Statistical Natural Language Processing (SS-2018)
Exercise 3

(Prepared in format inspired by Dolores Umbridge, Hogwarts)

Submission Deadline: 12.05.2018, 23:59

Entropy

1) (1.5 points) Let X and Y be discrete random variables. The joint entropy $H(X,Y)$ measures the entropy of a joint probability distribution $P(x,y)$, and is given by the following formula:

$$H(X,Y) = -\sum_x \sum_y P(x,y) log_2[P(x,y)]$$

Prove the following identity between the joint and conditional entropy:

$$H(X,Y) = H(X) + H(Y|X)$$

.

Please note that conditional entropy $H(Y|X)$ is defined as:

$$H(Y|X) = \sum_x P(x) H(Y|X=x)$$

2) (0.5 points) Assuming that *all* the alphabets in German are equally likely to occur (though in reality that is not the case), please compute the entropy H. (Don't forget to write the unit in your final result!)

3) (1 point) From the lecture, we learned about *Cross-Entropy* ($H(p,q)$, between two probability distributions p and q). In the context of *language modeling*, briefly explain what is p and q. Also, explain how can we compute CE in practice.

Kullback-Leibler Divergence

Also know as **Relative Entropy**, it is defined as:

$$D(p||q) = \sum_x p(x) log \frac{p(x)}{q(x)}$$

4) (4 points) For this exercise, you should select three texts, from a source such as NLTK or Project Gutenberg. Choose two in the same language (for example, English), and one in a different language (for example, German). We will consider the token-level distribution, using the maximum likelihood estimation with *Lidstone smoothing**.

- What is *smoothing* and why do we need it? *Hint*: Think about where would it fail if we don't do smoothing.
- Pre-process (lowercase, remove punctuation) and tokenize[1] the text. You may use any functions from NLTK for this task.

[1] `https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html`

- Implement a function[2] that calculates the **entropy** of a given text, using the formula from the slides. Run it on the three texts that you have selected, and report the results. Comment briefly on how to interpret the numbers from the results.
- Implement a function that takes two texts as arguments, and calculates the KL divergence $D(p||q)$ between them. For this task, you will need to use the probabilities after applying *Lidstone smoothing*. Calculate and report the **KL divergence** between the two texts of the same language, and between two of the texts in different languages. Comment on any difference in the results.

*The formula for *Lidstone smoothing* is:

$$P_{lidstone}(w) = \frac{count(w) + \alpha}{N + \alpha V} \tag{1}$$

where N is the total number of tokens, and V the size of the vocabulary. We'll use $\alpha = 0.1$ for this example.

Text Compression

5) (0.5 points) In the lecture, we have encountered *Lagrange Multipliers* (denoted as λ). Briefly explain the scenarios where *Lagrange Multipliers* could be helpful in achieving our goal(s).

6) (2.5 points) In this task, we will determine an encoding for the following string of characters:

aaabacddbabacdd

- (1.5 points) Use the Huffman coding algorithm (explained here: `https://en.wikipedia.org/wiki/Huffman_coding#Informal_description`) to calculate, by hand, the encoding for the string above. Report the code obtained for each character, and use it to encode the string above.
- (1 point) Using the formula in the slides, calculate the optimal length of the code. Comment on how this relates to the length of the code you found in the previous question.

---

[2]Implementation hint: You may want to begin both of these tasks by writing a function that takes a single text as an argument, and returns a dictionary with words as keys, and their probabilities as values.

Submission Instructions

- You must form groups of 2 to 3 people
- Submit only 1 archive file in the ZIP format with name containing the MN of all the team members, e.g.:

  Exercise_03_MatriculationNumber1_MatriculationNumber2.zip

- Provide in the archive:
  i. your code, accompanied with sufficient comments
  ii. a PDF report with answers, solutions, plots and brief instructions on executing your code
  iii. a README file with the group member names, matriculation numbers and emails
  iv. Data necessary to reproduce your results

- 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 MatriculationNumber1 MatriculationNumber2

- Depending on your tutorial group, please send your assignment to the **corresponding tutor**:
  * Mo. 16-18: Lukas Lange *s9lslang@stud.uni-saarland.de*
  * Th. 14-16: Harshita Jhavar *snlp18.thursday@gmail.com*
  * Fr. 10-12: Marimuthu Kalimuthu *mkalimuthu@lsv.uni-saarland.de*