Hidden Markov Models

STATS 305C: Applied Statistics

Scott Linderman

May 9, 2022

Gaussian Mixture Models

Recall the basic Gaussian mixture model,

$$z_t \overset{\text{iid}}{\sim} \mathrm{Cat}(\pi) \tag{1}$$
$$x_t \mid z_t \sim \mathcal{N}(\boldsymbol{\mu}_{z_t}, \boldsymbol{\Sigma}_{z_t}) \tag{2}$$

where

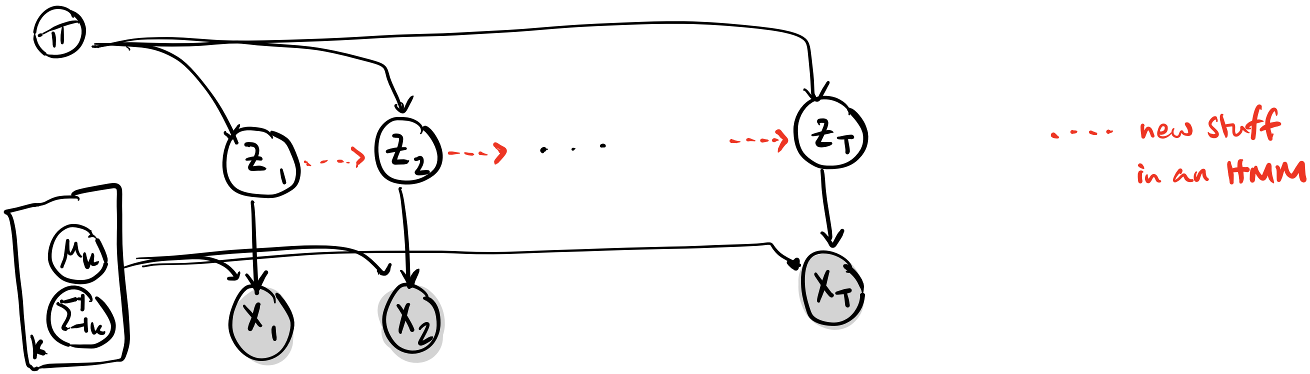
- $z_t \in \{1, \ldots, K\}$ is a **latent mixture assignment**
- $x_t \in \mathbb{R}^D$ is an **observed data point**
- $\pi \in \Delta_K$, $\boldsymbol{\mu}_k \in \mathbb{R}^D$, and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{D\times D}_{\succeq 0}$ are parameters

(Here we've switched to indexing data points by t rather than n.)

Let $\boldsymbol{\Theta}$ denote the set of parameters. We can be Bayesian and put a prior on $\boldsymbol{\Theta}$ and run Gibbs or VI, or we can point estimate $\boldsymbol{\Theta}$ with EM, etc.

Gaussian Mixture Models II

Draw the graphical model.



Gaussian Mixture Models III

Recall the EM algorithm for mixture models,

- **E step:** Compute the posterior distribution

$$
\begin{aligned}
q(\boldsymbol{z}_{1:T}) &= p(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T}; \boldsymbol{\Theta}) && (3)\\
&= \prod_{t=1}^{T} p(z_t \mid \boldsymbol{x}_t; \boldsymbol{\Theta}) && (4)\\
&= \prod_{t=1}^{T} q_t(z_t) && (5)
\end{aligned}
$$

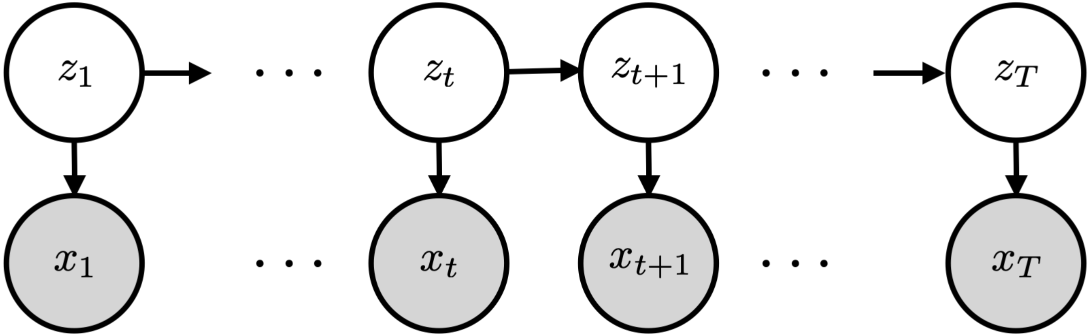
- **M step:** Maximize the ELBO wrt $\boldsymbol{\Theta}$,

