Introduction to Hidden Markov Models

Antonio Artés-Rodríguez
Unviersidad Carlos III de Madrid

2nd MLPM SS, September 17, 2014

Outline

Section 1

Markov and Hidden Markov Models

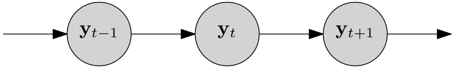
Markov processes

Joint distribution of a sequence $\mathbf{y}_{1:T}$

$$p(\mathbf{y}_{1:T}) = p(\mathbf{y}_1)p(\mathbf{y}_2|\mathbf{y}_1)\dots p(\mathbf{y}_t|\mathbf{y}_{1:t-1})\dots p(\mathbf{y}_T|\mathbf{y}_{1:T-1})$$

- First order Markov process

$$p(\mathbf{y}_{1:T}) = p(\mathbf{y}_1)p(\mathbf{y}_2|\mathbf{y}_1)\dots p(\mathbf{y}_t|\mathbf{y}_{t-1})\dots p(\mathbf{y}_T|\mathbf{y}_{T-1})$$



  - Second order Markov process

$$p(\mathbf{y}_{1:T}) = p(\mathbf{y}_1)p(\mathbf{y}_2|\mathbf{y}_1)\dots p(\mathbf{y}_t|\mathbf{y}_{t-1},\mathbf{y}_{t-2})\dots p(\mathbf{y}_T|\mathbf{y}_{T-1},\mathbf{y}_{T-2})$$

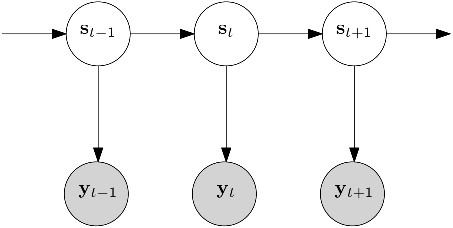
- First order homogeneous Markov process

$$p(\mathbf{y}_2|\mathbf{y}_1) = \cdots = p(\mathbf{y}_t|\mathbf{y}_{t-1}) = \cdots = p(\mathbf{y}_T|\mathbf{y}_{T-1})$$

Hidden Markov processes

If the observed sequence $\mathbf{y}_{1:T}$ is a noisy version of the (first order) Markov process $\mathbf{s}_{1:T}$

$$p(\mathbf{y}_{1:T}, \mathbf{s}_{1:T}) = p(\mathbf{y}_1|\mathbf{s}_1)p(\mathbf{s}_1)\dots p(\mathbf{y}_t|\mathbf{s}_t)p(\mathbf{s}_t|\mathbf{s}_{t-1})\dots$$
$$\dots p(\mathbf{y}_T|\mathbf{s}_T)p(\mathbf{s}_T|\mathbf{s}_{T-1})$$



- Discrete s_t: Hidden Markov Model (HMM)
- Continuous $\mathbf{s}_t$: State Space Model (SSM)
    - e.g. AR models

Coin Toss Example

(from [Rabiner and Juang, 1986])

- The result of tossing one-or-multiple fair-or-biased coins is

$$y_{1:T} = \mathtt{hhtttthtth} \cdots \mathtt{h}$$

- Possible models:
  - 1-coin model (not hidden):

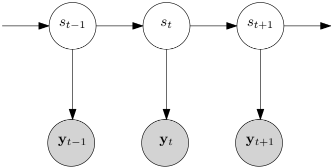
$$p(y_t = \mathtt{h} | y_{t-1} = \mathtt{h}) = p(y_t = \mathtt{h} | y_{t-1} = \mathtt{t}) = \\ 1 - p(y_t = \mathtt{t} | y_{t-1} = \mathtt{h}) = 1 - p(y_t = \mathtt{t} | y_{t-1} = \mathtt{t})$$

  - 2-coin model:

$$\begin{aligned} p(y_t = \mathtt{h} | s_t = 1) &= p_1 & \quad p(y_t = \mathtt{t} | s_t = 1) &= 1 - p_1 \\ p(y_t = \mathtt{h} | s_t = 2) &= p_2 & \quad p(y_t = \mathtt{t} | s_t = 2) &= 1 - p_2 \\ p(s_t = 1 | s_{t-1} = 1) &= a_{11} & \quad p(s_t = 2 | s_{t-1} = 1) &= a_{12} \\ p(s_t = 1 | s_{t-1} = 2) &= a_{21} & \quad p(s_t = 2 | s_{t-1} = 2) &= a_{22} \end{aligned}$$

  - ...

