Imperial College London

COURSEWORK 2: HIDDEN MARKOV MODELS

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

495 Advanced Statistical Machine Learning and Pattern Recognition

Author:
Thomas Teh (CID: 0124 3008)

Date: February 21, 2017

1 Exercise I: Implementation Exercise

1.1 Files Submitted

Apart from the `HMMGenerateData.m`, I have implemented the HMM in files listed below:

m-files	Brief Description
`HMMGenerateData`	Data generator provided by course instructor.
`HMMExpectationDiscrete`	Filtering, smoothing and expectation for discrete case.
`HMMExpectationContinuous`	Filtering, smoothing and expectation for continuous case.
`HMMMaximizationDiscrete`	Maximization step for discrete case.
`HMMMaximizationContinuous`	Maximization step for continuous case.
`HMMViterbiDiscrete`	Viterbi decoding for the discrete case.
`HMMViterbiContinuous`	Viterbi decoding for the continuous case.
`HMM`	Combined function for the EM and decoding for both cases.

2 Exercise II: Parameter Estimation

2.1 Part i

Based on the stochastic automaton, we have the following transition probability $P(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$

$$p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \begin{bmatrix} 0 & a_{12} & 0 & 0 & 0 \\ 0 & a_{22} & a_{23} & 0 & 0 \\ 0 & 0 & a_{33} & a_{34} & a_{35} \\ 0 & 0 & 0 & a_{44} & a_{45} \\ 0 & 0 & 0 & 0 & a_{55} \end{bmatrix}$$

Hence, we will have the following constraints when we get the parameter estimates via maximum likelihood

$$\begin{aligned} a_{12} = 1 &\Rightarrow a_{1j} = 0, \forall j = 1,3,4,5 \\ a_{22} + a_{23} = 1 &\Rightarrow a_{2j} = 0, \forall j = 1,4,5 \\ a_{33} + a_{34} + a_{35} = 1 &\Rightarrow a_{3j} = 0, \forall j = 1,2 \\ a_{44} + a_{45} = 1 &\Rightarrow a_{4j} = 0, \forall j = 1,2,3 \\ a_{55} = 1 &\Rightarrow a_{5j} = 0, \forall j = 1,2,3,4 \end{aligned}$$

Since the model given is a bigram model, we have the following distribution of each character $\boldsymbol{x}_t$

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \prod_{j=1}^{5}\prod_{k=1}^{5} a_{jk}^{x_{t-1,j}x_{t,k}}$$

$$p(\boldsymbol{x}_1|\boldsymbol{\pi}) = \prod_{k=1}^{5} \pi_k^{x_{1,k}}$$

For each sequence $\boldsymbol{D}_l$, we have the probability of the sequence to be

$$\begin{aligned} p(\boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_T^l) &= p(\boldsymbol{x}_1^l) \prod_{t=2}^{T} p(\boldsymbol{x}_t^l|\boldsymbol{x}_{t-1}^l) \\ &= \prod_{k=1}^{5} \pi_k^{x_{1,k}^l} \prod_{t=2}^{T} \prod_{j=1}^{5} \prod_{k=1}^{5} a_{jk}^{x_{t-1,j}^l x_{t,k}^l} \end{aligned}$$

Since we have N sequences, the probability distribution can be written as

$$\begin{aligned} p(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_N) &= \prod_{l=1}^{N} (\boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_T^l) \\ &= \prod_{l=1}^{N} \prod_{k=1}^{5} \pi_k^{x_{1,k}^l} \prod_{t=2}^{T} \prod_{j=1}^{5} \prod_{k=1}^{5} a_{jk}^{x_{t-1,j}^l x_{t,k}^l} \end{aligned}$$

