Exercises on Hidden Markov Model

immediate

April 9, 2018

1 Demonstration of KL Divergence

- Convex function
$$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda)f(b) \ (\lambda \geq 1)$$
- By induction, we obtain the Jensen's inequality
$$f\left(\sum_{i=1}^{M} \lambda_i x_i\right) \leq \sum_{i=1}^{M} \lambda_i f(x_i)$$
where $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$ for any set of points $\{x_i\}$.
- Interpret λ_i as the probability distribution over a discrete variable x taking the values $\{x_i\}$
$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$$
- For continuous variables, Jensen's inequality
$$f\left(\int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x}\right) \leq \int f(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$
- $-\ln x$ is a convex function
- Apply Jensen's inequality to the KL divergence
$$KL(p\|q) = -\int p(\boldsymbol{x}) \ln\left\{\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right\} d\boldsymbol{x} \geq -\ln \int q(\boldsymbol{x}) d(\boldsymbol{x}) = 0$$

2 Sample

Q: Verify the conditional distribution for $\boldsymbol{x}_n$ given all of the observations up to time n is given by

$$p(\boldsymbol{x}_n | \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n-1}) = p(\boldsymbol{x}_n | \boldsymbol{x}_{n-1})$$

A: We We first of all find the joint distribution $p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ by marginalizing over the variables $\boldsymbol{x}_{n+1}, \cdots, \boldsymbol{x}_N$, to give

$$\begin{aligned}
p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) &= \sum_{\boldsymbol{x}_{n+1}} \cdots \sum_{\boldsymbol{x}_N} p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) \\
&= \sum_{\boldsymbol{x}_{n+1}} \cdots \sum_{\boldsymbol{x}_N} p(\boldsymbol{x}_1) \prod_{m=2}^{N} p(\boldsymbol{x}_m | \boldsymbol{x}_{m-1}) \\
&= p(\boldsymbol{x}_1) \prod_{m=2}^{n} p(\boldsymbol{x}_m | \boldsymbol{x}_{m-1})
\end{aligned}$$

Now we evaluate the required conditional distribution

$$p(\boldsymbol{x}_n | \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n-1}) = \frac{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)}{\sum_{\boldsymbol{x}_n} p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)} = \frac{p(\boldsymbol{x}_1) \prod_{m=2}^{n} p(\boldsymbol{x}_m | \boldsymbol{x}_{m-1})}{\sum_{\boldsymbol{x}_n} p(\boldsymbol{x}_1) \prod_{m=2}^{n} p(\boldsymbol{x}_m | \boldsymbol{x}_{m-1})}$$

Note that any factors which do not depend on $\boldsymbol{x}_n$ will cancel between numerator and denominator, giving

$$p(\boldsymbol{x}_n | \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n-1}) = \frac{p(\boldsymbol{x}_n | \boldsymbol{x}_{n-1})}{\sum_{\boldsymbol{x}_n} p(\boldsymbol{x}_n | \boldsymbol{x}_{n-1})} = p(\boldsymbol{x}_n | \boldsymbol{x}_{n-1})$$

3 Exercise 1

Q: Show that second-order Markov chain described by the joint distribution

$$p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) = p(\boldsymbol{x}_1) p(\boldsymbol{x}_2 | \boldsymbol{x}_1) \prod_{n=3}^{N} p(\boldsymbol{x}_n | \boldsymbol{x}_{n-1}, \boldsymbol{x}_{n-2})$$

satisfies the conditional independent property

$$p(\boldsymbol{x}_n | \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n-1}) = p(\boldsymbol{x}_n | \boldsymbol{x}_{n-1}, \boldsymbol{x}_{n-2})$$

A: The marginal distribution over the variables $p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ is given by

$$\begin{aligned}
p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) &= \sum_{\boldsymbol{x}_{n+1}} \cdots \sum_{\boldsymbol{x}_N} p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) \\
&= \sum_{\boldsymbol{x}_{n+1}} \cdots \sum_{\boldsymbol{x}_N} p(\boldsymbol{x}_1) p(\boldsymbol{x}_2 | \boldsymbol{x}_1) \prod_{m=3}^{N} p(\boldsymbol{x}_m | \boldsymbol{x}_{m-1}, \boldsymbol{x}_{m-2}) \\
&= p(\boldsymbol{x}_1) p(\boldsymbol{x}_2 | \boldsymbol{x}_1) \prod_{m=3}^{n} p(\boldsymbol{x}_m | \boldsymbol{x}_{m-1}, , \boldsymbol{x}_{m-2})
\end{aligned}$$

