Advanced Methods in Machine Learning

Exercise 2 – conclusion

Student: Hodaya Koslowsky

The accuracy on the test for the three models:

(The accuracies are chosen as the highest from several runs of the trainings and tests)

Model 1: 63.8 %

Model 2: 75.4 %

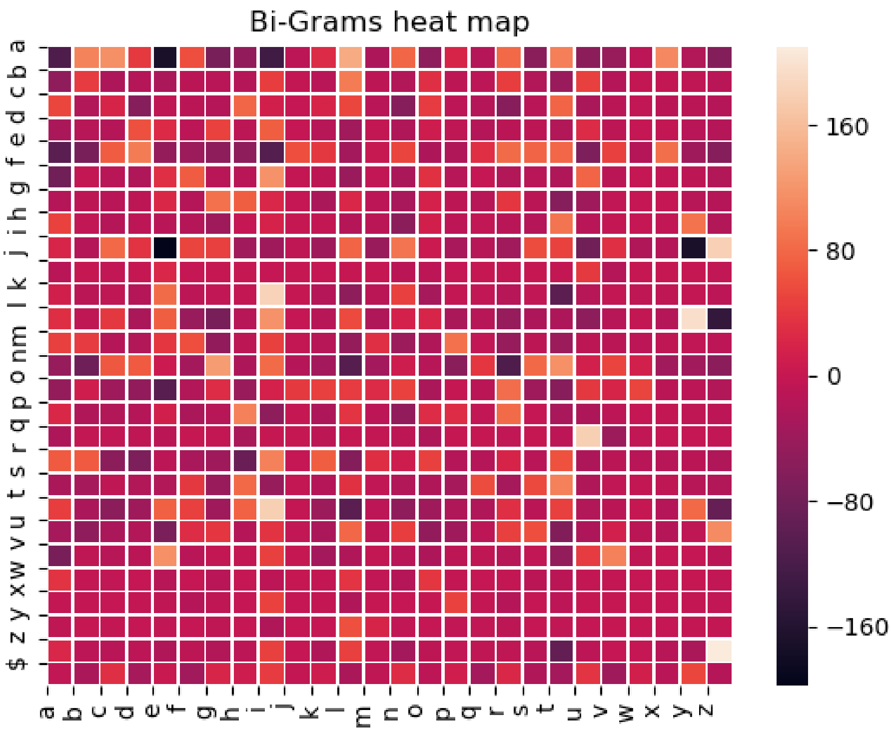
Model 3: 78.0%

The accuracy achieved by the third model is the highest. It uses the maximum amount of information that is available in the examples, between all 3 options.

There shouldn't be a large difference between the first and second model, since they use the same information from the examples, and the inference and update rules are essentially the same.

Bonus 1: Bi-Grams heat-map:

The heat map shows the weights that model 3 learned. (It is not normalized)



We can see that the higher scores are given to common bi-grams, such as 'ly', 'qu' and 'ng', and the lowest scores are for uncommon bi-grams, such as 'ie', 'iy', 'ae', 'lz'.

Not all the scores make sense when looking at this heat-map, which is reasonable since the accuracy is not very high.

Bonus 2: Structured SVM

The main difference is that we look at the problem as an optimization problem, of minimizing a certain task loss function of the true tags and the predictions, plus regularize W.

So, the goal is to minimize the following:

$$\frac{|W|^2}{\lambda}+\sum_{i=1}^{n}\max_{y'}\left(0,\tilde{l}(y',y_i)+W*\varphi(x_i,y')-W*\varphi(x_i,y_i)\right)$$

(Where n is the number of examples, x_i and y_i are the examples, y' is from the possible tag classes, φ is the feature function, W is the weights matrix that we learn, λ is the regularization factor, $\tilde{l}$ is a loss function)

Unlike the structured perceptron, we use a loss function for a possible prediction y' and the real tag y_i. For example, the hamming function: the number of features that are not equal in y' and y_i. In addition, we want to minimize the difference between $W*\varphi(x_i,y')$ of the prediction and $W*\varphi(x_i,y_i)$ of the true class.