The model



- $S = \{s_1, s_2, \ldots, s_T : s_t \in 1, \ldots, I\}$: hidden state sequence.
- $Y = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T : \mathbf{y}_t \in \mathbb{R}^M\}$: observed continuous sequence
- $\mathbf{A} = \{a_{ij} : a_{ij} = P(s_{t+1} = j | s_t = i)\}$: state transition probabilities.
- $\mathbf{B} = \{b_i : P_{b_i}(\mathbf{y}_t) = P(\mathbf{y}_t | s_t = i)\}$: observation emission probabilities.
- $\boldsymbol{\pi} = \{\pi_i : \pi_i = P(s_1 = i)\}$: initial state probability distribution.
- $\theta = \{\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}\}$: model parameters.

Applications of HMMs

- Automatic speech recognition
  - s corresponds to phonemes or words and $\mathbf{y}$ to features extracted from the speech signal
- Activity recognition
  - s corresponds to activities or gestures and $\mathbf{y}$ to features extracted from video or sensors signals
- Gene finding
  - s corresponds to the location of the gene and $\mathbf{y}$ to DNA nucleotides
- Protein sequence alignment
  - s corresponds to the matching to the latent consensus sequence and $\mathbf{y}$ to aminoacids

Section 2

Inference in HMM

Three Inference Problems for HMMs

Problem 1: Given Y and θ, determine $p(Y|\theta)$.

$$p(Y|\theta) = \sum_S p(Y, S|\theta) \qquad \mathcal{O}(I^T)$$

- $p(Y|\theta) = \sum_{s_T} p(Y, s_T|\theta)$ ($\mathcal{O}(I^2 T)$) (Forward algorithm)

Problem 2: Given Y and θ, determine the "optimal" S.

- $p(s_t|Y, \theta)$ ($\mathcal{O}(I^2 T)$) (Forward-Backward algorithm)
- $\underset{S}{\operatorname{argmax}}\, p(Y|S, \theta)$ ($\mathcal{O}(I^2 T)$) (Viterbi algorithm)

Problem 3: Determine θ to maximize $p(Y|\theta)$.

Forward-Backward Algorithm

$$
\begin{aligned}
P(s_t = i | Y) = \gamma_t(i) &= \frac{P(Y, s_t = i)}{P(Y)} \\
&= \frac{P(\mathbf{y}_{t+1:T} | s_t = i) P(\mathbf{y}_{1:t}, s_t = i)}{P(Y)} \\
&= \frac{\beta_t(i) \alpha_t(i)}{P(Y)}
\end{aligned}
$$

- Forward:
  - $\alpha_1(i) = \pi_i P_{b_i}(\mathbf{y}_1) \qquad 1 \le i \le I$
  - $\alpha_t(i) = \left(\sum_{j=1}^{I} \alpha_{t-1}(j) a_{ji} \right) P_{b_i}(\mathbf{y}_t) \qquad 1 \le i \le I,\ 1 < t \le T$

Forward-Backward Algorithm

$$
\begin{aligned}
P(s_t = i|Y) = \gamma_t(i) &= \frac{P(Y, s_t = i)}{P(Y)} \\
&= \frac{P(\mathbf{y}_{t+1:T}|s_t = i)P(\mathbf{y}_{1:t}, s_t = i)}{P(Y)} \\
&= \frac{\beta_t(i)\alpha_t(i)}{P(Y)}
\end{aligned}
$$

- Forward:
  - $\alpha_1(i) = \pi_i P_{b_i}(\mathbf{y}_1)$ $\qquad 1 \le i \le I$
  - $\alpha_t(i) = \left(\sum_{j=1}^{I} \alpha_{t-1}(j) a_{ji}\right) P_{b_i}(\mathbf{y}_t)$ $\qquad 1 \le i \le I,\ 1 < t \le T$
- Backward:
  - $\beta_T(i) = 1$ $\qquad 1 \le i \le I$
  - $\beta_t(i) = \sum_{j=1}^{I} a_{ij} P_{b_j}(\mathbf{y}_{t+1})\beta_{t+1}(j)$ $\qquad 1 \le i \le I,\ 1 \le t < T$

