UNIVERSITÄT DES SAARLANDES

Statistical Natural Language Processing

Summer Semester 2020
Prof. Dr. Dietrich Klakow

Exercise Sheet V

Submission Deadline: June 12th, 23:59

1 Mutual Information

1.1 Independence of variables (2 points)

Let X_1, X_2 and Y be binary random variables. Assume that we know that $\mathrm{I}(X_1;\, Y) = 0$ and $\mathrm{I}(X_2;\, Y) = 0$. Is it correct to say that $\mathrm{I}(X_1,\, X_2;\, Y) = 0$? Prove or provide a counter example.

1.2 Mutual information and entropy (3 points)

Imagine that you play a game with your friend who chooses an object in the room and you need to guess which object it is. You are allowed to ask the friend questions and the answer is a deterministic function of both the object o and your question q, i.e. $f(o, q)$. The output of this function can be represented as a random variable A.

Suppose that the object O and question Q are also random variables that are independent of each other. Then $I(O; Q, A)$ can be interpreted as the amount of uncertainty about O that we remove if the values of question Q and answer A are known. Show that $I(O; Q, A) = H(A|Q)$. Provide an interpretation to this equation.

Hint: $H(X) = 0$ if X is deterministic.

2 Encoding

2.1 Huffman encoding (1.5 points)

The Huffman encoding algorithm is an optimal compression algorithm when only the frequency of individual letters are used to compress the data. The main idea is that more frequent letters get shorter code. You can find a detailed explanation here: `https://en.wikipedia.org/wiki/Huffman_coding#Informal_description`.

Suppose that you have a string *BBBDBACCDBDBACC*. Use the Huffman algorithm to calculate the binary encoding for this string. Report the code you obtained for each character, and use it to encode the string.

Based on the formula from the lecture, calculate the optimal code length. How does it relate to the actual length of the code you obtained?

2.2 Entropy and encoding (1.5 points)

Let's assume that you encountered a new language which has 3 vowels and 3 consonants.

a) First, you counted the individual letter frequencies and got the following result:

p	t	k	a	i	u
1/8	1/4	1/8	1/4	1/8	1/8

Using the distribution from above, compute per-letter entropy $H(L)$ for the given language.

b) After doing some research, you realized that the language has a syllable structure. All words consist of CV (Consonant Vowel) syllables. Now you have a better model of the language and the joint distribution P(C,V) looks as follows:

	p	t	k
a	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{1}{16}$
i	$\frac{1}{16}$	$\frac{3}{16}$	0
u	0	$\frac{3}{16}$	$\frac{1}{16}$

Compute the per-letter $H(L)$ and per-syllable $H(C, V)$ entropy using the probabilities from the table above.

Note that in order to obtain the probabilities of individual letters, you will need to estimate the marginal probabilities from the table above and then divide the marginal probabilities by two. This is needed because the table presents per-syllable, not per-letter statistics.

c) Compare per-syllable probabilities based on the probability distribution in (a) to the result you obtained in (b). Explain the difference.

3 Out-of-Vocabulary Words

OOV and vocabulary size (2 points)

In this exercise you will use the multilingual corpus provided in `exercise5_corpora` to investigate the relation between the vocabulary size and the OOV rate in different languages. For each corpus, lower-case the text, remove non-alphabetical characters, and apply whitespace-based tokenization.

a) Partition each corpus into a training corpus (80% of the word tokens) and a test corpus (20% of the word tokens).

b) Construct a vocabulary for each language by taking the most frequent 10k (10,000) word types in the training corpus. The vocabulary set should be ranked by frequency from the most frequent to the least frequent.

c) Compute OOV rate (percentage) on the test corpus as the vocabulary grows by 1k words. This means that you should increase the vocabulary size starting from 1k words to 2k, 3k, 4k etc.

d) For each language, plot a logarithmic curve where the x-axis represents the size of the vocabulary and the y-axis represents the OOV rate. Each curve should have a legend to identify the language of the text corpus. Write your observations regarding the OOV rate for different languages.

Your solution should include the plot for part (d) and the source code to reproduce the results.

4 Bonus

Many questions (2 points)

Similar to part 1.2, imagine that you play a game with your friend and try to identify an object by asking questions and the friend gives you binary answers (yes or no). Let's assume that you need to ask 6.5 questions on average to guess the object and that all your questions are good (i.e. reduce the search space in an optimal way). Can you find a lower bound to the number of objects in the game? Justify your answer.

Hint: you can use the theorem about entropy-bound from the lecture slides.

Submission Instructions

The following instructions are mandatory. Please read them carefully. If you do not follow these instructions, the tutors can decide not to correct your exercise solutions.

- You have to submit the solutions of this exercise sheet as a team of 2 students.
- If you submit source code along with your assignment, please use Python unless otherwise agreed upon with your tutor.
- NLTK modules are not allowed, and not necessary, for the assignments unless otherwise specified.
- Make a single `ZIP` archive file of your solution with the following structure
  - A `source_code` directory that contains your well-documented source code and a `README` file with instructions to run the code and reproduce the results.
  - A `PDF` report with your solutions, figures, and discussions on the questions that you would like to include. You may also upload scans or photos of high quality.
  - A `README` file with group member names, matriculation numbers and emails.

- Rename your `ZIP` submission file in the format

  `exercise05_id#1_id#2.zip`

  where `id#n` is the matriculation number of every member in the team.

- Your exercise solution must be uploaded by only one of your team members under *Assignments* in the *General* channel on Microsoft Teams.

- If you have any problems with the submission, contact your tutor before the deadline.