A Spectral Algorithm for Learning Hidden Markov Models

Based on Daniel Hsu, Sham Kakade, and Tong Zhang
arXiv:0811.4413

Kentaro Hanaki

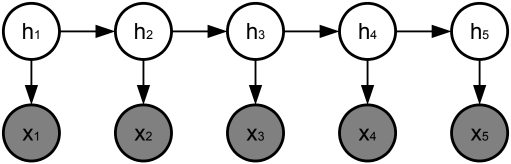
New York University

May 5, 2015

Table of Contents

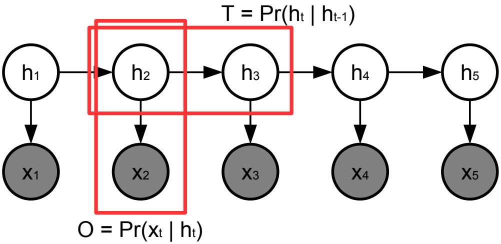
Table of Contents

What is HMM?



- Probabilistic model for sequential observations $\Pr[x_1, \cdots, x_T]$
- Observations are generated by their underlying hidden states
- Hidden states follows Markov assumptions
- Low-dim. hidden states $\rightarrow$ Dynamics easier to understand
- Speech recognition, NLP (PoS tagging, NER, MT, ...), etc.

Parameters of HMM



- $\Pr[x_1, \cdots, x_T] = \sum_{h_1,\cdots,h_T} (\prod_{t=1}^{T} \Pr[h_t|h_{t-1}] \Pr[x_t|h_t]) \Pr[h_1]$
- Observation matrix $O_{i,a} = \Pr[x_t = i | h_t = a]$
- Transition matrix $T_{a,b} = \Pr[h_t = a | h_{t-1} = b]$
- Prior probability for the initial hidden states $\pi_a = \Pr[h_t = a]$

Training HMM — Traditional Approaches

Natural loss for probabilistic model is negative log likelihood (NLL)

1. Find the global minimum of NLL
   - NLL is not convex due to the presence of hidden states
   - Solving non-convex optimization problem is extremely hard
2. Heuristically minimize NLL using EM algorithm
   - NLL is convex once the hidden states are fixed
   - E-step: Infer and fix the hidden states based on the parameters
   - M-step: Minimize NLL with respect to parameters
   - Tend to get stuck into local minima
   - Works fine in practice, hard to analyze learning guarantees

Contribution of the Paper

The paper proposed an efficient training method that admits a unique solution and learning guarantees

Table of Contents

Overview of the Algorithm

- Learn model for predicting joint probabilities for observations
- Learning is hard due to the presence of hidden states
- Two steps solution without directly referring to hidden states:
  - Map to observation space so that probs can be estimated
  - Find subspace that is tractable

Observation Operator Representation

In HMM, joint probability can be written as

$$\Pr[x_1, \cdots, x_t] = \vec{1}_m^T A_{x_t} \cdots A_{x_1} \vec{\pi}$$

A_x is the observation operator

$$A_x = T \mathrm{diag}(O_{x,1}, \cdots, O_{x,m})$$

Proof

Proof for $t = 2$ (Generalization is straightforward)

$$
\begin{aligned}
\Pr[x_1 = i, x_2 = j] &= \sum_{a,b} \Pr[x_1 = i, x_2 = j | h_1 = a, h_2 = b] \Pr[h_1 = a, h_2 = b] \\
&= \sum_{a,b} \Pr[x_1 = i, x_2 = j | h_1 = a, h_2 = b] T_{b,a} \pi_a \\
&= \sum_{a,b} \Pr[x_1 = i | h_1 = a] \Pr[x_2 = j | h_2 = b] T_{b,a} \pi_a \\
&= \sum_{a,b} O_{j,b} T_{b,a} O_{i,a} \pi_a \\
&= \vec{1}_m^T T \operatorname{diag}(O_{j,1}, \cdots, O_{j,m}) T \operatorname{diag}(O_{i,1}, \cdots, O_{i,m}) \vec{\pi} \\
&= \vec{1}_m^T A_j A_i \vec{\pi}
\end{aligned}
$$

Observation Operator Representation

Problem: Estimation of $\vec{\pi}$ and A_x requires inferring hidden states

- $\vec{\pi}$: Prior for hidden states
- A_x: Transition between hidden states

Solution: Map everything into the space in which

- each component can be estimated from observation
- dynamics as easy as in hidden state space

Mapping to Subspace

- First map everything into observation space using O

$$\vec{\pi} \to O\vec{\pi}, \quad \vec{1}_m \to O\vec{1}_m, \quad A_x \to OA_xO^{-1}$$

