Machine learning for computational linguistics: Assignment 1

Deadline: May 11, 2016

The exercises you need to complete for this assignment are below after the general guidelines about all homework submissions. Please first read these guidelines carefully.

General information

- Homeworks are to be done individually. You are encouraged to discuss your solutions with others, and study together. However, the answers to the questions have to be your own words, and the scripts/programs should be your own work.

Submission information

Please follow the steps below carefully:

- Submit your homeworks via email to `ccoltekin@sfs.uni-tuebingen.de`.
- The 'From' address of the email has to contain your full real name, preferably sent from your university email address.
- The subject should contain the string `ML assignment 1`.
- Your submissions should include two attachments:
  - A *plain text* or *PDF* (if you need to include formulas, figures etc.) file answering the questions in the assignment text. Please do not send any other format. Convert to PDF if you prepare it in Word, Open/Libre office etc. Do not forget to state your name on this document too. This file should contain *only and all* answers to the questions asked (no code listing, additional data or information).
  - A single file containing the code you have written for the project. If the code contains multiple files, please put them together in a single archive file using a common archiving utility. The archive should also contain a file, named `README`, describing each file in the archive briefly.

Evaluation

- You get a mark between 6 to 10 (inclusive) if your homework is satisfactory (generally this means you made an effort to do it completely) and on time.
- You get 0 if you do not submit the homework or your homework is not satisfactory.
- Late homeworks are not accepted.
- The average of the individual homework scores will determine the contribution of homeworks to the final score from the course.

Assignment 1: probability and information theory

Deadline: May 11, 2016

Data

For this homework, we will use the following three files containing POS tag sequences from three different languages (click on the link to download them):

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

Each line in these files contains POS tags of a single sentence. The POS tags in each line are separated by spaces. The data is from Universal Dependencies project and the same POS tag set is used for all three languages. Some POS labels are combined in two of the languages to make the exercises below easier.

Exercises

1. Create a table containing relative frequencies of each POS tag for each language. Your table should look like:

POS	English	German	Japanese
`ADJ`			
`ADV`			
...			

and each cell should contain the relative frequency of corresponding POS tag and the language.

2. Taking the relative frequencies calculated in question 1 as the probability estimates, calculate the entropies of POS distribution for each language.

3. Calculate and list the KL divergence (in both directions) of the POS tag distributions between all all language pairs as well as the uniform POS tags distribution.

4. For bigrams '`DET NOUN`', '`ADP DET`', '`VERB PUNCT`', calculate and list

   - *Joint probability* of the tags occuring in the given order:
     $P(\text{first tag} = t_1, \text{second tag} = t_2)$, short notation: $P(t_1, t_2)$
   - *Conditional probability* of the second tag of the bigram given the first:
     $P(\text{second tag} = t_2 | \text{first tag} = t_1)$, short notation: $P(t_2|t_1)$
   - *Pointwise mutual information* (PMI) of the POS tag bigram:
     $PMI(\text{first tag} = t_1, \text{second tag} = t_2)$, short notation: $PMI(t_1, t_2)$

Present your result in a table like following for each language:

Bigram	$P(t_1, t_2)$	$P(t_2\|t_1)$	$PMI(t_1, t_2)$
`DET NOUN`			
`ADP DET`			
`VERB PUNCT`			

where, t_1 represents the first POS tag in the bigram, t_2 is the second POS tag in the bigram.