Third Inference Problem

Joint distribution of S and Y and log-likelihood for N sequences

$$p(S, Y) = \prod_{n=1}^{N} \left(p(s_1^n) \prod_{t=2}^{T_n} p(s_t^n | s_{t-1}^n) \right) \left(\prod_{t=1}^{T_n} p(\mathbf{y}_t^n | s_t^n) \right)$$

- EM (Baum-Welch) [Baum et al., 1970]
- Bayesian inference methods:
  - Gibbs sampler [Robert et al., 1993]
  - Variational Bayes [MacKay, 1997]

Baum-Welch (EM) Algorithm

Joint distribution of S and Y and log-likelihood for N sequences

$$p(S, Y) = \prod_{n=1}^{N} \left(p(s_1^n) \prod_{t=2}^{T_n} p(s_t^n | s_{t-1}^n) \right) \left(\prod_{t=1}^{T_n} p(\mathbf{y}_t^n | s_t^n) \right)$$

$$\log p(S, Y|\theta) = \sum_{n=1}^{N} \Bigg(\sum_{i=1}^{I} I(s_1^n = i|Y, \theta) \log \pi_i +$$

$$\sum_{t=2}^{T_n} \sum_{i=1}^{I} \sum_{j=1}^{I} I(s_{t-1}^n = i, s_t^n = j|Y, \theta) \log a_{ij} + \sum_{t=1}^{T_n} \sum_{i=1}^{I} I(s_t^n = i|Y, \theta) \log p(\mathbf{y}_t^n | b_i) \Bigg)$$

$$= \sum_{i=1}^{I} \left(\sum_{n=1}^{N} I(s_1^n = i|Y, \theta) \right) \log \pi_i$$

$$+ \sum_{i=1}^{I} \sum_{j=1}^{I} \left(\sum_{n=1}^{N} \sum_{t=2}^{T_n} I(s_{t-1}^n = i, s_t^n = j|Y, \theta) \right) \log a_{ij}$$

$$+ \sum_{i=1}^{I} \left(\sum_{n=1}^{N} \sum_{t=1}^{T_n} I(s_t^n = i|Y, \theta) \right) \log p(\mathbf{y}_t^n | b_i)$$

Baum-Welch (EM) Algorithm (II)

$$\log p(S, Y|\theta) = \sum_{i=1}^{I}\left(\sum_{n=1}^{N} I(s_1^n = i|Y,\theta)\right)\log \pi_i$$
$$+ \sum_{i=1}^{I}\sum_{j=1}^{I}\left(\sum_{n=1}^{N}\sum_{t=2}^{T_n} I(s_{t-1}^n = i, s_t^n = j|Y,\theta)\right)\log a_{ij}$$
$$+ \sum_{i=1}^{I}\left(\sum_{n=1}^{N}\sum_{t=1}^{T_n} I(s_t^n = i|Y,\theta)\right)\log p(\mathbf{y}_t^n|b_i)$$

E step

- $E\left(\sum_{n=1}^{N} I(s_1^n = i|Y,\theta)\right) = \sum_{n=1}^{N}\gamma_{n,1}(i)$
- $E\left(\sum_{n=1}^{N}\sum_{t=2}^{T_n} I(s_{t-1}^n = i, s_t^n = j|Y,\theta)\right) = \sum_{n=1}^{N}\sum_{t=2}^{T_n}\xi_{n,t}(i,j)$
- $E\left(\sum_{n=1}^{N}\sum_{t=1}^{T_n} I(s_t^n = i|Y,\theta)\right) = \sum_{n=1}^{N}\sum_{t=1}^{T_n}\gamma_{n,t}(i)$

$$\xi_{n,t}(i,j) = P(s_{t-1}^n = i, s_t^n = j|Y) = \alpha_t(i)a_{ij}P_{b_j}(\mathbf{y}_{t+1})\beta_{t+1}(j)$$

Baum-Welch (EM) Algorithm (III)

$$\log p(S, Y|\theta) = \sum_{i=1}^{I} \left(\sum_{n=1}^{N} I(s_1^n = i|Y, \theta) \right) \log \pi_i$$
$$+ \sum_{i=1}^{I} \sum_{j=1}^{I} \left(\sum_{n=1}^{N} \sum_{t=2}^{T_n} I(s_{t-1}^n = i, s_t^n = j|Y, \theta) \right) \log a_{ij}$$
$$+ \sum_{i=1}^{I} \left(\sum_{n=1}^{N} \sum_{t=1}^{T_n} I(s_t^n = i|Y, \theta) \right) \log p(\mathbf{y}_t^n|b_i)$$