$$
\begin{aligned}
\mathscr{L}(\boldsymbol{\Theta}, q) &= \mathbb{E}_{q(\boldsymbol{z}_{1:T})}\left[\log p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}; \boldsymbol{\Theta}) - \log q(\boldsymbol{z}_{1:T})\right] && (6)\\
&= \mathbb{E}_{q(\boldsymbol{z}_{1:T})}\left[\log p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}; \boldsymbol{\Theta})\right] + c. && (7)
\end{aligned}
$$

For exponential family mixture models, the M-step only requires expected sufficient statistics.

Hidden Markov Models

Hidden Markov Models (HMMs) are like mixture models with temporal dependencies between the mixture assignments.



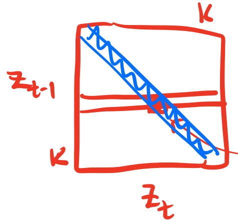
This graphical model says that the joint distribution factors as,

$$p(z_{1:T}, \boldsymbol{x}_{1:T}) = p(z_1) \prod_{t=2}^{T} p(z_t \mid z_{t-1}) \prod_{t=1}^{T} p(\boldsymbol{x}_t \mid z_t). \tag{8}$$

We call this an HMM because the *hidden* states follow a Markov chain, $p(z_1) \prod_{t=2}^{T} p(z_t \mid z_{t-1})$.

Hidden Markov Models II

$z_t \in \{1, \dots, K\}$



$P_{ij} = P(z_t = j \mid z_{t-1} = i)$

row-stoch: $\sum_j P_{ij} = 1$

An HMM consists of three components:

1. **Initial distribution:** $z_1 \sim \mathrm{Cat}(\pi_0)$
2. **Transition matrix:** $z_t \sim \mathrm{Cat}(\boldsymbol{P}_{z_{t-1}})$ where $\boldsymbol{P} \in [0,1]^{K \times K}$ is a *row-stochastic* transition matrix with rows $\boldsymbol{P}_k$.
3. **Emission distribution:** $\boldsymbol{x}_t \sim p(\cdot \mid \boldsymbol{\theta}_{z_t})$

$P_i \in \Delta_K$

① $P_i \overset{iid}{\sim} \mathrm{Dir}(\alpha 1_K)$

② $P_0 \sim \mathrm{Dir}(\alpha_0 1_K)$

$P_i \sim \mathrm{Dir}(\alpha P_0 + \gamma e_i)$

"hierarchical Dirichlet prior"

Hidden Markov Models III

We are interested in questions like:

- What are the *predictive distributions* of $p(z_{t+1} \mid \boldsymbol{x}_{1:t})$? [filtering]
- What is the *posterior marginal* distribution $p(z_t \mid \boldsymbol{x}_{1:T})$? [smoothing]
- What is the *posterior pairwise marginal* distribution $p(z_t, z_{t+1} \mid \boldsymbol{x}_{1:T})$?
- What is the *posterior mode* $z^\star_{1:T} = \arg\max p(z_{1:T} \mid \boldsymbol{x}_{1:T})$? [viterbi]
- How can we *sample the posterior* $p(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T})$ of an HMM?
- What is the *marginal likelihood* $p(\boldsymbol{x}_{1:T})$? Worst case: $O(K^T)$
- How can we *learn the parameters* of an HMM?

