Information Theory

Lab 4: Conditional entropy of natural languages

Description of the tasks

All tasks realized during classes are `.pdf` files formatted similarly as this document. The tasks will be of different kinds. Every task will be appropriately marked:

- Tasks to be realized during classes are marked with $\square$ – you won't get points for them, but you still need to do them.
- Pointed tasks to be realized during classes are marked with $\diamond$ – you need to do them during class and show to your teacher, and in the case you don't manage to do so (or are absent) they become your homework ($\star$).
- Homeworks are marked with $\star$ – they also have assigned a number of points, and you need to deliver them to your teacher before a deadline (usually before the next class).
- You may use any programming language you like for the programming tasks, but you are limited to only the standard library of that language and libraries widely accepted as standard.

Objective

During this class we will apply in practice conditional entropy to the analysis of natural languages.

How do we know that hieroglyphs are a meaningful text?

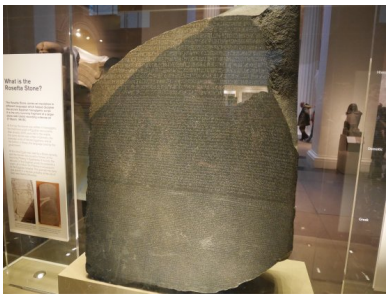

Figure 1: Rosetta Stone (British Museum)

Finding the Rosetta stone was crucial for deciphering Egyptian hieroglyphs. It was found in 1799 during Napoleon's expedition to Egypt, and later in 1801 taken by Great Britain as a result of peace negotiations.

The stone tablet contains the same text written in three ways: hieroglyphs, demotic ("everyday-use hieroglyphs"), and Greek. Based on these findings, the bigger part of the hieroglyphic writing system was deciphered by 1822.

Finding similar translations allowed to decipher some other ancient languages. The Mesopotamian language was deciphered in 1857, and the Mayan language in 1952. However, there are still other ancient languages that remain incomprehensible to us.

Figure 2: Clay tablet of Indus Valley Civilization

The main question for those incomprehensible texts is: are those pictograms a system of writing at all? For example, the Indus Valley Civilization, which existed between 3300 and 1300 BC, left behind clay tablets containing several symbols each. Archeologists long disputed, whether it is writing system at all. The conditional entropy analysis gave the positive answer to this question.

Conditional entropy (recap)

Conditional entropy $H(Y|X)$ measures the amount of uncertainty (entropy) of a random variable Y that remains, if we know the value of a random variable X. We can compute it using the formula:

$$H(Y|X) = -\sum_{x \in X, y \in Y} p(x,y) \cdot \log p(y|x)\,, \tag{1}$$

where $p(x,y)$ is the joint probability of events x and y, and $p(y|x)$ is the probability of an event y dependent on x. The above formula can be derived from:

$$\begin{aligned} H(Y|X) &= \sum_{x \in X} p(x) \cdot H(Y|X=x) \\ &= \sum_{x \in X} p(x) \cdot \left(-\sum_{y \in Y} p(y|x) \cdot \log p(y|x) \right) \end{aligned}$$

Conditional entropy is used, for example, in the decision tree learning algorithms to (sort of greedily) select at each step an attribute which will minimize the conditional entropy of a decision attribute.

When applying the above definition to the analysis of language, $p(x,y)$ is the joint probability of an n-gram x and a following character/word y against all other possible $(n{+}1)$-grams in the corpus, and $p(y|x)$ is the probability of y occurring after x. In the homework, we will consider the conditional entropy of rank n, where n denotes the length of an n-gram x.

Conditional entropy $H(Y|X)$ can be also computed as a difference between joint entropy $H(Y,X)$ and (marginal) entropy $H(X)$:

$$H(Y|X) = H(Y,X) - H(X) = -\sum_{x \in X, y \in Y} p(x,y) \log p(y,x) \; + \; \sum_{x \in X} p(x) \log p(x)$$

The language of dolphins and the search for extraterrestrial civilizations

When researching the language of the civilization of ancient Indus, as well as all modern languages, we will notice that conditional entropy of different languages is similar and decreases as the rank becomes bigger. This is true for both the conditional entropy of characters and words.

