Pattern and Speech Recognition Tutorial
Exercise 12

Exercise 1 (2 points)

In HMMs we did inference over the joint probability distribution:

$$log(P(x_1, x_2..x_n, \pi_1, \pi_2..\pi_n)) = log(P(\pi_0)) + \sum_{i=1}^{n} log(P(\pi_i|\pi_{i-1}) + \sum_{i=1}^{n} log(P(x_i|\pi_i)) \quad (1)$$

Show that you can rewrite this distribution the following way:

$$log(P(\pi_0)) + \sum_{i=1}^{n} log(P(\pi_i|\pi_{i-1}) + \sum_{i=1}^{n} log(P(x_i|\pi_i)) = \sum_{i=0}^{n} \sum_{j=1}^{N} \lambda_j f_j(\pi_i, \pi_{i-1}, x, i) \quad (2)$$

Where x is the observed sequence, n is the length of the observation and $f_j(\pi_i, \pi_{i-1}, x, i) \in \{0, 1\}$ are feature functions. The first sum runs over each position i in the observation,the second sum runs over each feature function j . $\lambda_j \in \mathbb{R}$ are weights corresponding to each feature function.

`Example feature functions`

- $f_1(\pi_i, \pi_{i-1}, x, i) = 1$ if $\pi_i = s_1$, 0 else.
- $f_2(\pi_i, \pi_{i-1}, x, i) = 1$ if $\pi_i = s_1$ and $\pi_{i-1} = s_2$, 0 else.
- $f_3(\pi_i, \pi_{i-1}, x, i) = 1$ if $x_i = a$, 0 else.

Specify the feature functions and the corresponding weights for **??**.

Exercise 2 (2 points)

Install CRF++. Read the documentation. For the template file:

```
# Unigram
U:%x[0,0]
# Bigram
B
```

and training data:

```
NN B
IN O
DT B
```

List the feature functions that the software is going to generate.

Exercise 3 (3 points)

Train a model for Spanish named entity recognition with the provided template file and data. The first two columns are the features, the last column is the true labeling. Evaluate on the test data and report the F-score (use the python file evaluate.py).

Exercise 4 (2 points)

$x = 2, 1, 5, 6, 2, 1, 3, 1, 6, 2$ $\pi = F, F, L, F, F, F, L, F, L, F$
Compute the transition matrix and the emission matrix using maximum likelihood. What can you observe? What is it that the HMM fails to capture?

Submission architecture

You have to generate a **single ZIP file** respecting the following architecture:

```
tutorial1_<matriculation_nb1>_<matriculation_nb2>_<matriculation_nb3>
|
+--- source
|      |
|      +------- file 1
|      +------- file 2
|      +------- ...
+--- rapport.pdf
+--- README.txt
```

where

- **source** contains the source code of your project,
- **rapport.pdf** is the report where you present your solution with **the explanations (!)** and the plots,
- **README** which contains group member informations (name, matriculation numbers and emails) and a **clear** explanation about how to compile and run your source code

The ZIP filename has to be :

`tutorial11_<matriculation_nb1>_<matriculation_nb2>_<matriculation_nb3>.zip`

You have to choose between the following languages **python** or **matlab**. Other languages won't be accepted.

Some hints

We advice you to follow the following guidelines in order to avoid problems :

- Avoid building complex systems. The exercises are simple enough.
- Do not include any executables in your submission, as this will cause the e-mail server to reject it.

Grading

Send your assignment to the tutor who is responsible of your group:

- Gerrit Gromann `gerritgr@gmail.com`
- Sbastien Le Maguer `slemaguer@coli.uni-saarland.de`
- Kata Naszdi `b.naszadi@gmail.com`

The email subject should start with `[PSR TUTORIAL 12]`