UNIVERSITÄT DES SAARLANDES

Statistical Natural Language Processing

Summer Semester 2020
Prof. Dr. Dietrich Klakow

Exercise Sheet IV

Submission Deadline: June 5th, 23:59

Information Theory

1) **Information content** (1 point)
Assume that the probability of being male is $p(M) = 0.5$ and so likewise for being female $p(F) = 0.5$. Suppose that 20% of males are T (i.e. tall) and that 6% of females are tall.
Calculate the probability that if somebody is "tall" (meaning taller than 6 ft or whatever), that person must be male.

Now, if you know that somebody is male, how much information do you gain (in bits) by learning that he is also tall? How much do you gain by learning that a female is tall? Finally, how much information do you gain from learning that a tall person is female?

2) **Entropy** (3 points)
Consider a binary symmetric communication channel (see figure 1), whose input source is the alphabet $X = 0, 1$ with probabilities $0.5, 0.5$; whose output alphabet is $Y = 0, 1$ and the channel is:

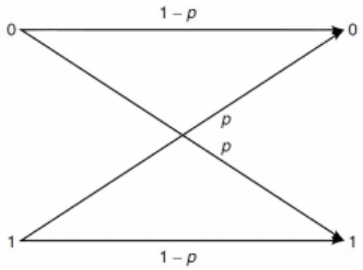

Figure 1: Binary symmetric communication channel with inputs X on the left side, outputs Y on the right side and p is the probability of transmission error.

- (0.5 points) What is the entropy of the source, $H(X)$?
- (1 point) What is the probability distribution of the outputs, $p(Y)$, and the entropy of this output distribution, $H(Y)$?
- (1 point) What is the joint probability distribution for the source and the output, $p(X, Y)$, and what is the joint entropy, $H(X, Y)$?
- (0.5 points) What is the mutual information of this channel, $I(X; Y)$? See: https://en.wikipedia.org/wiki/Mutual_information

3) **Kullback-Leibler Divergence** (2 points)
First, provide proof that the Kullback-Leibler Divergence does not follow the triangle inequality and thus cannot be considered a true distance metric (1 point).
Now, let the random variable X have three possible outcomes a, b, c. Consider two distributions on this random variable:

Symbol	$p(x)$	$q(x)$
a	$\frac{1}{2}$	$\frac{1}{3}$
b	$\frac{1}{4}$	$\frac{1}{3}$
c	$\frac{1}{4}$	$\frac{1}{3}$

Calculate $H(p)$, $H(q)$, $D(p||q)$ and $D(q||p)$. Verify that in this case $D(p||q) \neq D(q||p)$ (1 point).

4) **Codes** (2 points)
Consider the following source alphabet and its letter probabilities:

A	B	C	D
$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{8}$

- (0.5 points) What is the entropy H, in bits, of the above source alphabet?
- (1 point) Why are fixed length codes inefficient for alphabets whose letters are not equiprobable? Discuss this in relation to Morse Code.
- (0.5 points) Offer an example of a uniquely decodable prefix code for the above alphabet. What features make it a uniquely decodable prefix code?

5) Martian codes and Kraft's inequality (2 points)

Martians have landed on earth and you have found one of their code books. In it you find a code entry of the form:

$$S = \begin{vmatrix} S_1, ..., S_m \\ p_1, ..., p_m \end{vmatrix}$$

The S_i's are encoded into strings from a D-symbol output alphabet in a uniquely decodable manner. If $m = 6$ and the code word lengths are $(l_1, l_2, ..., l_6) = (1, 1, 2, 3, 2, 3)$, what is the most likely base for the Martian code (i.e. what is a good lower bound for D)?

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

`exercise02_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.