M step

- $\hat{\pi}_i = \left(\sum_{n=1}^{N} \gamma_{n,1}(i) \right) / N$
- $\hat{a}_{ij} = \left(\sum_{n=1}^{N} \sum_{t=2}^{T_n} \xi_{n,t}(i,j) \right) / \left(\sum_{j=1}^{I} \sum_{n=1}^{N} \sum_{t=2}^{T_n} \xi_{n,t}(i,j) \right)$
- Gaussian emission probabilities:
  - $\hat{\boldsymbol{\mu}}_i = \left(\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma_{n,t}(i)\, \mathbf{y}_t^n \right) / \left(\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma_{n,t}(i) \right)$
  - $\hat{\boldsymbol{\Sigma}}_i = \dfrac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma_{n,t}(i)\, \mathbf{y}_t^n \mathbf{y}_t^{n*} - \sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma_{n,t}(i)\, \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{\mu}}_i^*}{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma_{n,t}(i)}$

Bayesian Inference Methods for HMM

- Priors:
  - Independent Dirichlet distributions on the rows of $\mathbf{A}$, $\mathbf{a}_i = [a_{i1} \cdots a_{iI}]$
  - If possible, conjugate priors on emission probability parameters: Dirichlet for discrete observations, Normal-Invert Wishart for Gaussian observations, ...

Bayesian Inference Methods for HMM

- Priors:
  - Independent Dirichlet distributions on the rows of $\mathbf{A}$, $\mathbf{a}_i = [a_{i1} \cdots a_{iI}]$
  - If possible, conjugate priors on emission probability parameters: Dirichlet for discrete observations, Normal-Invert Wishart for Gaussian observations, ...
- Inference methods
  - Gibbs sampler: iterative sampling from $\{p(s_t|Y, S_{-t}, \theta): \ t = 1, \ldots, T\}$, $p(\mathbf{A}|S)$, $p(\mathbf{B}|Y, S)$, $p(\boldsymbol{\pi}|S)$
    - Samples from $\{p(s_t|Y, S_{-t}, \theta): \ t = 1, \ldots, T\}$ can be efficiently generated using the Forward-Filtering Backward-Sampling (FF-BS) algorithm [Frühwirth-Schnatter, 2006]
  - Variational Bayes: maximization of the Evidence Lower BOund (ELBO) obtained by assuming independence among S, $\mathbf{A}$, $\mathbf{B}$, and $\boldsymbol{\pi}$

Section 3

Variations on HMMs

From Gaussian to Mixture of Gaussian Emission Probabilities

$$\log p(\mathbf{y}_t^n|b_i) = \log \prod_{k=1}^{K} N(\mathbf{y}_t^n|\boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik})^{z_t^n} = \sum_{k=1}^{K} z_t^n \log N(\mathbf{y}_t^n|\boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik})$$

$$\sum_{i=1}^{I} \left(\sum_{n=1}^{N} \sum_{t=1}^{T_n} I(s_t^n = i|Y, \theta) \right) \log p(\mathbf{y}_t^n|b_i) =$$

$$\sum_{i=1}^{I} \sum_{k=1}^{K} \left(\sum_{n=1}^{N} \sum_{t=1}^{T_n} I(s_t^n = i|Y, \theta) I(z_t^n = k|\mathbf{y}_t^n, \theta) \right) \log N(\mathbf{y}_t^n|\boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik})$$

E step

$$E \left(\sum_{n=1}^{N} \sum_{t=1}^{T_n} I(s_t^n = i|Y, \theta) I(z_t^n = k|\mathbf{y}_t^n, \theta) \right) \propto$$

$$\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma_{n,t}(i) c_{ik} N(\mathbf{y}_t^n|\boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}) \doteq \sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma_{n,t}(i, k)$$