The required conditional distribution is then given by

$$p(\boldsymbol{x}_n | \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n-1}) = \frac{p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)}{\sum_{\boldsymbol{x}_n} p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)} = \frac{p(\boldsymbol{x}_1) p(\boldsymbol{x}_2 | \boldsymbol{x}_1) \prod_{m=3}^{n} p(\boldsymbol{x}_m | \boldsymbol{x}_{m-1}, \boldsymbol{x}_{m-2})}{\sum_{\boldsymbol{x}_n} p(\boldsymbol{x}_1) p(\boldsymbol{x}_2 | \boldsymbol{x}_1) \prod_{m=3}^{n} p(\boldsymbol{x}_m | \boldsymbol{x}_{m-1}, \boldsymbol{x}_{m-2})}$$

Again, cancelling factors independent of $\boldsymbol{x}_n$ between numerator and denominator we obtain

$$p(\boldsymbol{x}_n | \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n-1}) = \frac{p(\boldsymbol{x}_n | \boldsymbol{x}_{n-1}, \boldsymbol{x}_{n-2})}{\sum_{\boldsymbol{x}_n} p(\boldsymbol{x}_n | \boldsymbol{x}_{n-1}, \boldsymbol{x}_{n-2})} = p(\boldsymbol{x}_n | \boldsymbol{x}_{n-1}, \boldsymbol{x}_{n-2})$$

4 Exercise 2

Q: The joint probability distribution over both latent and observed variables of HMM is

$$p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) = p(\boldsymbol{z}_1|\boldsymbol{\pi}) \left[\prod_{t=2}^{T} p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, \boldsymbol{A})\right] \prod_{t=1}^{T} p(\boldsymbol{x}_t|\boldsymbol{z}_t, \boldsymbol{\phi})$$

where $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$, $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$, and $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{\phi}\}$ denotes the set of parameters governing the model. Given the marginal distribution

$$\gamma(z_{tk}) = \mathbb{E}[z_{tk}] = \sum_{\boldsymbol{z}} \gamma(\boldsymbol{z}) z_{tk}$$
$$\xi(z_{t-1,i}, z_{tj}) = \mathbb{E}[z_{t-1,i}, z_{tj}] = \sum_{\boldsymbol{z}} \gamma(\boldsymbol{z}) z_{t-1,i} z_{tj}$$

The expectation of the logarithm of the complete-data likelihood function defined by

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) &= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \\ &= \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \sum_{t=2}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K} \xi(z_{t-1,i}, z_{t,j}) \ln A_{ij} + \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma(z_{tk}) \ln p(\boldsymbol{x}_t|\boldsymbol{\phi}_k) \end{aligned}$$

A: In the M step, we maximize $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ with respect to the parameters $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{\phi}\}$ in which we treat $\gamma(\boldsymbol{z}_t)$ and $\xi(\boldsymbol{z}_{t-1}, \boldsymbol{z}_t)$ as constant. Maximization with respect to $\boldsymbol{\phi}$ and $\boldsymbol{A}$ is easily achieved using appropriate Lagrange multipliers.

Maximization with respects to $\boldsymbol{\pi}$: Take account of the summation constraint

$$\sum_{k=1}^{K} \pi_k = 1 \tag{1}$$

We there first omit terms from $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ which are independent of $\boldsymbol{\pi}$, and then add a Lagrange multiplier term to enforce the constraint, giving the following function to be maximized

$$\tilde{Q} = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \lambda \left(\sum_{k=1}^{K} \pi_k - 1\right) \tag{2}$$

Setting the derivative with respect to π_k equal to zero we obtain

$$0 = \gamma(z_{1k}) \frac{1}{\pi_k} + \lambda \tag{3}$$

We now multiply through by π_k and then sum over k and make use of the summation constraint to give

$$\lambda = -\sum_{k=1}^{K} \gamma(z_{1k}) \tag{4}$$

Substituting back into (2)

$$\tilde{Q} = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k - \sum_{k=1}^{K} \gamma(z_{1k}) \left(\sum_{k=1}^{K} \pi_k - 1 \right) \tag{5}$$

and solving for π_k we obtain

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{i=1}^{K} \gamma(z_{1i})} \tag{6}$$