This is an expected result. Imagine that our task is to guess a word selected from a book. Our options are limited and we can only guess randomly, alternatively choose the most frequent word. If instead we were supposed to do the same but given the word that was before, our probability of success would be much higher. The task will be even simpler when we are provided with two previous words. The set of possible words decreases (for example, because of different parts of speech), certain words become much more

probable (for example, “wise man” vs “wise politician”), or both of those effects at the same time. Since our uncertainty decreases, we can infer that the entropy also does so.

A similar analysis was conducted on the noises generated by different species of animals in order to compare the complexity of their communication. The animals with high intelligence (or rather: the intelligence we as humans would be more apt to call as such), like for example dolphins, have the similar entropy characteristic of their communication system as human languages.

This line of research interested the astronomer and astrophysicist Frank Drake. In 1959 he initiated the SETI (Search for Extraterrestrial Intelligence) project, which is continued to this day. The goal of this project is to find extraterrestrial civilizations by capturing and analyzing the radio and light signals coming to us from space. This search is conducted using, among others, the conditional entropy of signals.

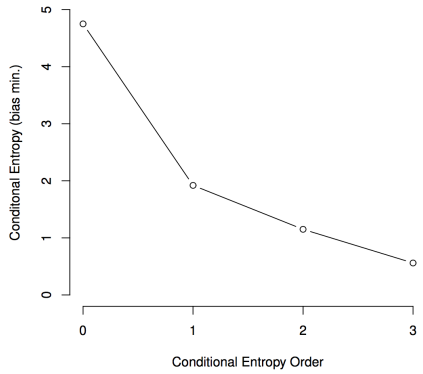

Figure 3: The conditional entropy of the language used by dolphins

1 Detecting if a text was written in a natural language *10pt*◇

Task

For this task will be needed text corpuses, which can be downloaded from http://www.cs.put.poznan.pl/ibladek/students/timkod/lab3.zip

Compute the entropies of characters and words, and their conditional entropies of higher ranks (it may be useful to think here about standard entropy as simply a conditional entropy of rank 0) for a sample of English (file `norm_wiki_en.txt`). ***(2pt)***

Then do the same for Latin (file `norm_wiki_lo.txt`). ***(2pt)***

You can also analyze, for your own curiosity, samples of other languages:

- Esperanto (file `norm_wiki_eo.txt`),
- Estonian (file `norm_wiki_et.txt`),
- Somali (file `norm_wiki_so.txt`),
- Haitian (file `norm_wiki_ht.txt`),
- Navaho (file `norm_wiki_nv.txt`),

Using the observed values of conditional entropy of different ranks, answer if the following text samples contain the natural language or not (drawing a plot might help): ***(6pt)*** (***(1pt)*** for each correctly guessed file)

- sample0.txt,
- sample1.txt,
- sample2.txt,
- sample3.txt,
- sample4.txt,
- sample5.txt.

Additional remarks

- Some of the languages contain additional letters, which were normalized to their closest equivalent in the Latin alphabet, so that files contain only 26 small letters of Latin alphabet, digits, and spaces.
- In order to make this easier for your teacher to check, as a unit of entropy use bits.

- Checked will be the correctness of numerical results, your answers, and their justifications (please write for each sample, why do you think that it is a language or not).
- The tasks can, but not have to, be tricky.

Sources

- https://en.wikipedia.org/wiki/Entropy_(information_theory)
- https://en.wikipedia.org/wiki/Conditional_entropy
- http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf
- www.mdpi.com/1099-4300/16/1/526/pdf
- https://en.wikipedia.org/wiki/Rosetta_Stone
- https://en.wikipedia.org/wiki/Indus_Valley_Civilisation
- https://pl.wikipedia.org/wiki/Search_for_Extraterrestial_Inteligence
- http://www.inference.org.uk/itprnn/book.pdf