From Gaussian to Mixture of Gaussian Emission Probabilities (II)

$$\sum_{i=1}^{I}\left(\sum_{n=1}^{N}\sum_{t=1}^{T_n} l(s_t^n = i|Y,\theta)\right)\log p(\mathbf{y}_t^n|b_i) =$$

$$\sum_{i=1}^{I}\sum_{k=1}^{K}\left(\sum_{n=1}^{N}\sum_{t=1}^{T_n} l(s_t^n = i|Y,\theta) l(z_t^n = k|\mathbf{y}_t^n,\theta)\right)\log N(\mathbf{y}_t^n|\boldsymbol{\mu}_{ik},\boldsymbol{\Sigma}_{ik})$$

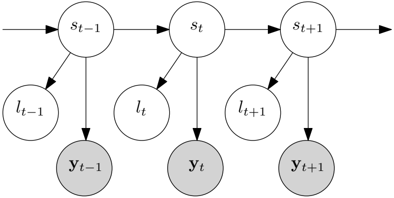
M step

$$\hat{c}_{ik} = \frac{\sum_{n=1}^{N}\sum_{t=1}^{T_n}\gamma_{n,t}(i,k)}{\sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{t=1}^{T_n}\gamma_{n,t}(i,k)}$$

$$\hat{\boldsymbol{\mu}}_{ik} = \frac{\sum_{n=1}^{N}\sum_{t=1}^{T_n}\gamma_{n,t}(i,k)\,\mathbf{y}_t^n}{\sum_{n=1}^{N}\sum_{t=1}^{T_n}\gamma_{n,t}(i,k)}$$

$$\hat{\boldsymbol{\Sigma}}_{ik} = \frac{\sum_{n=1}^{N}\sum_{t=1}^{T_n}\gamma_{n,t}(i,k)\,\mathbf{y}_t^n\mathbf{y}_t^{n*} - \sum_{n=1}^{N}\sum_{t=1}^{T_n}\gamma_{n,t}(i,k)\,\hat{\boldsymbol{\mu}}_i\hat{\boldsymbol{\mu}}_i^*}{\sum_{n=1}^{N}\sum_{t=1}^{T_n}\gamma_{n,t}(i,k)}$$

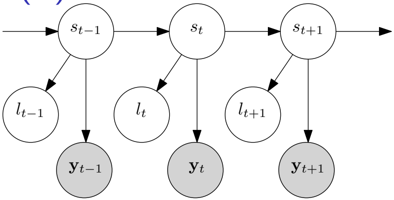
HMM with labels



- $L = \{l_1, l_2, \ldots, l_T : l_t \in 1, \ldots, J\}$: label's sequence.
- $\mathbf{D} = \{d_{im} : d_{im} = P(l_t = m | s_t = i)\}$: label emission probabilities.

$$p(S, Y, L) = \prod_{n=1}^{N} \left(p(s_1^n) \prod_{t=2}^{T_n} p(s_t^n | s_{t-1}^n) \right) \left(\prod_{t=1}^{T_n} p(\mathbf{y}_t^n | s_t^n) \right) \left(\prod_{t=1}^{T_n} p(l_t^n | s_t^n) \right)$$

HMM with labels (II)



- $L = \{l_1, l_2, \ldots, l_T : l_t \in 1, \ldots, J\}$: label's sequence.
- $\mathbf{D} = \{d_{im} : d_{im} = P(l_t = m | s_t = i)\}$: label emission probabilities.

$$\begin{aligned}
\log p(S, Y, L|\theta) = &\sum_{i=1}^{I} \left(\sum_{n=1}^{N} I(s_1^n = i | Y, L, \theta) \right) \log \pi_i \\
&+ \sum_{i=1}^{I} \sum_{j=1}^{I} \left(\sum_{n=1}^{N} \sum_{t=2}^{T_n} I(s_{t-1}^n = i, s_t^n = j | Y, L, \theta) \right) \log a_{ij} \\
&+ \sum_{i=1}^{I} \left(\sum_{n=1}^{N} \sum_{t=1}^{T_n} I(s_t^n = i | Y, L, \theta) \right) \left(\log p(\mathbf{y}_t^n | b_i) + \sum_{j=1}^{J} \log d_{ij} \right)
\end{aligned}$$