- O^{-1} not well-defined and dynamics complicated in this space
- Find U such that U^TO is invertible (i.e., preserves hidden state dynamics) and

$$\vec{\pi} \to \vec{b}_1 = (U^TO)\vec{\pi}, \quad \vec{1}_m \to \vec{b}_\infty = (U^TO)\vec{1}_m$$

$$A_x \to B_x = (U^TO)A_x(U^TO)^{-1}$$

$$\Pr[x_1, \cdots, x_t] = \vec{b}_\infty^T B_{x_t} \cdots B_{x_1} \vec{b}_1$$

Estimating b_1, b_∞ and B

b_1, b_∞ and B_x can be estimated using

$$
\begin{aligned}
(P_1)_i &= \Pr[x_1 = i] \\
(P_{2,1})_{i,j} &= \Pr[x_2 = i, x_1 = j] \\
(P_{3,x,1})_{i,j} &= \Pr[x_3 = i, x_2 = x, x_1 = j]
\end{aligned}
$$

Lemma

b_1, b_∞, B_x can be expressed as

$$
\begin{aligned}
b_1 &= U^T P_1 \\
b_\infty &= (P_{2,1}^T U)^+ P_1 \\
B_x &= (U^T P_{3,x,1})(U^T P_{2,1})^+
\end{aligned}
$$

Finding U

Lemma

Assume $\vec{\pi} > 0$ and that O and T have column rank m (i.e., any two hidden states are not identical). If U is the matrix of left singular vectors of $P_{2,1}$ corresponding to non-zero singular values, then $U^T O$ is invertible.

Proof for the Existence of $(U^T O)^{-1}$

$P_{2,1}$ can be rewritten as

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

So, $\mathrm{rank}(P_{2,1}) \leq \mathrm{rank}(O)$. Also, $T\mathrm{diag}(\vec{\pi})O^T$ has full row rank from assumptions. This implies

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

Therefore, $\mathrm{rank}(O) \subset \mathrm{rank}(P_{2,1})$. So,

$$\mathrm{rank}(O) = \mathrm{rank}(P_{2,1}) = \mathrm{rank}(U)$$

and $U^T O$ has rank m and invertible.

Learning Algorithm

1. Randomly sample N observation triples (x_1, x_2, x_3) and estimate P_1, $P_{2,1}$ and $P_{3,x,1}$
2. Compute SVD of $P_{2,1}$ and let U be the matrix consisting of left singular vectors corresponding to m largest singular values
3. Compute model parameters b_1, b_∞ and B_x

Table of Contents

Learning Bound for Joint Probabilities

Theorem

There exists a constant $C > 0$ s.t.

$$\Pr\left[\sum_{x_1,\cdots,x_t} |\Pr[x_1,\cdots,x_t] - \hat{\Pr}[x_1,\cdots,x_t]| \geq \epsilon\right] \leq \exp\left(-\frac{\sigma_m(O)^2\sigma_m(P_{2,1})^4 N\epsilon^2}{C(1+n_0(\epsilon)\sigma_m(P_{2,1})^2)t^2}\right)$$

where

$$\begin{aligned} n_0(\epsilon) &= \min\{k : \epsilon(k) \leq \epsilon/4\} \\ \epsilon(k) &= \min\left\{\sum_{j\in S} \Pr[x_2 = j] : S \subset [n], |S| = n-k\right\} \end{aligned}$$

and σ_m is the m-th largest singular value.

Note: Bound gets looser as the length of sequence gets longer!!!

Brief Idea of the Proof

1. Compute the sampling error for P_1, $P_{2,1}$ and $P_{3,x,1}$
2. Compute how the sampling error is propagated to accuracy
   - Compute the approximation error for U
   - Compute the approximation error for b_1, b_∞ and B_x

Learning Bound for Conditional Probabilities

Conditional probability

$$\Pr[x_T | x_{T-1}, \cdots, x_1] = \frac{b_\infty^T B_{x_T} b_T}{\sum_x b_\infty^T B_x b_T}$$

$$b_{t+1} = \frac{B_{x_{T+1}} b_T}{b_\infty^T B_{x_T} b_T}$$

Conditional probability also have learning guarantee

- Bound independent of the length of the sequence
- Two more parameters compared to joint probability bound
  - α: Smallest value of transition matrix A_x
  - γ: Error for hidden states measured in observation space

Table of Contents

Conclusion

- Spectral learning for HMM yields a unique solution
- Joint probability estimated without inferring hidden states
- Learning guarantee for joint probability depends on the length, but that for conditional probability does not