Lecture Graphical Models
`https://ml01.zrz.tu-berlin.de/wiki/Main/SS09_GraphicalModels`
Machine Learning Group, TU Berlin

Instructors: Dr. Ulf Brefeld, Dr. Marc Toussaint
Tutor: Tobias Lang, `lang@cs.tu-berlin.de`

Sheet 7

Due: 16 June 2009

1. Parameter learning for a Hidden Markov Model

Please take a look at the "Spanish News Wire" data-set containing 3146 sentences of a Spanish new agency which you can find on the course website. The format looks as follows:

```
♯♯ sequence xx     // marks beginning of new sentence

<token> <label> <feature-vector>     // each line = word of a sentence
```

The following labels are used:

- 0 location-begin
- 1 other
- 2 company-begin
- 3 person-begin
- 4 person-inside
- 5 misc-begin
- 6 company-inside
- 7 location-inside
- 8 misc-inside

A token is a word or a punctuation symbol. You can ignore the feature vector.

Please write a program (in Java or C/C++) that learns a HMM from these data. You need to parse the data-set and estimate the label-label transition probabilities and the label-token emission probabilities. With regard to exercise 2, make use of a parameter S that specifies how many sentences are used for learning, namely the first S sentences. Please send your program to `brefeld@cs.tu-berlin.de`.

2. Evaluation of HMM learning

In this exercise, you have the pleasure to investigate the dependency of the learned HMM parameters on the number of training examples. In particular, for different choices of $S \in \{100, 500, 1000 \dots\}$, learn a HMM by means of your program of exercise 1. Use the resulting HMM to predict the labels of the remaining sentences in the data-set by applying the Viterbi you programmed for exercise sheet 6.

Compare your predictions with the true labellings of the test sentences. In particular, make use of two evaluation metrices: (1) the percentage of fully correctly labeled sentences; (2) the percentage of correctly labeled tokens over the whole test data-set.

Please visualize your results in diagrams with S on the x-axis and the evaluation metrics on the y-axis.

3. Perceptron learning: problem dual

Please confer the slides of lecture 7 and show that in perceptron learning the following identity holds in the dual problem formulation:

$$f(x, y) = f_1(x, y) + f_2(x, y)\,. \tag{1}$$