Question: Why might these sound like hard problems?

$$p(z_{1:T} \mid x_{1:T}) \propto p(z_{1:T}, x_{1:T})$$
$$\neq \prod_t p(z_t \mid x_{1:T})$$

Computing the predictive distributions

The predictive distributions give the probability of the latent state z_{t+1} given observations *up to but not including* time $t+1$. Let,

$$
\begin{aligned}
\alpha_{t+1}(z_{t+1}) &\triangleq p(z_{t+1}, \boldsymbol{x}_{1:t}) \qquad = \sum_{z_1} \cdots \sum_{z_t} p(z_{1:t+1}, x_{1:t}) && (9)\\
&= \sum_{z_1=1}^{K} \cdots \sum_{z_t=1}^{K} p(z_1) \prod_{s=1}^{t} p(\boldsymbol{x}_s \mid z_s)\, p(z_{s+1} \mid z_s) && (10)\\
&= \sum_{z_t=1}^{K} \left[\left(\sum_{z_1=1}^{K} \cdots \sum_{z_{t-1}=1}^{K} p(z_1) \prod_{s=1}^{t-1} p(\boldsymbol{x}_s \mid z_s)\, p(z_{s+1} \mid z_s) \right) p(\boldsymbol{x}_t \mid z_t)\, p(z_{t+1} \mid z_t) \right] && (11)\\
&= \sum_{z_t=1}^{K} \alpha_t(z_t)\, p(\boldsymbol{x}_t \mid z_t)\, p(z_{t+1} \mid z_t). && (12)
\end{aligned}
$$

We call $\alpha_t(z_t)$ the *forward messages*. We can compute them recursively! The base case is $p(z_1 \mid \varnothing) \triangleq p(z_1)$.

$\hookrightarrow = \alpha_1(z_1)$

Computing the predictive distributions II

We can also write these recursions in a vectorized form. Let

$$\boldsymbol{\alpha}_t = \begin{bmatrix} \alpha_t(z_t = 1) \\ \vdots \\ \alpha_t(z_t = K) \end{bmatrix} = \begin{bmatrix} p(z_t = 1, \boldsymbol{x}_{1:t-1}) \\ \vdots \\ p(z_t = K, \boldsymbol{x}_{1:t-1}) \end{bmatrix} \quad \text{and} \quad \boldsymbol{l}_t = \begin{bmatrix} p(\boldsymbol{x}_t \mid z_t = 1) \\ \vdots \\ p(\boldsymbol{x}_t \mid z_t = K) \end{bmatrix} \tag{13}$$

both be vectors in $\mathbb{R}_+^K$. Then,

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{P}^\top (\boldsymbol{\alpha}_t \odot \boldsymbol{l}_t) \tag{14}$$

[O(K²) complexity]
for one step

where $\odot$ denotes the Hadamard (elementwise) product and $\boldsymbol{P}$ is the transition matrix.

⇒ O(TK²) complexity
for whole forward pass

Computing the predictive distributions III

Finally, to get the predictive distributions we just have to normalize,

$$p(z_{t+1} \mid \boldsymbol{x}_{1:t}) \propto p(z_{t+1}, \boldsymbol{x}_{1:t}) = \alpha_{t+1}(z_{t+1}). \tag{15}$$

Question: What does the normalizing constant tell us?

$$p(z_{t+1} \mid x_{1:t}) = \frac{p(z_{t+1}, x_{1:t})}{p(x_{1:t})} = \frac{p(z_{t+1}, x_{1:t})}{\sum_{k=1}^{K} p(z_{t+1}=k, x_{1:t})}$$

Computing the posterior marginal distributions

The posterior marginal distributions give the probability of the latent state z_t given *all the observations* up to time T.

$$
\begin{aligned}
p(z_t \mid \mathbf{x}_{1:T}) &= \sum_{z_1=1}^{K} \cdots \sum_{z_{t-1}=1}^{K} \sum_{z_{t+1}=1}^{K} \cdots \sum_{z_T=1}^{K} p(\mathbf{z}_{1:T}, \mathbf{x}_{1:T}) \qquad (16) \\
&= \underbrace{\left[\sum_{z_1=1}^{K} \cdots \sum_{z_{t-1}=1}^{K} p(z_1) \prod_{s=1}^{t-1} p(\mathbf{x}_s \mid z_s)\, p(z_{s+1} \mid z_s) \right]}_{\alpha_t} \times \overset{\ell_t}{p(\mathbf{x}_t \mid z_t)} \\
&\quad \times \underbrace{\left[\sum_{z_{t+1}=1}^{K} \cdots \sum_{z_T=1}^{K} \prod_{u=t+1}^{T} p(z_u \mid z_{u-1})\, p(\mathbf{x}_u \mid z_u) \right]}_{\beta_t} \qquad (17) \\
&= \alpha_t(z_t) \times p(\mathbf{x}_t \mid z_t) \times \beta_t(z_t) \qquad (18)
\end{aligned}
$$

where we have introduced the *backward messages* $\beta_t(z_t)$.