Taking the log, we have

$$\begin{aligned} \ln p(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_N) &= \sum_{l=1}^{N} \sum_{k=1}^{5} x_{1,k}^l \ln \pi_k + \sum_{l=1}^{N} \sum_{t=2}^{T} \sum_{j=1}^{5} \sum_{k=1}^{5} x_{t-1,j}^l x_{t,k}^l \ln a_{jk} \\ &= \sum_{k=1}^{5} \left(\sum_{l=1}^{N} x_{1,k}^l \right) \ln \pi_k + \sum_{j=1}^{5} \sum_{k=1}^{5} \left(\sum_{l=1}^{N} \sum_{t=2}^{T} x_{t-1,j}^l x_{t,k}^l \right) \ln a_{jk} \\ &= \sum_{k=1}^{5} N_k^1 \ln \pi_k + \sum_{j=1}^{5} \sum_{k=1}^{5} N_{jk} \ln a_{jk} \end{aligned}$$

where

$$\begin{aligned} N_k^1 &= \sum_{l=1}^{N} x_{1,k}^l \\ N_{jk} &= \sum_{l=1}^{N} \sum_{t=2}^{T} x_{t-1,j}^l x_{t,k}^l \end{aligned}$$

Formulating the Lagrangian, we have

$$\begin{aligned} \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{A}) = &\sum_{k=1}^{5} N_k^1 \ln \pi_k + \sum_{j=1}^{5} \sum_{k=1}^{5} N_{jk} \ln a_{jk} - \lambda \left(\sum_{k=1}^{5} \pi_k - 1 \right) - \sum_{j=1}^{5} \left(\gamma_j \sum_{k=1}^{5} a_{jk} - 1 \right) \\ &- \sum_{j=1,3,4,5} \mu_{1j} a_{1j} - \sum_{j=1,4,5} \mu_{2j} a_{2j} - \sum_{j=1,2} \mu_{3j} a_{3j} - \sum_{j=1,2,3} \mu_{4j} a_{4j} - \sum_{j=1,2,3,4} \mu_{5j} a_{5j} \end{aligned}$$

Taking the derivatives and setting them to zero

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \frac{N_k^1}{\pi_k} - \lambda = 0 \Rightarrow \lambda \pi_k = N_k^1$$
$$\lambda \sum_{k=1}^{5} \pi_k = \sum_{k=1}^{5} N_k^1 \Rightarrow \lambda = \sum_{k=1}^{5} N_k^1$$
$$\pi_k = \frac{N_k^1}{\sum_{j=1}^{5} N_j^1}$$

Likewise, we have

$$\frac{\partial \mathcal{L}}{\partial a_{jk}} = \frac{N_{jk}}{a_{jk}} - \gamma_j = 0 \Rightarrow \gamma_j a_{jk} = N_{jk}$$
$$a_{jk} = \frac{N_{jk}}{\sum_{k=1}^{5} N_{jk}}$$

$$p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & \frac{N_{22}}{N_{22}+N_{23}} & \frac{N_{23}}{N_{22}+N_{23}} & 0 & 0 \\ 0 & 0 & \frac{N_{33}}{N_{33}+N_{34}+N_{35}} & \frac{N_{34}}{N_{33}+N_{34}+N_{35}} & \frac{N_{35}}{N_{33}+N_{34}+N_{35}} \\ 0 & 0 & 0 & \frac{N_{44}}{N_{44}+N_{45}} & \frac{N_{45}}{N_{44}+N_{45}} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

2.2 Part II

2.2.1 Formulation

Consider a Hidden Markov Model where the observed variable $\boldsymbol{x}_t$ has 5 different states and the latent variable $\boldsymbol{z}_t$ has K different states:

$$p(\boldsymbol{z}_1|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{1,k}}$$
$$p(\boldsymbol{x}_t|\boldsymbol{z}_t) = \prod_{j=1}^{5} \prod_{k=1}^{K} b_{j,k}^{x_{t,j} z_{t,k}}$$
$$p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \prod_{j=1}^{K} \prod_{k=1}^{K} a_{jk}^{z_{t-1,j} z_{t,k}}$$

