A Spectral Algorithm for Learning Hidden Markov Models

By: Mohsen Nabian

Summary:

This paper provides efficient algorithms for learning hidden markov models (HMMs) from Data. HMMs are generally computationally hard to be learnt and sometimes require to use heuristics which themselves makes optimization to trap in local optimum.
However this paper with the algorithm proposed, with the constraint on the problems to be in "natural separation condition" learning procedure is simple. it employs only a singular value decomposition and matrix multiplications.
Also the complexity of the algorithm implicitly depends number of observation through spectral properties of underlying HMM.

Hidden Markov Model

HMM defined as a sequences of hidden states (h_t) and observations (x_t).

~~hidden states: $h_1, h_2, \dots h_t$~~

Set of hidden states : $[m] = \{1, 2, \dots, m\}$ $\quad n \geq m$

" " observations : $[n] = \{1, 2, \dots, n\}$

$$T_{ij} = Pr\left[h_{t+1} = i \mid h_t = j\right] \qquad T \in \mathbb{R}^{m \times m}$$

$$O_{ij} = Pr\left[x_t = i \mid h_t = j\right] \qquad O \in \mathbb{R}^{n \times n}$$

HMM satisfies conditional independence properties.

$\vec{\pi}$ is the initial state distribution $\vec{\pi} \in \mathbb{R}^m$

with $\vec{\pi}_i = Pr[h_1 = i]$

Lemma 1) : For $x = 1, 2, \dots, n$ define

$$A_x = T\,\text{diag}\,(O_{x,1}, \dots, O_{x,m}) \qquad \text{ex: } A_1 = T\,\text{dig}\,(O_{11}, O_{12} \dots O_{1m})$$

for any t :

$$Pr[x_1, \dots, x_t] = \underbrace{\vec{1}_m^{\,T}}_{\substack{\text{all ones} \\ \text{vector } \in \mathbb{R}^m}} A_{x_t} \cdots A_{x_1} \vec{\pi}$$

Assumption :

Condition 1: $\vec{\pi} > 0$ and $\vec{O}$ and $\vec{T}$ are rank m.

Learning Model:

The goal is to derive cumulative distribution $\Pr[x_{1:t}]$ and conditional distribution $\Pr[x_t \mid x_{1:t-1}]$ for any sequence length t.

definitions :

$[P_1]_i = \Pr[x_1 = i]$

$[P_{2,1}]_{ij} = \Pr[x_2 = i, x_1 = j]$

$[P_{3,x,1}]_{ij} = \Pr[x_3 = i, x_2 = x, x_1 = j] \quad \forall x \in [n]$

where $P_1 \in \mathbb{R}^n$ is a vector, $P_{2,1} \in \mathbb{R}^{n \times n}$ and $P_{3,x,1} \in \mathbb{R}^{n \times n}$ are matrices.

Also we define $U \in \mathbb{R}^{n \times m}$ so that $U^T O$ is invertable.
an exaple of U is 'thin' SVD of $P_{2,1}$.

Condition 2

$$[P_{21}]_{ij} = \sum_{k=1}^{m} \sum_{l=1}^{m} Pr[x_2 = i, x_1 = j, h_2 = k, h_1 = l]$$

$$= \sum_{k=1}^{m} \sum_{l=1}^{m} O_{ik} T_{kl} \vec{\pi}_l [O^T]_{lj}$$

equality by Lemma 2

or in general:

$$P_{2,1} = \cancel{OT}\ O T \,\mathrm{diag}(\vec{\pi})\, O^T$$

$$O = P_{2,1} (T \,\mathrm{diag}(\vec{\pi})\, O^T)^{+}$$

X^{+} denots the Moore-Penrose pseduo-inverse of Matrix X

So: $$\begin{cases} \mathrm{range}(O) \subseteq \mathrm{range}(P_{2,1}) \\ \mathrm{rank}(P_{2,1}) = \mathrm{rank}(O) = m \\ \mathrm{range}(U) = \mathrm{range}(P_{2,1}) = \mathrm{range}(O) \end{cases}$$

So in each iteration:

→ Compute SVD($P_{2,1}$) → discover U that satisfies Condition 2

→ we define b, B

$$\begin{cases} \vec{b}_1 = U^T P_1 \\ \vec{b}_\infty = (P_{2,1}^T U)^{+} P_1 \\ B_x = (U^T P_{3,x,1})(U^T P_{2,1})^{+} \quad \forall x \in [n] \end{cases}$$

1) Spectral Learning of Hidden Markov Models:

Algorithm:

we want to predict probability of a sequence:

$$\Pr[x_1, \dots, x_t] = \hat{b}_\infty^T \hat{B}_{x_t} \cdots \hat{B}_{x_1} \hat{b}_1 .$$

given observation x_t, the 'internal state' update is:

$$\hat{b}_{t+1} = \frac{\hat{B}_{x_t} \hat{b}_t}{\hat{b}_\infty^T \hat{B}_{x_t} \hat{b}_t}$$

So Algorithm: Learn HMM(m, N)

input: m - number of states, N - sample size

Return: HMM model parametrized by $\{\hat{b}_1, \hat{b}_\infty, \hat{B}_x \ \forall x \in [n]\}$

① Independently sample N observation triples (x_1, x_2, x_3) from HMM to form emperical estimates:

$\hat{P}_1, \hat{P}_{2,1}, \hat{P}_{3,x,1} \ \forall x \in [n]$ of $P_1, P_{2,1}, P_{3,x,1} \ \forall x \in [n]$.

② Compare the SVD of $\hat{P}_{2,1}$ and let $\hat{U}$ be the matrix of left singular vectors correspondig to the m largest singular values

(3) Compute model parameters:

(a) $\hat{b}_1 = \hat{U}^T \hat{P}_1$

(b) $\hat{b}_\infty = (\hat{P}_{2,1}^T \hat{U})^+ P_1$

(c) $\hat{B}_x = \hat{U}^T P_{3,x,1} (\hat{U}^T \hat{P}_{2,1})^+ \quad \forall x \in [n]$

Now to predict conditional Probability:

$$\hat{P}_r [x_t | x_{1:t-1}] = \frac{\hat{b}_\infty^T \hat{B}_{x_t} \hat{b}_t}{\sum_x \hat{b}_\infty^T \hat{B}_x \hat{b}_t}$$

↳ Conditional Probabity

and remineling

$$\hat{P}_r [x_1, \ldots, x_t] = \hat{b}_\infty^T \hat{B}_{x_t} \ldots \hat{B}_{x_1} \hat{b}_1$$

↳ joint probability

and at the end the author provides ~~the~~ some upper bounds for the Learning error (~~or~~ accuracy) ε for the case of Joint probability and Conditional Probability