# Discrete Time Discrete Symbol Sequence Prediction Using HMM

**Zhengli Zhao, Karthik Prasad, Abhisaar Sharma**

## Abstract

This paper describes our graphical-model approach to the data-oriented project SPiCe (The Sequence PredictIction ChallengE). Hidden Markov model (HMM) is a statistical Markov model in which the system being modelled is assumed to be a Markov process with latent states and can be presented as the simplest Bayesian network. We show our approach in modelling the problem as a HMM and using learning techniques like Baum Welch algorithm and spectral learning methods to learn the parameters of this graphical model.

## 1 Introduction

The Sequence PredictIction ChallengE (SPiCe) is a competition where the aim is to learn a model that allows the ranking of potential next symbols for a given prefix and submitting a ranking of the 5 most probable next symbols. Training datasets consist of variable length sequences with a fixed number of symbols. The competition uses real-world data from different fields like Natural Language Processing, Biology, Signal Processing, Software Verification.

We can use a Hidden Markov Model to model this problem, where the training sequences can be treated as discrete time observables (emission variables) and the unobserved latent states can be used to capture the intrinsic sequence structure. A hidden Markov model can be considered a generalization of a mixture model where the hidden variables which control the mixture component to be selected for each observation, are related through a Markov process rather than independent of each other.

## 2 Methodology

While trying to model the training sequence as an HMM, we would like to learn the parameters of this graphical model. We will present two approaches that we explored, the first one is the BaumWelch algorithm which uses the EM algorithm to find the maximum likelihood estimate of the parameters of a hidden Markov model given a set of observed feature vectors. Let $X_t$ be a discrete hidden random variable with $N$ possible values. We assume the $P(X_t|X_{t-1})$ is independent of time t. Let the state transitions be described by $A$ which is a homogeneous time independent stochastic transition matrix.

$A = \{a_{ij}\} = P(X_t = j|X_{t-1} = i)$

The initial state distribution is given by

$\pi_i = P(X_1 = i)$ The observation variables $Y_t$ can take one of $K$ possible values. The probability of a certain observation at time $t$ for state $j$ is given by

$B = \{b_j(y_t)\} = P(Y_t = y_t|X_t = j)$

Write the algorithm and the update

## 3 Implementation

There were some obstacles we overcame

- Training with python library took a lot of time - wrote code in java
- Sequences too long, added minibatch type descent
- Mixed the weights together with step size of 1/n
- Underflow problem in forward backward algorithm, the forward problem algorithm values quickly became very small, so modified the algorithm and implemented the normalized version of it. Other things tried for underflow which were unsuccessful was using BigDecimal class of java

which was too slow and turning the probabilities as logarithmic since there were quantities involved with sum inside the log.

### 3.1 SECOND LEVEL HEADING

Second level headings must be flush left, all caps, bold and in point size 10. One line space before the second level heading and 1/2 line space after the second level heading.

#### 3.1.1 Third Level Heading

Third level headings must be flush left, initial caps, bold, and in point size 10. One line space before the third level heading and 1/2 line space after the third level heading.

Fourth Level Heading

Fourth level headings must be flush left and initial caps. One line space before the fourth level heading and 1/2 line space after the fourth level heading.

### 3.2 CITATIONS, FIGURES, REFERENCES

#### 3.2.1 Citations in Text

Citations within the text should include the author's last name and year, e.g., (Cheesman, 1985). Reference style should follow the style that you are used to using, as long as the citation style is consistent.

For the original submission, take care not to reveal the authors' identity through the manner in which one's own previous work is cited. For example, writing "In (Bovik, 1970), we studied the problem of AI" would be inappropriate, as it reveals the author's identity. Instead, write "(Bovik, 1970) studied the problem of AI."

#### 3.2.2 Footnotes

Indicate footnotes with a number[1] in the text. Use 8 point type for footnotes. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a 0.5 point horizontal rule 1 inch (6 picas) long.[2]

#### 3.2.3 Figures

All artwork must be centered, neat, clean, and legible. Figure number and caption always appear below the figure. Leave 2 line spaces between the figure and the caption. The figure caption is initial caps and each figure numbered consecutively.

Make sure that the figure caption does not get separated from the figure. Leave extra white space at the bottom of the page rather than splitting the figure and figure caption.

Figure 1: Sample Figure Caption

#### 3.2.4 Tables

All tables must be centered, neat, clean, and legible. Table number and title always appear above the table. See Table 1.

One line space before the table title, one line space after the table title, and one line space after the table. The table title must be initial caps and each table numbered consecutively.

Table 1: Sample Table Title

| PART | DESCRIPTION |
|---|---|
| Dendrite | Input terminal |
| Axon | Output terminal |
| Soma | Cell body (contains cell nucleus) |

[1] Sample of the first footnote

[2] Sample of the second footnote

### Acknowledgements

Use unnumbered third level headings for the acknowledgements title. All acknowledgements go at the end of the paper.

### References

References follow the acknowledgements. Use unnumbered third level heading for the references title. Any choice of citation style is acceptable as long as you are consistent.

J. Alspector, B. Gupta, and R. B. Allen (1989). Performance of a stochastic learning microchip. In D. S. Touretzky (ed.), *Advances in Neural Information Processing Systems 1*, 748-760. San Mateo, Calif.: Morgan Kaufmann.

F. Rosenblatt (1962). *Principles of Neurodynamics.* Washington, D.C.: Spartan Books.

G. Tesauro (1989). Neurogammon wins computer Olympiad. *Neural Computation* **1**(3):321-323.