram model, n-gram model, custom-feature set model, and finally a combination model.

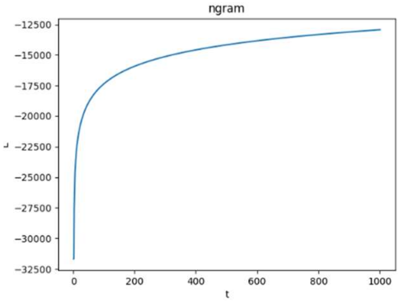
Methods

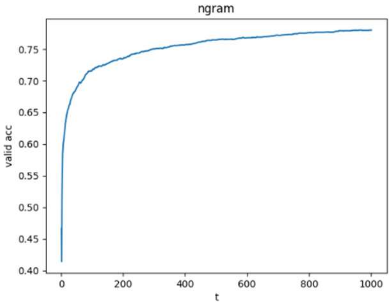
Since all the models follow the same concept of a discriminative classifier, we will explore the general steps taken to achieve optimal predictions as in the source code. First, the dataset is fed through the model where its data rows are parsed and featured in the model feature space. The data is converted into a numpy matrix of [row_size x feature_size] with a corresponding weight matrix of [class_size x feature_size]. The model will use these to step through gradient ascent, where the partial gradient with respect to each output class is added to the weight matrix. The procedure continues until the L-2 norm between the previous and new weight matrices crosses a threshold. The final output will rely on the final weight matrix state.

Results

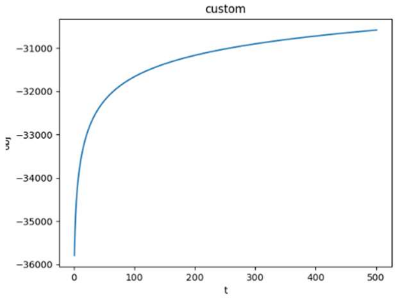


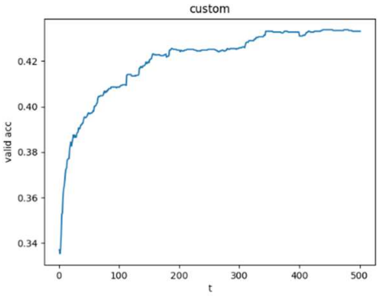
The first naïve model to be explored is the Unigram model where each character of a string is considered a feature in the classifier. For example, “Xylec” will be broken down into the following features: {‘X’, ‘y’, ‘l’, ‘e’, ‘c’}. As shown in the figure above, the model converges at an objective function of around -29,000. The validation accuracy is 47.9694%.



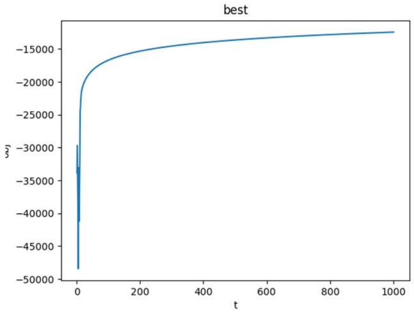


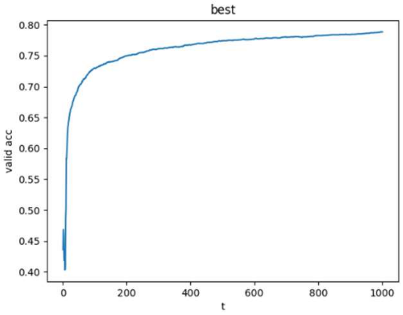
The second model to be explored is an n-gram model where n > 1. In our case, we conjoined the bigram and trigram model into one. The figure above shows the model converging at a rapid rate towards -12500, a much higher objective function over the unigram model. The model achieved a staggering improvement with validation accuracy of 77.9813%.





The next model that was considered was a custom model whose feature space was hand-picked. Features such as prefixes and suffixes as well as entity names such as ‘corporation’ were selected upon perusing through the dataset. The feature size had a size of only 104, compared to a similar unigram feature set of size 118 and a much bigger n-gram feature size of 20,702. The model achieved a modest accuracy of ___ that is similar to the unigram model.





Finally, our best model combines the features of the ngram and custom models. It achieves a slightly higher validation accuracy of 78.8455%

Our final test set accuracy using the best model is TBD.