Computing the backward messages

The backward messages can be computed recursively too,

$$
\begin{aligned}
\beta_t(z_t) &\triangleq \sum_{z_{t+1}=1}^{K} \cdots \sum_{z_T=1}^{K} \prod_{u=t+1}^{T} p(z_u \mid z_{u-1})\, p(\boldsymbol{x}_u \mid z_u) && (19)\\
&= \sum_{z_{t+1}=1}^{K} p(z_{t+1} \mid z_t)\, p(\boldsymbol{x}_{t_1\,\color{red}{t+1}} \mid z_{t+1}) \left(\sum_{z_{t+2}=1}^{K} \cdots \sum_{z_T=1}^{K} \prod_{u=t+2}^{T} p(z_u \mid z_{u-1})\, p(\boldsymbol{x}_u \mid z_u) \right) && (20)\\
&= \sum_{z_{t+1}=1}^{K} p(z_{t+1} \mid z_t)\, p(\boldsymbol{x}_{t_1} \mid z_{t+1})\, \beta_{t+1}(z_{t+1}). && (21)
\end{aligned}
$$

For the base case, let $\beta_T(z_T) = 1$.

Computing the backward messages (vectorized)

Let

$$\boldsymbol{\beta}_t = \begin{bmatrix} \beta_t(z_t = 1) \\ \vdots \\ \beta_t(z_t = K) \end{bmatrix} \tag{22}$$

be a vector in $\mathbb{R}_+^K$. Then,

$$\boldsymbol{\beta}_t = \boldsymbol{P}(\boldsymbol{\beta}_{t+1} \odot \boldsymbol{l}_{t+1}). \qquad O(K^2) \text{ per step} \to O(TK^2) \text{ total} \tag{23}$$

Let $\boldsymbol{\beta}_T = \mathbf{1}_K$.

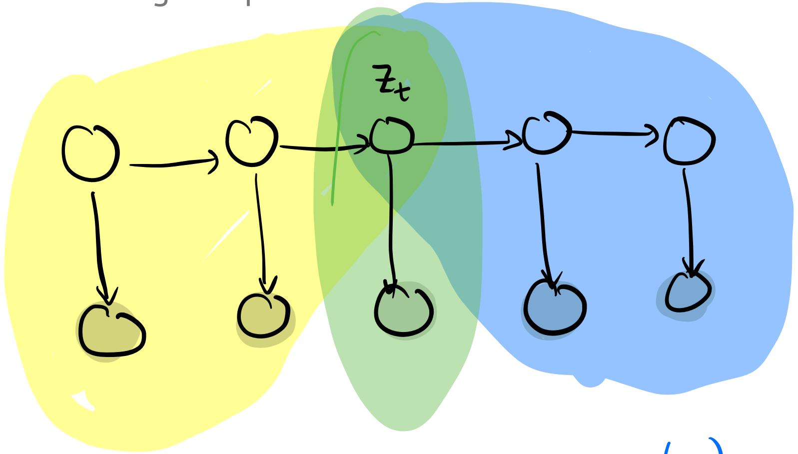
Now we have everything we need to compute the posterior marginal,

$$p(z_t = k \mid \boldsymbol{x}_{1:T}) = \frac{\alpha_{t,k}\, l_{t,k}\, \beta_{t,k}}{\sum_{j=1}^K \alpha_{t,j} l_{t,j} \beta_{t,j}}. \tag{24}$$

We just derived the **forward-backward algorithm** for HMMs [Rabiner and Juang, 1986].

What do the backward messages represent?

Question: If the forward messages represent the predictive probabilities $\alpha_{t+1}(z_{t+1}) = p(z_{t+1}, \mathbf{x}_{1:t})$, what do the backward messages represent?



