Residential time HMM

László Keresztes

Applied Mathematics MSc, ELTE-TTK

Supervisor: Balázs Csanád Csáji, SZTAKI, ELTE

December 16, 2021

Hidden Markov Models

A Hidden Markov Model (HMM) is a discrete-time, discrete Markov chain $z_t \in \{1, \ldots, N\}$ with a parametric observation model $p(x_t|z_t = j)$ for $j = 1, \ldots, N$.

Observe the x_t sequence and infer the z_t hidden sequence. Match the hidden states with the states of a real process.

Joint distribution:

$$p(z_{1:T}, x_{1:T}) = p(z_1) \prod_{t=2}^{T} p(z_t|z_{t-1}) \prod_{t=1}^{T} p(x_t|z_t)$$

Start probabilities and observation model

Start probabilities $\pi_i = p(z_1 = i)$, probability distribution on $\{1, \dots, N\}$.

Observation model comes from a parametric family: e.g.

$$p(x_t | z_t = j) = \mathcal{N}(x_t | \mu_j, \Sigma_j)$$

or

$$p(x_t | z_t = j) = Cat(p_{j,1}, \dots, p_{j,L})$$

Great flexibility on the observation generation.

Transition model

Transition model: time independent $N \times N$ matrix

$$A_{ij} \doteq p(z_t = j | z_{t-1} = i)$$

Residential time in a hidden state T_i always has a geometric distribution.

$$T_i \sim Geo(1 - A_{ii})$$

Limited flexibility in the non-transition generation.

HMM inference and learning

Inference and learning - computing the following:

- $\alpha_t(i) = p(z_t = i | x_{1:t})$
- $\beta_t(j) = p(x_{t+1:T} | z_t = j)$
- $\gamma_t(i) = p(z_t = i | x_{1:T})$
- $\xi_{t,t+1}(i,j) = p(z_t = i, z_{t+1} = j | x_{1:T})$

EM learning in HMM

Expectation-Maximization algorithm is used for approximate ML estimation by iteratively reestimating the hidden variables and the parameters.

EM in HMM (Baum-Welch):

1. E-step - computing γ_t and $\xi_{t,t+1}$ values using previous parameters
2. M-step - update parameters (separately), e.g. for Categorical observation model:
    - $\hat{\pi}_i \propto \gamma_1(i)$
    - $\hat{A}_{kj} \propto \sum_{t=2}^{T} \xi_{t-1,t}(k,j)$
    - $\hat{B}_{kl} \propto \sum_{t=1}^{T} \gamma_t(k)\mathbb{I}(x_t = l)$

Graph representation of distributions

Representation graph: the Markov-chain of the graph has a distribution T on the first arrival to the ending (absorption) state.

The geometric family $Geo(p)$ has the following representation:

- Nodes: r, v_1, s
- Edges:
  - $p(v_1|r) = 1$
  - $p(v_1|v_1) = p$
  - $p(s|v_1) = 1 - p$

Results on representation graphs

Representative families:

- geometric family with parameter p
- negative binomial family of fixed order N with parameter p
- categorical family on $\{1, \ldots, D\}$
- mixture of representative families

Non-representative distributions:

- light-tailed (lighter than exponential) distributions
- truncated Poisson distribution

Representing negative binomials

Example: consider the representation graph $G(p)$ with nodes r, v_1, v_2, v_3, s and with the following non-zero probabilities:

- $p(v_1|r) = 1$
- $p(v_1|v_1) = p$
- $p(v_2|v_1) = 1 - p$
- $p(v_2|v_2) = p$
- $p(v_3|v_2) = 1 - p$
- $p(v_3|v_3) = p$
- $p(s|v_3) = 1 - p$

It is not hard to see, that G represents the family of negative binomial distributions of fixed order 3.

Residential time HMM

RT-HMM is HMM variant with counter states representing the residential process in each state. With maximum duration D and number of hidden states M the efficiency forward-backward variant takes $\mathcal{O}((M^2 + MD)T)$ time.

With representation graphs we can generalize, extend the concept of RT-HMM while maintaining the efficiency to the categorical case, meanwhile giving lower bound on the efficiency.

RT-HMM from representation graphs

We can construct RT-HMM (as large HMM model) from the representation graphs by connecting the entry and exit points properly.

Two level of representation:

- lower level representation: i_d, Markov-chain
- higher level representation: $i \leftrightarrow \{i_1, \ldots, i_{D_i}\}$, original hidden states

Learning the parameters of RT-HMM

By modifying the Baum-Welch algorithm to accept special parameter tyings, it possible to learn the parameters.

Time complexity of steps remains (as in simple HMM):

- E-step: $\mathcal{O}(T\#\{\text{non-zero edges}\})$ (including forward-backward)
- M-step: $\mathcal{O}(T\#\{\text{parameters}\})$

Efficiency of representation

Efficiency is defined as the number of non-zero edges in RT-HMM. We can measure the efficiency of representing distribution families:

$$E(G, \{T(\theta)\})(M) = Me(G) + M(M-1)e_{out}(G)e_{in}(G)$$

where

- M is the number of hidden states
- $e_{in} \doteq |\{i : p(v_i|r) \not\equiv 0\}|$ the number of incoming edges
- $e_{out} \doteq |\{i : p(s|v_i) \not\equiv 0\}|$ the number of outgoing edges
- $e \doteq |\{i, j : p(v_j|v_i) \not\equiv 0\}|$ the number of inner edges

The main result: it is possible to represent the $Cat(\{1, \ldots, D\})$ with efficiency $\mathcal{O}(MD + M^2)$, and it is optimal. It gives the same $\mathcal{O}((M^2 + MD)T)$ on the forward-backward algorithm.