SNLP

Exercise 4

Submission date: 26.05.2016, 23:59

1 (2 points)

The random variable y has a probability density function,

$$\begin{aligned} p(y) &= (1-\theta) + 2\theta y, && 0 < y < 1 \\ &= 0 && otherwise \end{aligned} \tag{1}$$

for $-1 < \theta < 1$. There are n observations y_i, $i = 1, 2, ..., n$ drawn independently from this distribution.

- Write the cumulative distribution function(`https://en.wikipedia.org/wiki/Cumulative_distribution_function`) of y. (1 point)
- Derive the expected value for y. (1 point)

2 (2 points)

$y_1, ..., y_n$ are n independent draws from an exponential distribution. The probability density function for each y_i is $f(y_i|\theta) = \theta^{-1} exp(-y_i/\theta)$ where $y_i > 0$, $\theta > 0$. The exponential distribution has the property $E(y_i) = \theta$.
Derive

- The observation specific log-likelihood function $l_i(\theta)$. (1 point)
- The maximum likelihood estimator for θ. (1 point)

3 (3 points)

Construct a different vocabulary using each of the documents provided in the training folder in Materials, e.g., for n documents in the training folder you will create n different vocabularies.

- Now use the document provided in the test folder in Materials to compute **OOV** using each of the vocabularies that you created. (1 point)
- Plot **OOV** vs size-of-vocabulary and explain the plot. (1 point)
- How do Out-Of-Vocabulary words affect tasks like computing probability for a sequence of words? Explain a possible remedy. (1 point)

4 (3 points)

- From NLTK obtain the text “austen-emma.txt”. Start with text normalization and change all different versions of the word “you” like “your, you’ll, you’ve” into “you”. Print out the modified version in a text file. (1 point)
- Compute the correlation for the word “you” with different distances of 1 to 50 ($\forall d \in [\,1, 50]\,, d \in \mathbb{N}$). Use the correlation function provided on slide 47 of chapter 4 in SNLP. Now produce the plot correlation vs. distance. (1 +1 points)

5 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*

[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

6 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.**

- 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.