Consider a sequence D_l with length T

$$\begin{aligned} p(D_l) &= p(\boldsymbol{x}_1^l, \ldots, \boldsymbol{x}_T^l, \boldsymbol{z}_1^l, \ldots, \boldsymbol{z}_T^l) \\ &= \prod_{t=1}^{T} p(\boldsymbol{x}_t^l|\boldsymbol{z}_t^l) p(\boldsymbol{z}_1^l|\boldsymbol{\pi}) \prod_{t=2}^{T} p(\boldsymbol{z}_t^l|\boldsymbol{z}_{t-1}^l) \end{aligned}$$

$$= \prod_{t=1}^{T}\prod_{j=1}^{5}\prod_{k=1}^{K} b_{j,k}^{x_{t,j}z_{t,k}} \prod_{k=1}^{K} \pi_k^{z_{1,k}} \prod_{t=2}^{T}\prod_{j=1}^{K}\prod_{k=1}^{K} a_{jk}^{z_{t-1,j}z_{t,k}}$$

Since we are given N sequences of length T, we have the following

$$\begin{aligned} P(D_1,\ldots,D_N) &= \prod_{l=1}^{N} p(D_l) \\ &= \prod_{l=1}^{N}\left(\prod_{t=1}^{T}\prod_{j=1}^{5}\prod_{k=1}^{K} b_{j,k}^{x_{t,j}^l z_{t,k}^l} \prod_{k=1}^{K} \pi_k^{z_{1,k}^l} \prod_{t=2}^{T}\prod_{j=1}^{K}\prod_{k=1}^{K} a_{jk}^{z_{t-1,j}^l z_{t,k}^l}\right) \end{aligned}$$

Take log on the likelihood function

$$\ln P(D_1,\ldots,D_N) = \sum_{l=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{5}\sum_{k=1}^{K} x_{t,j}^l z_{t,k}^l \ln b_{j,k} + \sum_{l=1}^{N}\sum_{k=1}^{K} z_{1,k}^l \ln \pi_k + \sum_{l=1}^{N}\sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K} z_{t-1,j}^l z_{t,k}^l \ln a_{jk}$$

2.2.2 The Expectation Step

Take the expectation on the above log likelihood

$$\begin{aligned} \ell &= \mathbb{E}[\ln P(D_1,\ldots,D_N)] \\ &= \sum_{l=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{5}\sum_{k=1}^{K} \mathbb{E}\left[z_{t,k}^l\right] x_{t,j}^l \ln b_{j,k} + \sum_{l=1}^{N}\sum_{k=1}^{K} \mathbb{E}\left[z_{1,k}^l\right] \ln \pi_k + \sum_{l=1}^{N}\sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K} \mathbb{E}\left[z_{t-1,j}^l z_{t,k}^l\right] \ln a_{jk} \end{aligned}$$

Hence we need to compute

$$p(\boldsymbol{z}_t^l|\boldsymbol{x}_1^l,\ldots,\boldsymbol{x}_T^l) = \frac{p(\boldsymbol{x}_{1:T}^l|\boldsymbol{z}_t^l)p(\boldsymbol{z}_t^l)}{p(\boldsymbol{x}_{1:T}^l)} = \frac{p(\boldsymbol{x}_{1:t}^l|\boldsymbol{z}_t^l)p(\boldsymbol{x}_{t+1:T}^l|\boldsymbol{z}_t^l)p(\boldsymbol{z}_t^l)}{p(\boldsymbol{x}_{1:T}^l)} = \frac{p(\boldsymbol{x}_{1:t}^l,\boldsymbol{z}_t^l)p(\boldsymbol{x}_{t+1:T}^l|\boldsymbol{z}_t^l)}{p(\boldsymbol{x}_{1:T}^l)} = \frac{\alpha(\boldsymbol{z}_t^l)\beta(\boldsymbol{z}_t^l)}{p(\boldsymbol{x}_{1:T}^l)}$$