E step with labels

$$\begin{aligned}
\alpha_t(j) &= p(s_t = j|\mathbf{y}_{1:t}, l_{1:t}) = p(s_t = j|\mathbf{y}_t, \mathbf{y}_{1:t-1}, l_t, l_{1:t-1}) \\
&\propto p(\mathbf{y}_t|s_t = j)\, p(l_t|s_t = j)\, p(s_t = j|\mathbf{y}_{1:t-1}, l_{1:t-1}) \\
&= p(\mathbf{y}_t|s_t = j)\, p(l_t|s_t = j) \sum_{i=1}^{I} a_{ij}\alpha_{t-1}(i)
\end{aligned}$$

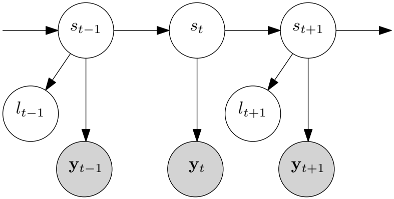
$$\begin{aligned}
\beta_{t-1}(i) &= p(\mathbf{y}_{t:T}, l_{t:T}|s_{t-1} = i) \\
&= \sum_{j=1}^{I} p(s_t = j, \mathbf{y}_t, \mathbf{y}_{t+1:T}, l_t, l_{t+1:T}|s_{t-1} = i) \\
&= \sum_{j=1}^{I} p(\mathbf{y}_{t+1:T}, l_{t+1:T}|s_t = j)\, p(s_t = j, \mathbf{y}_t, l_t|s_{t-1} = i) \\
&= \sum_{j=1}^{I} \beta_t(j)\, p(\mathbf{y}_t|s_t = j)\, p(l_t|s_t = j) a_{ij}
\end{aligned}$$

E step with labels (II)

$$\gamma_t(j) = p(s_t = j | \mathbf{y}_{1:T}, l_{1:T}) \propto p(s_t = j, \mathbf{y}_{t+1:T}, l_{t+1:T} | \mathbf{y}_{t:T}, l_{t:T}) =$$
$$= p(\mathbf{y}_{t+1:T}, l_{t+1:T} | s_t = j) p(s_t = j | \mathbf{y}_{t:T}, l_{t:T}) = \beta_t(j)\, \alpha_t(j)$$

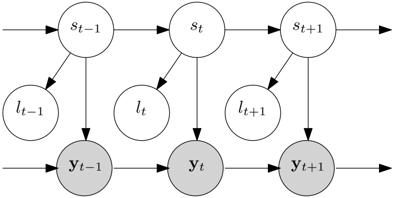
$$\begin{aligned}
\xi_{t+1}(i,j) &= p(s_t = i, s_{t+1} = j | \mathbf{y}_{1:T}, l_{1:T}) \\
&= p(s_{t+1} = j | s_t = i, \mathbf{y}_{1:T}, l_{1:T})\, p(s_t = i | \mathbf{y}_{1:T}, l_{1:T}) \\
&\propto p(\mathbf{y}_{t+1:T}, l_{t+1:T} | s_{t+1} = j)\, a_{ij}\, \alpha_t(i) \\
&= p(\mathbf{y}_{t+1}, l_{t+1} | s_{t+1} = j)\, p(\mathbf{y}_{t+2:T}, l_{t+2:T} | s_{t+1} = j)\, a_{ij}\, \alpha_t(i) \\
&= p(\mathbf{y}_{t+1} | s_{t+1} = j)\, p(l_{t+1} | s_{t+1} = j)\, \beta_{t+1}(j)\, a_{ij}\, \alpha_t(i)
\end{aligned}$$

Semi-supervised HMM



- To avoid the uncertainty in the labeling, the beginning and the end of each sequence can be let unlabeled
- The label emission probabilities are set *a priori*

Autoregressive HMM



$$p(\mathbf{y}_t|\mathbf{y}_{t-1}, s_t = i, \theta) = \sum_{k=1}^{K} c_{ik} N(\mathbf{y}_t|\mathbf{W}_i\mathbf{y}_{t-1} + \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik})$$

- E step

$$\gamma_{n,t}(i,k) = \gamma_{n,t}(i) c_{ik} N(\mathbf{y}_t^n - \mathbf{W}_i\mathbf{y}_{t-1}^n|\boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik})$$

- M step

$$\mathbf{C}_i = \frac{\sum_{n=1}^{N}\sum_{t=1}^{T_n}\sum_{k=1}^{K}\gamma_{n,t}(i,k)\left(\mathbf{y}_t^n - \boldsymbol{\mu}_{ik}\right)\left(\mathbf{y}_t^n - \boldsymbol{\mu}_{ik}\right)^*}{\sum_{n=1}^{N}\sum_{t=1}^{T_n}\sum_{k=1}^{K}\gamma_{n,t}(i,k)}$$

