Hidden Markov Models

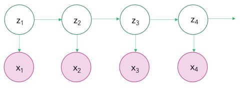
November 11, 2020

Acknowledgements

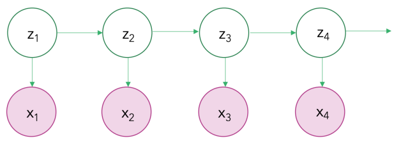
An important resource for these slides was Christopher Bishop's *Pattern Recognition and Machine Learning*.

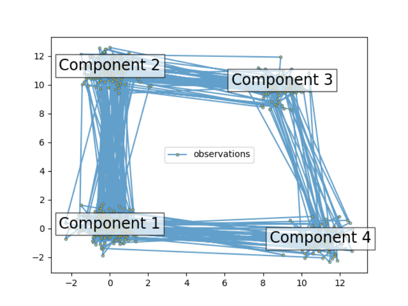
The model

- Observations $x_1, ..., x_n$.
- Latent variables $\mathbf{z}_n$ encode class (1-of-K encoding scheme).Assume a data point belongs to exactly one group or class out of K possibilities,
- Parameters $\theta = \{A, \pi, \phi\}$
  - Transition probabilities $p(\mathbf{z}_n|\mathbf{z}_{n-1})$ given by $\mathbf{A}$, where $A_{jk} = p(z_{nk} = 1|z_{n-1,j} = 1)$
  - Emission probabilities $p(\mathbf{x}_n|\mathbf{z_n}, \phi)$ governed by ϕ. (E.g. $\phi = \{\mu_k, \Sigma_k\}_{k=1}^K$)
  - Distribution of $\mathbf{z}$ given by π: $\pi_k = p(z_{1k} = 1)$



Hidden Markov Models





Likelihood

$$P(X|\theta) = \sum_{Z} P(X, Z|\theta)$$

What would be hard about maximizing the likelihood?

Hidden Markov Models

Likelihood

$$P(X|\theta) = \sum_{Z} P(X, Z|\theta)$$

What would be hard about maximizing the likelihood?

- No closed-form solution for maximum likelihood
- The number of terms in the summation goes as K^N

Complete data likelihood

$$p(X, Z|\theta) = p(\mathbf{z}_1|\pi) \prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1}, A) \prod_{m=1}^{N} p(\mathbf{x}_m|\mathbf{z}_m, \phi)$$

EM

E : Compute

$$Q(\theta, \theta^{(old)}) = \mathbb{E}_{Z|X,\theta^{(old)}} \ln p(X, Z|\theta).$$

M : Find θ to maximize $Q(\theta, \theta^{(old)})$.

$$
\begin{aligned}
p(X, Z|\theta) &= p(\mathbf{z}_1|\pi) \prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1}, A) \prod_{m=1}^{N} p(\mathbf{x}_m|\mathbf{z}_m, \phi) \\
&= (\prod_{k=1}^{K} \pi_k^{z_{1k}})(\prod_{n=1}^{N} \prod_{k=1}^{K} \prod_{j=1}^{K} A_{jk}^{z_{n-1,j} z_{nk}})(\prod_{m=1}^{N} \prod_{k=1}^{K} z_{mk} p(\mathbf{x}_n|\phi_k)
\end{aligned}
$$

E-step

Define: $\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{X}, \theta^{(old)})$ Note: $\gamma(z_{nk}) = \mathbb{E}[z_{nk}]$
$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X}, \theta^{(old)})$. Note: $\xi(z_{n-1,j}, z_{nk}) = \mathbb{E}[z_{n-1,j} z_{nk}]$
Evaluate γ, ξ (how? forward backward algorithm). Then can compute:

$$
\begin{aligned}
Q(\theta, \theta^{(old)}) =& \mathbb{E}_{Z|X,\theta^{(old)}} \ln p(X, Z|\theta) \\
&= \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j} z_{nk}) \ln A_{jk} \\
&+ \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(\mathbf{x}_n|\phi_k)
\end{aligned}
$$

$$
\begin{aligned}
Q(\theta, \theta^{(old)}) =& \mathbb{E}_{Z|X,\theta^{(old)}} \ln p(X, Z|\theta) \\
&= \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j} z_{nk}) \ln A_{jk} \\
&+ \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(\mathbf{x}_n|\phi_k)
\end{aligned}
$$

M-step: Find θ to maximize $Q(\theta, \theta^{(old)})$.

$\pi_k = \frac{\gamma(z_{1k})}{\sum_{k=1}^{K} \gamma(z_{1k})}$, $A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j} z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})}$
Result of maximization with respect to ϕ depends on choice of emission probabilities.

M-step: Find θ to maximize $Q(\theta, \theta^{(old)})$.

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{k=1}^{K} \gamma(z_{1k})}, A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j} z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})}$$

Result of maximization with respect to ϕ depends on choice of emission probabilities.

E.g if $p(\mathbf{x}|\phi_k) = \text{Normal}(\mathbf{x}|\mu_k, \Sigma_k)$

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})}, \Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

EM for HMMs summary:

Alternate between updating γ, ξ and updating π, A, ϕ.

How to efficiently calculate γ, ξ for the E-step?

Use a two-stage message passing algorithm, the *forward-backward algorithm* to compute.

For details and derivation, see, e.g. Christopher Bishop's *Pattern Recognition and Machine Learning* sec. 13.2.2.