We can compute the $\alpha(\boldsymbol{z}_t^l)$ and $\beta \boldsymbol{z}_t^l$ recursively.

$$\begin{aligned} \alpha(\boldsymbol{z}_t^l) &= p(\boldsymbol{x}_{1:t}^l,\boldsymbol{z}_t^l) \\ &= p(\boldsymbol{x}_{1:t}^l|\boldsymbol{z}_t^l)p(\boldsymbol{z}_t^l) \\ &= p(\boldsymbol{x}_t^l|\boldsymbol{z}_t^l)p(\boldsymbol{x}_{1:t-1}^l|\boldsymbol{z}_t^l)p(\boldsymbol{z}_t^l) \\ &= p(\boldsymbol{x}_t^l|\boldsymbol{z}_t^l)\sum_{\boldsymbol{z}_{t-1}^l} p(\boldsymbol{x}_{1:t-1}^l,\boldsymbol{z}_{t-1}^l,\boldsymbol{z}_t^l) \\ &= p(\boldsymbol{x}_t^l|\boldsymbol{z}_t^l)\sum_{\boldsymbol{z}_{t-1}^l} p(\boldsymbol{x}_{1:t-1}^l,\boldsymbol{z}_t^l|\boldsymbol{z}_{t-1}^l)p(\boldsymbol{z}_{t-1}^l) \\ &= p(\boldsymbol{x}_t^l|\boldsymbol{z}_t^l)\sum_{\boldsymbol{z}_{t-1}^l} p(\boldsymbol{x}_{1:t-1}^l|\boldsymbol{z}_{t-1}^l)p(\boldsymbol{z}_t^l|\boldsymbol{z}_{t-1}^l)p(\boldsymbol{z}_{t-1}^l) \\ &= p(\boldsymbol{x}_t^l|\boldsymbol{z}_t^l)\sum_{\boldsymbol{z}_{t-1}^l} p(\boldsymbol{x}_{1:t-1}^l,\boldsymbol{z}_{t-1}^l)p(\boldsymbol{z}_t^l|\boldsymbol{z}_{t-1}^l) \end{aligned}$$

$$= p(\boldsymbol{x}_t^l|\boldsymbol{z}_t^l)\sum_{\boldsymbol{z}_{t-1}^l}\alpha(\boldsymbol{z}_{t-1}^l)p(\boldsymbol{z}_t^l|\boldsymbol{z}_{t-1}^l)$$

The initial condition, $\alpha(\boldsymbol{z}_1)$ can be computed

$$\alpha(\boldsymbol{z}_1^l) = p(\boldsymbol{x}_1^l|\boldsymbol{z}_1^l)p(\boldsymbol{z}_1^l) = \prod_{k=1}^{K}\left[\pi_k p(\boldsymbol{x}_1^l|\boldsymbol{z}_1^l)\right]^{z_{1,k}^l}$$