Other Generalizations of HMMs

- Hidden semi-Markov Model
- Input-Output HMM
- Hierarchical HMM
- Factorial HMM
- Coupled HMMs

Section 4

Extensions on classical HMM methods

Well known problems of HMM

- Model selection
  - Use your favorite complexity measure (BIC, AIC, ...) and train HMMs for different values of I
  - Infinite (Nonparametric) Hidden Markov Model [Beal et al., 2001] [Teh et al., 2006].
- Local maxima of likelihood
  - Reinitialize the algorithm several times
  - Spectral learning of HMMs [Hsu et al., 2012] [Song et al., 2010]

The Infinite Hidden Markov Model

- Bayesian HMM, discrete observation, single sequence
  - Priors

$$\mathbf{a}_i|\alpha, I \sim \mathrm{Dirichlet}(\alpha/I\,\mathbf{1}_I) \qquad \mathbf{b}_i|\beta, I \sim \mathrm{Dirichlet}(\beta)$$

  - Posteriors

$$n_{ij} = \sum_{t=2}^{T} I(s_{t-1} = i, s_t = j|Y, \theta) \qquad \mathbf{n}_i = [n_{i1} \cdots n_{iI}]$$

$$m_{ij} = \sum_{t=2}^{T} I(s_t = i, y_t = j|\theta) \qquad \mathbf{m}_i = [m_{i1} \cdots n_{iJ}]$$

$$\mathbf{a}_i|\mathrm{rest} \sim \mathrm{Dirichlet}(\alpha/I\,\mathbf{1}_I + \mathbf{n}_i) \qquad \mathbf{b}_i|\mathrm{rest} \sim \mathrm{Dirichlet}(\beta + \mathbf{m}_i)$$

The Infinite Hidden Markov Model (II)

- Hierarchical Dirichlet Process (IHMM)
  - $I = \infty$
  - Stick-breaking process

$$\hat{\epsilon}_i = \mathrm{Beta}(1, \gamma) \quad \epsilon_i = \hat{\epsilon}_i \prod_{l=1}^{i-1} (1 - \hat{\epsilon}_l) \quad \boldsymbol{\epsilon} \sim \mathrm{Stick}(\gamma)$$

  - Priors ($i \in \{1, \ldots, \infty\}$)

$$\boldsymbol{\epsilon} \sim \mathrm{Stick}(\gamma) \quad \mathbf{a}_i | \alpha, \boldsymbol{\epsilon} \sim \mathrm{Stick}(\alpha \boldsymbol{\epsilon}) \quad \mathbf{b}_i | \boldsymbol{\beta}, I \sim \mathrm{Dirichlet}(\boldsymbol{\beta})$$

  - Posteriors ($K \equiv$ number of active states, $\mathbf{a}_i = [a_{i1} \cdots a_{iK} \; \sum_{l=K+1}^{\infty} a_{il}]$, $\boldsymbol{\epsilon}_K = [\epsilon_i \cdots \epsilon_K \; \sum_{l=K+1}^{\infty} \epsilon_l]$)

$$\begin{aligned}
\mathbf{a}_i | \mathrm{rest} &\sim \mathrm{Dirichlet}(\alpha \boldsymbol{\epsilon}_K + \mathbf{n}_i) & \mathbf{b}_i | \mathrm{rest} &\sim \mathrm{Dirichlet}(\boldsymbol{\beta} + \mathbf{m}_i) \\
o_{ij} &\equiv \mathrm{resample}\ n_{ij} & &\mathrm{with\ Bernouilly}(\alpha \epsilon_j) \\
c_j &= \sum_i o_{ij} & \mathbf{c} &= [c_1 \cdots c_K \, \gamma] \\
\boldsymbol{\epsilon} | \mathrm{rest} &\sim \mathrm{Dirichlet}(\mathbf{c})
\end{aligned}$$

The Infinite Hidden Markov Model (III)

- Inference
  - Sampling S is challenging with $I = \infty$ (Forward-filtering Backward-sampling can not be employed)
  - Beam sampling make use of an auxiliary variable to work with a finite number of states [van Gael et al., 2008]

