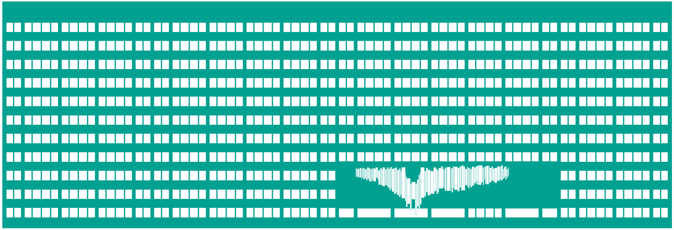
VŠB TECHNICKÁ UNIVERZITA OSTRAVA

VSB TECHNICAL UNIVERSITY OF OSTRAVA



www.vsb.cz

Analysis and Signal Compression

Markov models

Michal Vašinek

VŠB – Technická univerzita Ostrava

FEI/EA404

michal.vasinek@vsb.cz

2020, February 12th

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Recapitulation 1

- Entropy gives a fundamental limit to our ability to efficiently represent an input message.
- It expresess the average number of bits needed to represent one symbol.
- Each symbol x can be in theory compressed by $-\log p(x)$ bits.
- If we compress below the entropy we will be unable to reconstruct an original message.

Recapitulation 2

- Shannon equation:

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \tag{1}$$

- Better encoding can be achieved using conditional probability:

$$H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y|x) \tag{2}$$

Calculating zero-th order entropy

The zeroth order empirical entropy H_0 of a text T of length n over an alphabet $\Sigma = \{c_0, \ldots, c_{\sigma-1}\}$ of size σ is a lower bound for a compressor which encodes each symbol independently of any of its preceding or following symbols.

$$H_0(T) = \sum_{x \in \Sigma} \frac{f(x)}{n} \log \frac{n}{f(x)} \tag{3}$$

where $f(x)$ is a frequency of symbol x in text T.

More advanced compressors make use of the context before a symbol and choose different codes dependent on the preceding context. Say, the length of the context is of fixed length k, then the k-th order empirical entropy $H_k(T)$ is a lower bound for compressors which encode each symbol with a codeword that depends on the preceding k symbols.

$$H_k(T) = \frac{1}{n} \sum_{\omega \in \Sigma^k} |T_\omega| H_0(T_\omega) \quad (4)$$

where T_ω is the concatenation of all symbols in T which follow the occurences of the substring ω in T.

Calculating k-th order entropy

Alternatively:

$$\begin{aligned} H_k(T) &= -\sum_{\omega \in \Sigma^k} \sum_{x \in \Sigma} \frac{f(\omega x)}{n} \log \frac{f(\omega x)}{f(\omega)} \\ &= -\sum_{\omega \in \Sigma^k} \sum_{x \in \Sigma} p(\omega x) \log p(x|\omega) \end{aligned} \tag{5}$$

Models

- Information source generates sequences of letters (some data) over the input alphabet.
- Good model of the data can be useful in estimating the entropy of the source and may help to produce better compression algorithms.
- We would like to develop a mathematical model of the data.

Probability models - zero knowledge model

The simplest model - zero knowledge model

In the simplest model, we assume that each letter is generated by the source independent of every other letter and each occurs with the same probability.

- In the simplest model we assume that we don't know anything about the information source.

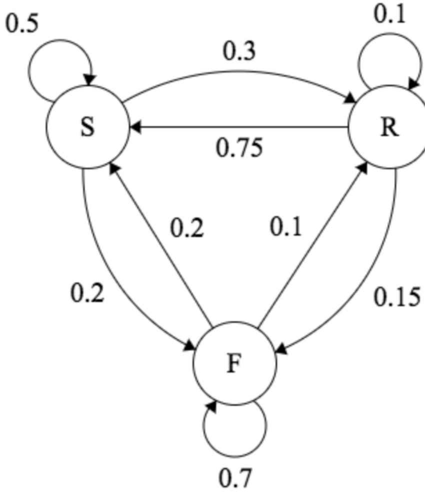
Probability models - i.i.d. models

i.i.d. model

In the i.i.d. (independent and identically distributed) model, we assume that each letter is generated by the source independent of every other letter but their probabilities may differ.

- We can easily compute the entropy of the source of this model using Shannon equation.
- If the assumption of the independency does not fit our observation of the data, we can generally find better compression schemes if we find a way how to describe the dependence of elements of the data sequence on each other.

Dependencies in the model



Markov model

- Also called a discrete time Markov chain or Markov process.
- A Markov model is a model that satisfies the Markov property.

Markov property

A model has the Markov property if the conditional probability distribution of future states of the process (conditional on both past and present values) depends only upon the present state; that is, given the present, the future does not depend on the past.

$$P(\mathit{future}|\mathit{present}, \mathit{past}) = P(\mathit{future}|\mathit{present}) \tag{6}$$

- Memoryless process.