Maximization with respect to $\boldsymbol{A}$: For the maximization with respect to $\boldsymbol{A}$ we follow the same steps and first omit terms from $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ which are independent of $\boldsymbol{A}$, and then add appropriate Lagrange multiplier terms to enforce the summation constraints. In this case there are K constraints to be satisfied since we must have

$$\sum_{j=1}^{K} A_{ij} = 1 \tag{7}$$

for $i = 1, \cdots, K$. We introduce K Lagrange multipliers λ_i for $i = 1, \cdots, K$, and maximize the following function

$$\hat{Q} = \sum_{t=2}^{T} \sum_{i=1}^{K} \sum_{j=1}^{K} \xi(z_{t-1,i}, z_{tj}) \ln A_{ij} + \sum_{i=1}^{K} \lambda_i \left(\sum_{j=1}^{K} A_{ij} - 1 \right) \tag{8}$$

Setting the derivative of $\hat{Q}$ with respect to A_{ij} to zero we obtain

$$0 = \sum_{t=2}^{T} \xi(z_{t-1,i}, z_{tj}) \frac{1}{A_{ij}} + \lambda_i \tag{9}$$

Again we multiply through by A_{ij} and then sum over j and make use of the summation constraint to give

$$\lambda_i = - \sum_{t=2}^{T} \sum_{j=1}^{K} \xi(z_{t-1,i}, z_{tj}) \tag{10}$$

Substituting for λ_i in (8) and solving for A_{ij} we obtain

$$A_{ij} = \frac{\sum_{t=2}^{T} \xi(z_{t-1,i}, z_{tj})}{\sum_{k=1}^{K} \sum_{t=2}^{T} \xi(z_{t-1,i}, z_{tk})} \tag{11}$$

Maximization with respect to $\boldsymbol{\phi}$: To maximize $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ with respect to $\boldsymbol{\phi}_k$, we notice that only the final term depends on $\boldsymbol{\phi}_k$.

In the case of Gaussian emission densities we have $p(\boldsymbol{x}|\boldsymbol{\phi}_k) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, and maximize the following function

$$\begin{aligned} \bar{Q} &= \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma(z_{tk}) \ln \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) &(12) \\ &= \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma(z_{tk}) \left(\frac{D}{2} \ln(2\pi) - \frac{1}{2} \ln |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\boldsymbol{x}_t - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_k) \right) &(13) \end{aligned}$$

Setting the derivative of $\bar{Q}$ with respect to $\boldsymbol{\mu}_k$ to zero we obtain

$$\sum_{t=1}^{T} \gamma(z_{tk}) \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_k) = 0 \tag{14}$$

$$\Rightarrow \qquad \sum_{t=1}^{T} \gamma(z_{tk}) \boldsymbol{x}_t = \boldsymbol{\mu}_k \sum_{t=1}^{T} \gamma(z_{tk}) \tag{15}$$

$$\Rightarrow \qquad \boldsymbol{\mu}_k = \frac{\sum_{t=1}^{T} \gamma(z_{tk}) \boldsymbol{x}_t}{\sum_{t=1}^{T} \gamma(z_{tk})} \tag{16}$$

Next if we define

$$N_k = \sum_{t=1}^{T} \gamma(z_{tk}) \tag{17}$$

$$\hat{\boldsymbol{S}}_k = \sum_{t=1}^{T} \gamma(z_{tk}) (\boldsymbol{x}_t - \boldsymbol{\mu}_k)(\boldsymbol{x}_t - \boldsymbol{\mu}_k)^T \tag{18}$$

then we can rewrite $\bar{Q}$ in the form

$$\bar{Q} = -\frac{N_k D}{2} \ln(2\pi) - \frac{N_k}{2} \ln|\boldsymbol{\Sigma}_k| - \frac{1}{2} \mathrm{Tr}(\boldsymbol{\Sigma}_k^{-1} \hat{\boldsymbol{S}}_k) \tag{19}$$

Setting the derivative of $\bar{Q}$ with respect to $\boldsymbol{\Sigma}_k^{-1}$ to zero we obtain

$$\frac{N_k}{2} \boldsymbol{\Sigma}_k^T - \frac{1}{2} \hat{\boldsymbol{S}}_k^T = 0 \tag{20}$$

$$\Rightarrow \qquad \boldsymbol{\Sigma}_k = \frac{\hat{\boldsymbol{S}}_k}{N_k} = \frac{\sum_{t=1}^{T} \gamma(z_{tk}) (\boldsymbol{x}_t - \boldsymbol{\mu}_k)(\boldsymbol{x}_t - \boldsymbol{\mu}_k)^T}{\sum_{t=1}^{T} \gamma(z_{tk})} \tag{21}$$