Graphical-Duration HMM

László Keresztes

Applied Mathematics MSc, ELTE-TTK

Supervisor: Balázs Csanád Csáji, SZTAKI, ELTE

May 18, 2022

Hidden Markov Models

A Hidden Markov Model is a hidden process, a discrete $z_t \in \{1, \dots, N\}$ Markov chain in discrete time ($t \in \{1, \dots, T\}$), and an observation model $p(x_t|z_t)$. The joint distribution has the form

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

An HMM (with categorical observations) has parameters $\theta = (\pi, A, B)$.

- $\pi_i = p(z_1 = i)$ initial distribution
- $A_{ij} = p(z_t = j|z_{t-1} = i)$ transition probabilities
- $B_{il} = p(x_t = l|z_t = i)$ emission probabilities

Hidden Markov Models

Given an HMM $\theta = (\pi, A, B)$ and observation sequence $x_{1:T}$.
Inference and learning (E-step):

- $\alpha_t(i) = p(z_t = i | x_{1:t})$ (forwards alg.)
- $\beta_t(j) = p(x_{t+1:T} | z_t = j)$ (backwards alg.)
- $\gamma_t(i) = p(z_t = i | x_{1:T}) \propto \alpha_t(i) \beta_t(i)$
- $\xi_{t,t+1}(i,j) = p(z_t = i, z_{t+1} = j | x_{1:T}) \propto \alpha_t(i) A_{ij} \beta_{t+1}(j) B_{j,x_{t+1}}$

Time complexity (altogether):

- $\mathcal{O}(TM^2)$
- $\mathcal{O}(TE)$ in a sparse graph with $E \ll M^2$

EM learning

Expectation-Maximization algorithm increases the likelihood and finds a local optima when exact maximum likelihood estimation is not possible. Complete data log likelihood:

$$l_c(\theta) = \log p(x_{1:T}, z_{1:T}|\theta)$$

Auxiliary function:

$$Q(\theta; \theta^{n-1}) = E_{z_{1:T} \sim p(z_{1:T}|x_{1:T}, \theta^{n-1})}\left[l_c(\theta)\right]$$

EM (using initial parameters θ^0):

1. E-step: compute $Q(\theta; \theta^{n-1})$
2. M-step:

$$\theta^n = \arg\max_{\theta} Q(\theta; \theta^{n-1})$$

EM learning

EM in HMM (Baum-Welch):

1. E-step - compute γ_t and $\xi_{t,t+1}$ values in the θ^{n-1} HMM
2. M-step - update parameters:
   - $\pi_i^n \propto \gamma_1(i)$
   - $A_{ij}^n \propto \sum_{t=2}^{T} \xi_{t-1,t}(i,j)$
   - $B_{il}^n \propto \sum_{t=1}^{T} \gamma_t(i)\mathbb{I}(x_t = l)$

Time complexity of Baum-Welch:

- $\mathcal{O}(TM^2)$
- $\mathcal{O}(TE)$ in a sparse graph with $E \ll M^2$

Graph representation of distributions

Representing duration distributions with graphs: the distribution of the first arrival to the ending (absorption) state in the graph (Markov chain).

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

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

Graph representation of distributions

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 distributions (including truncated Poisson distribution)

(All proved.)

Graphical-Duration Hidden Markov Model

HSMMs have counter states representing the residential process in each state and a maximum duration parameter D. Time complexity of forwards-backwards (E-step) $\mathcal{O}((M^2 + MD)T)$ (most efficient implementation).

HSMMs in general consider only categorical distributions on $\{1, \ldots, D\}$.

GD-HMM extends the concept to other families while maintaining the efficiency to the categorical case. With representation graphs, we could give a lower bound on the time complexity.

Graphical-Duration Hidden Markov Model

Let (π, A, θ_o) be an HMM model with M hidden states, and T_i is a duration distribution, represented with $G_i(\eta_i)$ $\forall i = 1, \ldots, M$.
The GD-HMM is a $(\tilde{\pi}, \tilde{A}, \tilde{\theta}_o)$ HMM model.
For $i = 1, \ldots, M$:

- hidden states: $i_d \in V_{inn}(G_i)$ for $d = 1, \ldots, D_i$
- transition probabilities
  - $\tilde{A}(i_k, i_l) \doteq p_{G_i(\eta_i)}(i_l|i_k)$ for $k, l = 1, \ldots, D_i$
  - $\tilde{A}(i_k, j_l) \doteq p_{G_i(\eta_i)}(s_i|i_k) A_{ij} p_{G_j(\eta_j)}(j_l|r_j)$ for $k = 1, \ldots, D_i$ for $l = 1, \ldots, D_j$ for $j \neq i$
- initial distribution $\tilde{\pi}(i_1) = \pi(i)$, $\tilde{\pi}(i_k) = 0$ for $k = 2, \ldots, D_i$
- observation model parameters $\tilde{\theta}_o(i_k) = \theta_o(i)$ for $k = 1, \ldots, D_i$

The parameters of the GD-HMM are $(\pi, A, \theta_o, (\eta_1, \ldots, \eta_M))$.

Graphical-Duration Hidden Markov Model

Two levels of representation:

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

The number of (non-zero) edges in a GD-HMM is:

$$E = \sum_{i=1}^{M} e^i + \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} e^i_{out} e^j_{in} \mathbb{I}(A_{ij} > 0)$$

- $e^i_{in} = e_{in}(G_i)$ the number of incoming edges
- $e^i_{out} = e_{out}(G_i)$ the number of outgoing edges
- $e^i = e(G_i)$ the number of inner edges

Learning the parameters of GD-HMM

Baum-Welch E-step applicable.

Baum-Welch M-step (complex modification):

- write $Q(\theta; \theta^n)$ as a function of parameters $(\tilde{\pi}, \tilde{A}, \tilde{\theta}_o)$
- rewrite it as a function of parameters $(\pi, A, \theta_o, (\eta_1, \ldots, \eta_M))$ using the definition GD-HMM
- group the terms to have a sum of $\sum_{i \in I} a_i \log p_i$ terms, where $(p_i : i \in I)$ is a probability distribution and $a_i \geq 0 \; \forall i \in I$
- compute a_i coefficients

Time complexity remains $\mathcal{O}(TE)$.

Efficiency of representation in GD-HMM

The number of (non-zero) edges is the key measure of the time complexity of the forwards-backwards algorithm (and EM algorithm) in any HMM.

We have an efficient representation if the final graph uses less edges (complete original HMM, same distribution family, same representation graph):

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

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

Efficiency of representation in GD-HMM

Állítás

(Optimal representation of categorical distributions)
Let $\{T(\theta) : \theta \in \Theta\}$ *is the family of categorical distributions on* $\{1, \ldots, D\}$, *with* $\theta = (p_1, \ldots, p_D)$. *Let* G *represents this family. Then*

$$E(G, \{T(\theta)\})(M) \geq M(D-1) + M(M-1)$$

- For the categorical distribution, a representation graph G has $E = M(2D-3) + M(M-1)2 = \mathcal{O}(MD + M^2)$ edges.
- The time complexity (of 1 iteration of EM) with this graph is $\mathcal{O}(T(MD + M^2))$ (same as the most efficient implementation of HSMM)
- No better time complexity could be achieved with representation graphs.