SNLP 2017
Exercise 8

Submission date: 23.06.2016, 23:59

EM Algorithm

1) (7 points) Suppose data $Z = (X, Y)$ is generated by $p(X, Y|\theta) = p(Y|X, \theta)p(X|\theta)$, where X is observed and Y is unobserved. EM algorithm functions as follows:

```
1)Random initialize $\theta^0$
2)for t=0..T
    E-step: compute $p(Y|X, \theta^t)$
    M-step: derive $\theta^{t+1} = arg\max_\theta \mathbb{E}_{p(Y|X,\theta^t)}[p(X, Y|\theta)] = arg\max_\theta[\sum_Y p(Y|X, \theta^t)p(X, Y|\theta)]$
```

- (1 points) Correlate the above notations with the algorithm in Slide 43, Chapter 7. Find the meaning of X,Y and θ in that environment.
- (2 point) Prove $p(X|\theta^{t+1}) \geqslant p(X|\theta^t)$

  Given two coins A and B, θ_A and θ_B refer to the probability that coin A and B will land on head. We do the following procedure 5 times: Randomly pick one of the two coins (equal chance), toss the coin 10 times, the results are as below:

  H T T T H H T H T H

  H H H H T H H H H H

  H H H H H H T T H H

  H T H T T T H H T T

  T H H H T H H H T H

  H means head and T means tail.

- (1 point) Apply EM algorithm to find θ_A and θ_B. Explain the meaning of X,Y and θ in this environment
- (2 point) Initialize with $\theta_A = 0.6, \theta_B = 0.5$. Compute the E-step and M-step by hand. What is the value of θ_A and θ_B after one iteration?
- (1 point) Write a function to compute the value of θ_A and θ_B after 10 iterations.

Information-theoretic Approach

2) (3 points) NLTK has a corpus reader for the Senseval 2 dataset. This data set contains data for four ambigous words: hard, line, serve and interest. The provided code loads the dataset for you and converts each sample into sample object. The sample class has two attributes: context and label. Label is the ground truth sense of the ambiguous word. Run the Flip-Flop Algorithm in Slide 40, Chapter 7 to partition these 4 words. For every ambiguous word, use the most frequent 10 context words as the indicator words. Output the partition results in the report file.

1 Submission Instructions: Read carefully

- You can form groups of maximum 2 people.
- Submit only 1 archive file in the ZIP format with name containing the MN of all the team members, e.g.:

```
Exercise_01_MatriculationNumber1_MatriculationNumber2.zip
```

- Provide in the archive:
  - your code, accompanied with sufficient comments,
  - a PDF report with answers, solutions, plots and brief instructions on executing your code,
  - a README file with the group member names, matriculation numbers and emails,
  - Data necessary to reproduce your results [1]
- The subject of your submission mail must contain the string [SNLP] (including the braces) and explicitly denoting that it is an exercise submission, e.g:

```
[SNLP] Exercise Submission 01
```

- Depending on your tutorial group, send your assignment to the corresponding tutor:
  - David Adelani: *s8daadel@stud.uni-saarland.de*
  - Rajarshi Biswas: *rbisw17@gmail.com*
  - Xiaoyu Shen: *s8xishen@stud.uni-saarland.de*

2 General Information

- In your mails to us regarding the tutorial please add the tag [SNLP] in the subject accompanied by an appropriate subject briefly describing the contents.
- Feel free to use any programming language of your liking. However we strongly advise in favor of Python, due to the abundance of available tools (also note that Python3 comes with an excellent native support of UTF8 strings).
- Avoid using libraries that solve what we ask you to do (unless otherwise noted).
- Avoid building complex systems. The exercises are simple enough.
- Do not include any executable files in your submission, as this may cause the e-mail server to reject it.
- **In case of copying, all the participants (including the original solution) will get 0 points for the whole assignment. Note: it is rather easy to identify a copied solution. Plagiarism is also not tolerated.**
- **Missing the deadline will result in 0 point for the whole assignment. So make sure to submit on time.**
- **Please submit in your solutions necessary to support your claims. Failure to do so, might results in reduction of points in the relevant questions.**

[1] If you feel that these files are beyond reasonable size for an email submission and also reasonably convenient, please provide a means for us to access them online

- Each assignment has 10 points and perhaps some bonus points (usually 2 or 3). In order to qualify for the exams, you need to have 2/3 of the total points. For example, in case there are 12 assignments, you need to collect at least 80 out of the 120 points to be eligible for the exams. A person that gets 10 plus 2 bonus points in every exercise, needs to deliver only 7 assignments in order to be eligible for the exams, since 7*12=84.
- Attending the tutorial gives 1 point increase for the corresponding assignment.
- Presenting a solution will give bonus equal to the point of the presented question (Up to 4 poins). The bonus is divided evenly in the group member. Each group can present up to 3 times.
- Exercise points (including any bonuses) guarantee only the admittance to the exam, however have no further effect on the final exam grade.