Similarly, we can manipulate $\beta(\boldsymbol{z}_t)$

$$\begin{aligned}
\beta(\boldsymbol{z}_t^l) &= p(\boldsymbol{x}_{t+1:T}^l|\boldsymbol{z}_t^l)\\
&= \sum_{\boldsymbol{z}_{t+1}^l} p(\boldsymbol{x}_{t+1:T}^l, \boldsymbol{z}_{t+1}^l|\boldsymbol{z}_t^l)\\
&= \sum_{\boldsymbol{z}_{t+1}^l} p(\boldsymbol{x}_{t+1:T}^l|\boldsymbol{z}_{t+1}^l, \boldsymbol{z}_t^l)p(\boldsymbol{z}_{t+1}^l|\boldsymbol{z}_t^l)\\
&= \sum_{\boldsymbol{z}_{t+1}^l} p(\boldsymbol{x}_{t+1:T}^l|\boldsymbol{z}_{t+1}^l)p(\boldsymbol{z}_{t+1}^l|\boldsymbol{z}_t^l)\\
&= \sum_{\boldsymbol{z}_{t+1}^l} p(\boldsymbol{x}_{t+1}^l, \boldsymbol{x}_{t+2:T}^l|\boldsymbol{z}_{t+1})p(\boldsymbol{z}_{t+1}^l|\boldsymbol{z}_t^l)\\
&= \sum_{\boldsymbol{z}_{t+1}^l} p(\boldsymbol{x}_{t+2:T}^l|\boldsymbol{z}_{t+1}^l)p(\boldsymbol{x}_{t+1}^l|\boldsymbol{z}_{t+1}^l)p(\boldsymbol{z}_{t+1}^l|\boldsymbol{z}_t^l)\\
&= \sum_{\boldsymbol{z}_{t+1}^l} \beta(\boldsymbol{z}_{t+1}^l)p(\boldsymbol{x}_{t+1}^l|\boldsymbol{z}_{t+1}^l)p(\boldsymbol{z}_{t+1}^l|\boldsymbol{z}_t^l)
\end{aligned}$$

Initial condition for $\beta(\boldsymbol{z}_T^l)$ is $\beta(\boldsymbol{z}_T^l) = 1$.

We also need the joint probability between the latent variable (smoothed transition probability)

$$\begin{aligned}
p(\boldsymbol{z}_{t-1}^l, \boldsymbol{z}_t^l|\boldsymbol{x}_{1:T}^l) &= \frac{p(\boldsymbol{z}_{t-1}^l, \boldsymbol{z}_t^l, \boldsymbol{x}_{1:T}^l)}{p(\boldsymbol{x}_{1:T}^l)}\\
&= \frac{p(\boldsymbol{x}_{1:T}^l|\boldsymbol{z}_{t-1}^l, \boldsymbol{z}_t^l)p(\boldsymbol{z}_{t-1}^l, \boldsymbol{z}_t^l)}{p(\boldsymbol{x}_{1:T}^l)}\\
&= \frac{p(\boldsymbol{x}_{1:t-1}^l|\boldsymbol{z}_{t-1}^l)p(\boldsymbol{x}_t^l|\boldsymbol{z}_t^l)p(\boldsymbol{x}_{t+1:T}^l|\boldsymbol{z}_t^l)p(\boldsymbol{z}_t^l|\boldsymbol{z}_{t-1}^l)p(\boldsymbol{z}_{t-1}^l)}{p(\boldsymbol{x}_{1:T}^l)}\\
&= \frac{\alpha(\boldsymbol{z}_{t-1}^l)p(\boldsymbol{x}_t^l|\boldsymbol{z}_t^l)p(\boldsymbol{z}_t^l|\boldsymbol{z}_{t-1})^l\beta(\boldsymbol{z}_t^l)}{p(\boldsymbol{x}_{1:T}^l)}
\end{aligned}$$

$$\mathbb{E}\left[z_{1,k}^l\right] = \frac{\alpha(\boldsymbol{z}_1^l)\beta(\boldsymbol{z}_1^l)}{p(\boldsymbol{x}_{1:T}^l)}$$

$$\mathbb{E}\left[z_{t,k}^{l}\right] = \frac{\alpha(\boldsymbol{z}_k^l)\beta(\boldsymbol{z}_k^l)}{p(\boldsymbol{x}_{1:T}^l)}$$

$$\mathbb{E}\left[z_{t-1,j}^{l} z_{t,k}^{l}\right] = \frac{\alpha(\boldsymbol{z}_{t-1}^l)\left[\prod_{j=1}^{5} b_{j,k}^{x_{t,j}^l}\right] a_{jk}\beta(\boldsymbol{z}_t^l)}{p(\boldsymbol{x}_{1:T}^l)}$$