$\alpha_t(z_t) = p(z_t, x_{1:t-1})$ $\quad \ell_t(z_t) = p(x_t \mid z_t)$ $\quad \beta_t(z_t) = p(x_{t+1:T} \mid z_t)$

junction-tree algorithm, complexity ~ tree-width

Computing the posterior pairwise marginals

Exercise: Use the forward and backward messages to compute the posterior pairwise marginals $p(z_t, z_{t+1} \mid \boldsymbol{x}_{1:T})$.

Normalizing the messages for numerical stability

If you're working with long time series, especially if you're working with 32-bit floating point, you need to be careful.

The messages involve products of probabilities, which can quickly overflow.

There's a simple fix though: after each step, re-normalize the messages so that they sum to one. I.e replace

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{P}^\top (\boldsymbol{\alpha}_t \odot \boldsymbol{l}_t) \tag{25}$$

$$p(x_{1:T}) = p(x_1)\, p(x_2 \mid x_1) \cdots p(x_T \mid x_{1:T-1})$$
$$= \prod_t A_t$$

with

$$\widetilde{\boldsymbol{\alpha}}_{t+1} = \frac{1}{A_t} \boldsymbol{P}^\top (\widetilde{\boldsymbol{\alpha}}_t \odot \boldsymbol{l}_t) \tag{26}$$

$$A_t = \sum_{k=1}^{K} \sum_{j=1}^{K} P_{jk} \widetilde{\alpha}_{t,j} l_{t,j} \equiv \sum_{j=1}^{K} \widetilde{\alpha}_{t,j} l_{t,j} \quad \text{(since } \boldsymbol{P} \text{ is row-stochastic).} \tag{27}$$

This leads to a nice interpretation: The normalized messages are predictive likelihoods $\widetilde{\alpha}_{t+1,k} = p(z_{t+1} = k \mid \boldsymbol{x}_{1:t})$, and the normalizing constants are $A_t = p(\boldsymbol{x}_t \mid \boldsymbol{x}_{1:t-1})$.

EM for Hidden Markov Models

Now we can put it all together. To perform EM in an HMM,

- **E step:** Compute the posterior distribution

$$q(\boldsymbol{z}_{1:T}) = p(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T}; \boldsymbol{\Theta}). \tag{28}$$

  (Really, run the **forward-backward algorithm** to get posterior marginals and pairwise marginals.)

- **M step:** Maximize the ELBO wrt $\boldsymbol{\Theta}$,

$$\mathscr{L}(\boldsymbol{\Theta}) = \mathbb{E}_{q(\boldsymbol{z}_{1:T})}\left[\log p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}; \boldsymbol{\Theta})\right] + c \tag{29}$$

$$= \mathbb{E}_{q(\boldsymbol{z}_{1:T})}\left[\sum_{k=1}^{K} \mathbb{I}[z_1 = k] \log \pi_{0,k}\right] + \mathbb{E}_{q(\boldsymbol{z}_{1:T})}\left[\sum_{t=1}^{T-1}\sum_{i=1}^{K}\sum_{j=1}^{K} \mathbb{I}[z_t = i, z_{t+1} = j] \log P_{i,j}\right]$$

$$+ \mathbb{E}_{q(\boldsymbol{z}_{1:T})}\left[\sum_{t=1}^{T}\sum_{k=1}^{K} \mathbb{I}[z_t = k] \log p(\boldsymbol{x}_t; \theta_k)\right] \tag{30}$$

For exponential family observations, the M-step only requires expected sufficient statistics.

What else?

- How can we sample the posterior?
- How can we find the posterior mode?
- How can we choose the number of states?
- What if my transition matrix is sparse?

$$z_T \sim p(z_T \mid x_{1:T})$$

$$z_{T-1} \sim \underbrace{p(z_{T-1} \mid x_{1:T-2})\, p(x_{T-1} \mid z_{T-1})}_{\alpha_{T-1}(z_{T-1})}\, p(z_T \mid z_{T-1})$$

$$\vdots$$

References I

Lawrence Rabiner and Biinghwang Juang. An introduction to hidden Markov models. *ieee assp magazine*, 3(1):4–16, 1986.