OPT13 - Information Theory
TP2: Compression, Prediction, Generation
Text Entropy

Stella Douka[*], Guillaume Charpiat[†]

Credits: Gaétan Marceau Caron

March 29th, 2024

In this TP we are interested in compressing and generating texts written in natural languages.

Given a text of length n, a sequence of symbols is just a vector $(x_1, \dots, x_n)$ where each x_i is a symbol i.e. $x_i =$ a, b, c, ...

In order to model the sequence of symbols we need a joint probability distribution for each symbol in the sequence, namely $p(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$. If our alphabet had M symbols, for modelling a sequence of length n we would need M^n probabilities. Thus some assumptions are required in order to reduce this dimensionality. In this case we will use two different models for p, the IID and the Markov Chain model.

IID Model

The first model assumes:

$$p(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} p(X_i = x_i) \tag{1}$$

i.e. that the symbols in a sequence are independent and identically distributed. In this can we need now only M probabilities, one for each symbol. One can generalize and use symbols not of a single character but of multiples ones. For example using 3 characters per symbol, the symbols would be of the form *aaa*,*aab*,...,*zzz*. When using k characters per symbols in an alphabet of M characters, the needed probabilities would be M^k.

Markov Chain Model

The Markov Chain model assume a limited range of dependence of the symbols.

[*] styliani.douka@inria.fr

[†] https://www.lri.fr/~gcharpia/informationtheory/

Indeed for an order k Markov Chain:

$$p(X_i|X_{i-1}, \ldots, X_1) = p(X_i|X_{i-1}, \ldots, X_{i-k}) \tag{2}$$

The meaning of the above structure is that the i-th symbol in the sequence depends only on the previous k symbols. We add the *time invariant* assumption, meaning that the conditional probabilities do not depend on the time index i i.e. $p(X_i|X_{i-1}, \ldots, X_{i-k}) = p(X_{k+1}|X_k, \ldots, X_1)$. The most common and widely used Markov Chain is the Markov Chain of order 1:

$$p(X_i|X_{i-1}, \ldots, X_1) = p(X_i|X_{i-1}) \tag{3}$$

In this case the conditional probability $p(X_i|X_{i-1})$ can be expressed using M^2 numbers.

Questions

1. Interpret the time invariant assumption associated to our Markov chains.
2. How can we rewrite a Markov chain of higher order as a Markov chain of order 1?
3. Given a probability distribution over symbols, how to use it for generating sentences?

In order to construct our IID and Markov Chain models we need some *text*. Our source will be a set of classical novels available here. We will use the symbols in each text to learn the probabilities of each model.

Practical For both models, perform the following steps:

1. For different orders of dependencies, train the model on a novel and compute the associated entropy. What do you observe as the order increases? Explain your observations.
2. Use the other novels as test sets and compute the cross-entropy for each model trained previously. How to handle symbols (or sequences of symbols) not seen in the training set?
3. For each order of dependencies, compare the cross-entropy with the entropy. Explain and interpret the differences.
4. Choose the order of dependencies with the lowest cross-entropy and generate some sentences.
5. Train one model per novel and use the KL divergence in order to cluster the novels.

Implementation hints

1. It is possible to implement efficiently the two models with dictionaries in Python. For the IID model, a key of the dictionary is simply a symbol and the value is the number of occurrences of the symbol in the text. For a Markov chain, a key of the dictionary is also a symbol, but the value is a vector that contains the number of occurrences of each character of the alphabet. Notice that a symbol may consist of one or several characters. Note also that there is no need to explicitly consider all possible symbols; the ones that are observed in the training set are sufficient.

2. A low probability can be assigned to symbols not observed in the training set. How to choose this probability?