2.2.3 The Maximization Step

From the above, we have

$$\begin{aligned}\ell &= \mathbb{E}\left[\ln P(D_1,\ldots,D_N)\right]\\ &= \sum_{l=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{5}\sum_{k=1}^{K}\mathbb{E}\left[z_{t,k}^l\right]x_{t,j}^l \ln b_{j,k} + \sum_{l=1}^{N}\sum_{k=1}^{K}\mathbb{E}\left[z_{1,k}^l\right]\ln \pi_k + \sum_{l=1}^{N}\sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K}\mathbb{E}\left[z_{t-1,j}^l z_{t,k}^l\right]\ln a_{jk}\end{aligned}$$

Now, we can formulate the Lagrangian

$$\mathcal{L}(\boldsymbol{\pi},\boldsymbol{b},\boldsymbol{A}) = \ell - \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right) - \sum_{k=1}^{K}\gamma_k\left(\sum_{j=1}^{5} b_{jk} - 1\right) - \sum_{j=1}^{K}\mu_j\left(\sum_{k=1}^{K} a_{jk} - 1\right)$$

We take the derivative with respect to the parameters

$$\frac{\partial \mathcal{L}(\boldsymbol{\pi},\boldsymbol{b},\boldsymbol{A})}{\partial \pi_k} = \sum_{l=1}^{N}\frac{\mathbb{E}\left[z_{1,k}^l\right]}{\pi_k} - \lambda = 0 \Rightarrow \sum_{l=1}^{N}\mathbb{E}\left[z_{1,k}^l\right] = \lambda\pi_k$$

$$\sum_{k=1}^{K}\sum_{l=1}^{N}\mathbb{E}\left[z_{1,k}^l\right] = \lambda \Rightarrow \pi_k = \frac{\sum_{l=1}^{N}\mathbb{E}\left[z_{1,k}^l\right]}{\sum_{k=1}^{K}\sum_{l=1}^{N}\mathbb{E}\left[z_{1,k}^l\right]}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\pi},\boldsymbol{b},\boldsymbol{A})}{\partial b_{jk}} = \sum_{l=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[z_{t,k}^l\right]\frac{x_{t,j}^l}{b_{j,k}} - \gamma_k = 0 \Rightarrow \sum_{l=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[z_{t,k}^l\right]x_{t,j}^l = \gamma_k b_{j,k}$$

$$\sum_{l=1}^{N}\sum_{t=1}^{T}\left(\sum_{j=1}^{5}x_{t,j}^l\right)\mathbb{E}\left[z_{t,k}^l\right] = \gamma_k\sum_{j=1}^{5}b_{j,k} \Rightarrow b_{jk} = \frac{\sum_{l=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[z_{t,k}^l\right]x_{t,j}^l}{\sum_{l=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[z_{t,k}^l\right]}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\pi},\boldsymbol{b},\boldsymbol{A})}{\partial a_{jk}} = \sum_{l=1}^{N}\sum_{t=2}^{T}\frac{\mathbb{E}\left[z_{t-1,j}^l z_{t,k}^l\right]}{a_{jk}} - \mu_j = 0 \Rightarrow \sum_{l=1}^{N}\sum_{t=2}^{T}\mathbb{E}\left[z_{t-1,j}^l z_{t,k}^l\right] = \mu_j a_{jk}$$

$$\sum_{k=1}^{K}\sum_{l=1}^{N}\sum_{t=2}^{T}\mathbb{E}\left[z_{t-1,j}^l z_{t,k}^l\right] = \mu_j\sum_{k=1}^{K}a_{jk} \Rightarrow a_{jk} = \frac{\sum_{l=1}^{N}\sum_{t=2}^{T}\mathbb{E}\left[z_{t-1,j}^l z_{t,k}^l\right]}{\sum_{k=1}^{K}\sum_{l=1}^{N}\sum_{t=2}^{T}\mathbb{E}\left[z_{t-1,j}^l z_{t,k}^l\right]}$$