ML4CL: Assignment 3

Zarah Weiß

July 31, 2016

Task 1

Train an evaluate a logistic regression classifier, based on both word unigrams, and word bigrams. (You can limit the features to the words that appear at least N (e.g., 5) times in the corpus.) Report average accuracy of your models.

Average accuracy for stratified 10-folds cross validation was 85.35%.

Task 2

Repeat the first exercise, but this time use word vectors, with a multi-layer perceptron. You can represent each document as the sum (or average) of all the word vectors in the document (for bigrams you can concatenate the vectors). Report average accuracy.

For unigrams, average accuracy was 69.75% for the best model. For bigrams, average accuracy was 67.20% for the best model.

Task 3

Design and train a convolutional neural network (CNN) for the same task. You are free to choose the architecture, but make sure your convolutions cover at least bigrams, and the model learns multiple convolutions (features maps). Briefly explain your network architecture, and report the accuracy of the model.

The CNN uses a three dimensional input matrix representing each of the overall 2,000 documents as a two dimensional matrix of word vectors. Word vector length was chosen to be 50 and documents were limited to the first 200 words. Shorter documents are treated as having an according number of zero word vectors. This leads to a $2{,}000 \times 200 \times 50$ input matrix.

The CNN itself works with two dimensional convolutions. For the first layer, ten output filters of size 3×50 are applied, i.e. they read three words at once, hence, consider trigrams. For the second layer, five output filters of size 2×50 are applied. Then, max pooling is performed, using a 2×2 pooling size, which reduces dimensionality by half. Then, dimensionality reduction is performed, in order to make the dimensions fit for a final, fully connected layer. For all these layers, the activation function was chosen to be *relu*, as this is the most commonly used activation function. Finally, the results are transformed to be interpreted in a binary distinction, by adding a fully connected output layer with a single output dimension using a sigmoid activation function to map the result to probabilities between 0 and 1.

Average accuracy for 10-folds cross validation for this model was 52.25%

Task 4

Briefly (not more than half a page) discuss the results you have obtained. Include comparison of each model for their accuracy as well as computational complexity.

For the task of binary movie review classification three machine learning algorithms were applied: Logistic regression, multilayer perceptrons, and convolutional neural networks, which were trained and tested using stratified 10-folds cross validation. **Logistic regression** achieves the highest accuracy with 85.35%. It is also the simplest model in terms of computational complexity: On the one hand, it relies on simple unigram and bigram counts, instead of employing word vectors. On the other hand, it uses fewer parameters, than the others. In fact, a **MLP** without any hidden layers would correspond to the logistic regression analysis. However, the two MLP implemented consist of multiple hidden layers. Their number was varied to find the best results. The unigram MLP has 302 hidden layers and 30,805 parameters, resulting in an accuracy of 69.75%. For bigrams, the highest accuracy of 67.20% is achieved for 420 hidden layers and 65,731 parameters. Most complex, however, is the **CNN**, which only achieves an accuracy of 52.25% with 69,026 parameters. It runs for over 5 hours.

So it turns out, that the least computationally complex model also performs best in this case. However, with more successfully tuned hyper-parameters, this would probably change and MLP and CNN would return better results. Unfortunately, these more suited hyper-parameters could not be found in the course of experimenting.