Solution to exercice 2

Model

- Model to predict punctuation:
  - 7 classes: . ! ? ; : , AnyOtherChar
- Model: $P(y|x_t, x_{t-1}, \cdots, x_{t-n}) \sim M(\theta)$
  - x_{t-i} = char at position $t-i$
  - y = punct. class computed from char at position $t+1$

Draw a graphical model

$X_i \to Y$ for all i

Is it generative or discriminative ?

Discriminative

How many parameters

Let m be the size of the char lexicon (=nb of different characters).

We use a multinomial distribution, so we have one parameter per possible configuration of $(x_t, x_{t-1}, \cdots, x_{t-n}, y)$.

We thus have:

$$m \times m \times \cdots \times m \times 7 = 7m^{n+1}$$

parameters.

Conditional likelihood

A training corpus is composed of T sequences $\{x_t, x_{t-1}, \cdots, x_{t-n}, y\}_t$

The conditional likelihood is the function of θ:

$$f(\theta) = P(y|x_t, x_{t-1}, \cdots, x_{t-n}; \theta)$$

For a single training example, we have:

$$f(\theta) = \theta_{x_t, x_{t-1}, \cdots, x_{t-n}, y}$$

So for the whole corpus:

$$f(\theta) = \prod_{t}^{T} \theta_{x_t, x_{t-1}, \cdots, x_{t-n}, y}$$

We want the optimum:

$$\hat{\theta} = \arg\max_{\theta} f(\theta) = \arg\max_{\theta} \log f(\theta)$$

Let $N_{x_t, x_{t-1}, \cdots, x_{t-n}, y}$ be the number of occurrences of one specific sequence $(x_t, x_{t-1}, \cdots, x_{t-n}, y)$ in the whole corpus. We can group together every similar sequence:

$$f(\theta) = \prod_{s=(x_t, x_{t-1}, \cdots, x_{t-n}, y)} \theta_s^{N_s}$$

Equivalently

$$\log f(\theta) = \sum_{s=(x_t, x_{t-1}, \cdots, x_{t-n}, y)} N_s \log \theta_s$$

We want to maximize

$$\log f(\theta)$$

under the constraints:

$$\sum_{y} \theta_{x,y} = 1$$

for all $x = (x_t, \cdots, x_{t-n})$.

Method of Lagrange Multipliers: maximize:

$$f'(\theta) = \log f(\theta) - \sum_{x} \lambda_x \left(\sum_{y} \theta_{x,y} - 1 \right)$$

The derivative of f' is:

$$\frac{\partial f'(\theta)}{\partial \theta_{x,y}} = \frac{\partial \log f(\theta)}{\partial \theta_{x,y}} - \lambda_x = \frac{N_{x,y}}{\theta_{x,y}} - \lambda_x = 0$$

So we have a system of equations, for all x, y:

$$\theta_{x,y} = \frac{N_{x,y}}{\lambda_x}$$

The last derivatives give us back the constraints:

$$\frac{\partial f'(\theta)}{\partial \lambda_x} = \sum_{y} \theta_{x,y} - 1 = 0$$

By replacing every $\theta_{x,y}$:

$$\sum_y \frac{N_{x,y}}{\lambda_x} = 1$$

So

$$\lambda_x = \sum_y N_{x,y}$$

By replacing λ_x in the system:

$$\theta_{x,y} = \frac{N_{x,y}}{\sum_z N_{x,z}}$$

Unobserved sequences

For $n = 1$, there shouldn't be any problem, but with greater n, we will find some $N_s = 0$ for sequences that are never observed in the training corpus.

Our model will always return a proba = 0 for such sequences. But this might be due to a "limited" training corpus, so it's best to still give a small $\epsilon > 0$ probability for such sequences.

This is done by *smoothing* our trained model = adding the same artificial count $\eta > 0$ to every possible sequence, as if we merge into our training corpus another corpus composed of η occurrences of every s:

$$N_s \leftarrow N_s + \eta$$

for all possible s.

We can show that this is equivalent to adding a Dirichlet prior to our Multinomial distribution.