Markov model - formal definition

1-order Markov property:

$$P(x_n|x_{n-1},\ldots,x_0) = P(x_n|x_{n-1}) \tag{7}$$

- $x_{n-1},\ldots,x_0$ represents knowledge about past values.
- x_{n-1} current state.

k-order Markov property:

$$P(x_n|x_{n-1},\ldots,x_0) = P(x_n|x_{n-1},\ldots,x_{n-k}) \tag{8}$$

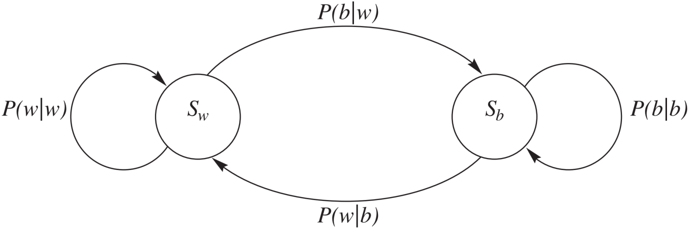
- In other words, knowledge of the past k symbols is equivalent to the knowledge of the entire past history of the process.
- Sequences $\{x_{n-1},\ldots,x_{n-k}\}$ are called the states of the process.

- If the size of the source alphabet is σ then the number of states is $M = \sigma^k$.
- The entropy of a finite state model with states S_i is the average value of the entropy at each state:

$$H = \sum_{i=1}^{M} P(S_i)H(S_i) \tag{9}$$

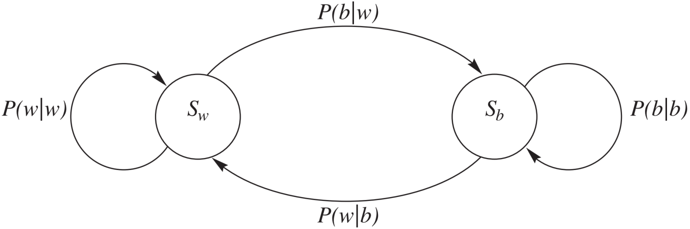
Markov model - two state example

- Suppose a two state Markov model, for instance black/white image.
- We assume that the occurence of the white(black) pixel depends on the current pixel color.

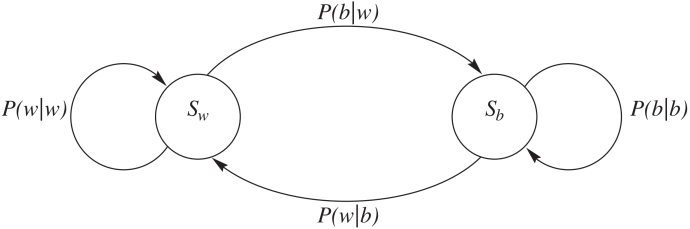


Markov model - two state example

- Define two state S_w and S_b representing that current pixel is white resp. black.
- Next, define transition probabilities that we go from state S_i to state S_j, i.e. $P(w|b)$ and $P(b|w)$ and the probability that of being in each state $P(S_w)$ and $P(S_b)$.



- What is the value of $P(w|w)$ and $P(b|b)$?

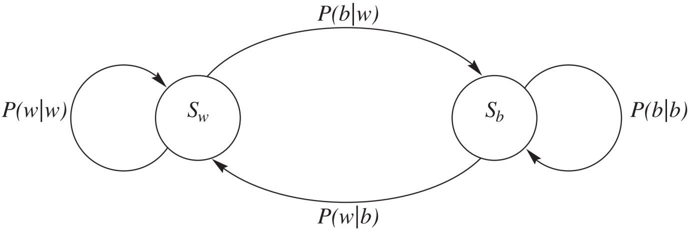


- What is the value of $P(w|w)$ and $P(b|b)$?

$$P(w|w) = 1 - P(b|w)$$

and

$$P(b|b) = 1 - P(w|b)$$



Markov model - two state example

- State probabilities are $P(S_w) = 30/31$, $P(S_b) = 1/31$.
- Transition probabilities are $P(w|b) = 0.01$, $P(b|w) = 0.3$.
- Entropy using i.i.d. assumption:

$$H = -30/31 \log 30/31 - 1/31 \log 1/31 = 0.206 \textit{bits}$$

- Entropy based on Markov model:

$$H(S_b) = -0.3 \log 0.3 - 0.7 \log 0.7 = 0.881 \textit{bits}$$

and

$$H(S_w) = -0.01 \log 0.01 - 0.99 \log 0.99 = 0.081 \textit{bits}$$

- The Markov model based entropy $H = 0.107$ bits.

Markov model in text compression

- In the text processing, probability of the next letter is influenced by preceding letters.
- k-th order Markov models knows as finite context models in the modern compresison algorithms.
- Shannon experiment comes with estimate of 0.6-1.3 bits per letter of english text.
- Rule of thumb: the longer the context the better is its predictive value and smaller the entropy.
- The problem is the representation of the model. If the model has to be transmitted with the data then for long context, e.g. k=5 and the alphabet size $\sigma = 52$: $52^5 = 380,204,032$ states.
- Modern algorithms adapt transition probabilities adaptively as they process letters.

Source literature and additional reading

- Introduction to data compression by Khalid Sayood.
- Introduction to statistical signal processing by Robert Gray.

DÄŻkuji za pozornost

Michal Vašinek

VŠB – Technická univerzita Ostrava

FEI/EA404

michal.vasinek@vsb.cz

2020, February 12th

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY