Machine learning for computational linguistics: Assignment 2

Deadline: June 8, 2016

The exercises you need to complete for this assignment are related to classification and cross validation. The submission procedure is the same as the first assignment, except of course you should change the assignment number in your email. Please follow the guidelines carefully.

Data

For this homework, we will use the same data set from the first assignment. You will need the following files:

- English: en-ud-pos.txt
- German: de-ud-pos.txt
- Japanese: ja-ud-pos.txt

Reminder: each line in these files contains POS tags of a single sentence. The POS tags in each line are separated by spaces.

Exercises

Read all the exercises first. It may save you some time during preprocessing.

1. Fit a logistic regression classifier distinguishing the sentences in German from the sentences in the other two languages using the number of words in a sentence as the only predictor. Do **not** use any regularization for exercises 1 to 3 (See the notes at the end of this document).
   - Write down the fitted model equation (estimated probability given the predictor)
   - Write a brief (with no more than three sentences) interpretation of the coefficients
   - Report *accuracy*, *precision*, *recall* and F_1*-score* of the fitted model on the training data
2. The model you fit in exercise 1 suffers from the class imbalance problem (besides the poor predictor). A workaround for this problem is to decide for a probability threshold other than 0.5.
   - Find and report the best threshold value that maximizes the F_1-score
   - Write down the discriminant function (the function $f(X)$ whose value is positive for the positive instances (sentences in German) and negative for the negative instances)
   - Report the *accuracy*, *precision*, *recall* and F_1*-score* at the best threshold value
3. This time, besides the sentence length, use 15 additional predictors that indicate the relative frequencies of POS unigrams within the sentence. Evaluate your model with probability threshold of 0.5, and report *accuracy*, *precision*, *recall* and F_1*-score* values. Does this model suffer from the class imbalance problem equally?
4. In this exercise you will fit two three-way classifiers predicting the language, rather than the binary distinction between German and others, using the same predictors as in exercise 3.
   - Fit two separate models to the complete data, one with L1, the other one with L2 regularization. For both models, use the regularization parameter $\lambda = 50$.
   - Briefly explain the differences between the coefficient values.
   - Calculate and compare accuracy of both L1 and L2 regularized models.
   - Tabulate the confusion matrix of the L2-regularized model you fit in the previous step.
5. Using the same model in exercise 4 with L2 regularization, evaluate the model accuracy using 10-fold cross validation, and report the average accuracy and its standard error.

Additional notes

- `glm()` in R does not perform any regularization, nor does it work with multiple classes. For these, you need an additional library. There are many, but you should be fine with *LiblineaR* for these exercises (and in many real world problems).

  If you are using Python *skleran* library, it does regularization by default. If you do not want it (or instructed not to use it), you can set the parameter `C` to a very high value.

  For others systems, you should consult the documentation and find out how to set the regularization parameter.

- The predictors (features) we used in the exercises above are chosen for demonstration purposes. These are far from the ideal predictors for the task. Although it is not part of this assignment, you are encouraged to explore other predictors. Here are some examples:
  - frequency or existence of all bigrams or larger ngrams (but higher level ngrams are unlikely to be useful, why?)
  - unigrams/bigrams that are found at the beginning/end of sentences
  - K-most frequent _n_grams (for various choices of K and n)
  - ratio of certain POS tags to all others or another tag (e.g., some languages have fewer or no determiners, others do not use many adpositions due to morphologically specified case)
  - ...

- The class imbalance problem demonstrated in the first two exercises is accented by the fact that the class overlap is extremely high, given the predictor we used in the first two exercises. Another interesting point is that since our sample has no reasonable prior distribution of classes, biasing towards majority class is not useful as it might have been if the data was sampled from a distribution of interest.

  In many reasonable data sets and tasks, the problem is not as severe.