Statistical Methods in Natural Language Processing (NLP)

Class 14: Machine Learning: Markov Chains, Hidden Markov Models

Charalambos (Haris) Themistocleous

Department of Philosophy, Linguistics and Theory of Science, Centre for Linguistic Theory and Studies in Probability

Machine Learning

1. Markov Chains
2. Hidden Markov Models

Applications of Hidden Markov Models

1. Speech Recognition.
2. They are a probabilistic function of a Markov Process.
3. They are related to Markov Processes/Markov Chains/Markov Models

Markov Models

1. Sequence of Non-Independent Random Variables
2. Need to Know the current random variable to predict the future and we do not need to know the values of all the past random variables in the sequence.
3. Nice metaphor from Manning and Schütze: "If you want to predict the number of books in a library it is enough to know how many books are in the library today and not in the previous weeks".

Example

Suppose $X = (X_1, \ldots, X_T)$ is a sequence of random variables taking values in some finite set, the state space:

$$S = \{S_1, \ldots, S_N\}$$

Markov Properties

Limited Horizon:

$$P(X_{t+1} = S_k|X_1, /ldots, X_t) = P(X_{t+1} = S_k|X_t)$$

Time invariant (stationary):

$$= P(X_2 = s_k|X_1)$$

X is a Markov chain, or to have the Markov property.
One can describe a Markov chain by a stochastic transition matrix A:

$$a_{ij} = P(X_{t+1} = S_j|X_t = S_i)$$

Here, $a_{ij} \geq 0, \forall_{i,j} and \sum_{j=1}^{N} = a_{ij} = 1, \forall_i$

Markov Chains

Also, we have to specify specify Π, the probabilities of different initial states for the Markov chain:

$$\pi_i = P(X_1 = S_i)$$

where
$\sum_{j=1}^{N} \pi = 1$
$\rightarrow$ This vector can be avoided by specifying that the Markov model always starts off in a certain extra initial state, S_0, and then using transitions from that state contained within the matrix A to specify the probabilities that used to be recorded in Π.

Markov Chain as a (nondeterministic) finite state automaton