Spectral Learning of HMMs

- Discrete observations, $J \geq I$

$$
\begin{aligned}
p(Y) &= \sum_{s_{T+1}} \sum_{s_T} p(s_{T+1}|s_T) p(y_T|s_T) \cdots \sum_{s_1} p(s_2|s_1) p(y_1|s_1) p(s_1) \\
&= \mathbf{1}^T \mathbf{A} \mathrm{diag}(\mathbf{b}_{y_T}) \cdots \mathbf{A} \mathrm{diag}(\mathbf{b}_{y_1}) \boldsymbol{\pi} \\
&= \mathbf{1}^T \mathbf{A}_{y_T} \cdots \mathbf{A}_{y_1} \boldsymbol{\pi} \\
&= \mathbf{1}^T \mathbf{A}_{y_{T:1}} \boldsymbol{\pi} \\
&= \mathbf{c}_\infty^T \mathbf{C}_{y_{T:1}} \mathbf{c}_1
\end{aligned}
$$

$$
\begin{array}{lll}
\mathbf{p}_1 = p(y_1) & \mathbf{P}_{21} = p(y_2, y_1) & \mathbf{P}_{31}^x = p(y_3, y_1)|_{y_2 = x} \\
\hat{\mathbf{p}}_1 = \overline{p(y_1)} & \hat{\mathbf{P}}_{21} = \overline{p(y_2, y_1)} & \hat{\mathbf{P}}_{31}^x = \overline{p(y_3, y_1)}|_{y_2 = x}
\end{array}
$$

$$
\mathbf{P}_{21} = \mathbf{U} \boldsymbol{\Sigma} \mathbf{U}^T
$$

Spectral Learning of HMMs (II)

$$
\begin{aligned}
\mathbf{c}_1 &= \mathbf{U}^T\mathbf{p}_1 && = \mathbf{U}^T\mathbf{B}\boldsymbol{\pi} \\
\mathbf{c}_\infty &= \mathbf{P}_{21}^T\mathbf{U}\mathbf{p}_1 && = \mathbf{1}^T(\mathbf{U}^T\mathbf{B})^{-1} \\
\mathbf{C}_x &= (\mathbf{U}^T\mathbf{P}_{31}^x)(\mathbf{U}^T\mathbf{P}_{21})^+ && = (\mathbf{U}^T\mathbf{B})\mathbf{A}_x(\mathbf{U}^T\mathbf{B})^{-1}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{c}_{t+1} &= \frac{\mathbf{C}_{y_t}\mathbf{c}_t}{\mathbf{c}_\infty^T\mathbf{C}_{y_t}\mathbf{c}_t} \\
\mathbf{c}_t &= \mathbf{U}^T\mathbf{B}\alpha_t \\
p(y_t|y_{1:t-1}) &= \mathbf{c}_\infty\mathbf{C}_{y_t}\mathbf{c}_t
\end{aligned}
$$

- No local maxima
- Kernelized version for continuous observations [Song et al., 2010]

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970).
A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains.
The annals of mathematical statistics, 41(1):164–171.

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2001).
The infinite hidden Markov model.
In *Advances in Neural Information Processing Systems*.

Frühwirth-Schnatter, S. (2006).
Finite mixture and Markov switching models.
Springer Series in Statistics. Springer, New York.

Hsu, D., Kakade, S. M., and Zhang, T. (2012).
A spectral algorithm for learning Hidden Markov Models.
Journal of Computer and System Sciences.

MacKay, D. J. C. (1997).
Ensemble learning for hidden Markov models.

Rabiner, L. R. and Juang, B.-H. (1986).
An introduction to hidden Markov models.

IEEE ASSP Magazine, 3(1):4–16.

Robert, C. P., Celeux, G., and Diebolt, J. (1993).
Bayesian Estimation of Hidden Markov Chains: A Stochastic Implementation.
Statistics & Probability Letters, 16(1):77–83.

Song, L., Boots, B., Siddiqi, S. M., Gordon, G., and Smola, A. J. (2010).
Hilbert space embeddings of hidden Markov models.
In *International Conference on Machine learning*.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006).
Hierarchical dirichlet processes.
Journal of the american statistical association, 101(476).

van Gael, J., Saatci, Y., Teh, Y. W., and Ghahramani, Z. (2008).
Beam sampling for the infinite hidden Markov model.
In *International Conference on Machine learning*.