# Assignment 2: HMM for Categorical Data Sequences

Daniel Barrejón Moreno

January 28, 2019

## 1 Complete data log-likehood $l_c(\boldsymbol{\theta})$ for the N sequences

The expression of the complete data log-likelihood will have the following form

$$l_c(\boldsymbol{\theta}) = \log p(S, Y|\boldsymbol{\theta}) = \log \prod_{n=1}^{N} \left( p\left(s_1^n|\boldsymbol{\pi}\right) \prod_{t=2}^{T_n} p\left(s_t^n|s_{t-1}^n, \mathbf{A}\right) \right) \left( \prod_{t=1}^{T_n} p\left(\mathbf{y}_t^n|s_t^n, \mathbf{B}\right) \right), \quad (1)$$

where log operator is the Naperian logarithm, $S$ represents the hidden states of the model, $Y$ is the observed continuous sequence, $\mathbf{A}$ stands for the state transition probabilities,$B$ the observatoin emission probabilities and $\boldsymbol{\pi}$ is the initial state probability distribution.

The parameters of the model are

$$\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}\}, \quad (2)$$

where $\mathbf{A}$ is the state transition matrix, $\mathbf{B}$ is the emission matrix and $\boldsymbol{\pi}$ represents the probability of each state. Equation 1 is composed by three main terms: the $\boldsymbol{\pi}$, $\mathbf{A}$ and $\mathbf{B}$ ones which can be rewritten as

$$p\left(s_1^n|\boldsymbol{\pi}\right) = \prod_{k=1}^{K} \pi_k^{\mathbb{I}\{s_1^n = k|Y,\boldsymbol{\theta}\}}, \quad (3)$$

$$p\left(s_t^n|s_{t-1}^n, \mathbf{A}\right) = \prod_{k=1}^{K} \prod_{k'=1}^{K} a_{k,k'}^{\mathbb{I}\{s_{t-1}^n = k, s_t^n = k'|Y,\boldsymbol{\theta}\}}, \quad (4)$$

$$p\left(\mathbf{y}_t^n|s_t^n, \mathbf{B}\right) = \prod_{k=1}^{K} p\left(\mathbf{y}_t^n|\boldsymbol{b}_k\right) = \prod_{k=1}^{K} p\left(\mathbf{y}_t^n|\boldsymbol{\theta}_k\right). \quad (5)$$

In this case, $\mathbb{I}$ represents an indicator function, $k = \{1, ..., K\}$ the current latent state of the model, $a_{k,k'}$ the $k$th row and $k'$th column element of the forementioned matrix $\mathbf{A}$ and $t = \{1, ..., T_n\}$ the position of the state in sequence $n$. Please notice that now, $\mathbf{b}_k$ (that belonged to $\mathbf{B}$) becomes $\boldsymbol{\theta}_k$, which denotes the hyperparameters of the categorical distribution that the data follow.

Since our data follow a categorical distribution, Equation 5 can be expressed as

$$p\left(\mathbf{y}_t^n|\boldsymbol{\theta}_k\right)=\prod_{j=1}^{Dt}\mathrm{Cat}(y_{j,t}|\boldsymbol{\theta}_k). \tag{6}$$

With a further development we can get

$$p\left(\mathbf{y}_t^n|\boldsymbol{\theta}_k\right)=\prod_{j=1}^{Dt}\mathrm{Cat}(y_{j,t}|\boldsymbol{\theta}_k)=\prod_{m=1}^{I}\prod_{j=1}^{Dt}\theta_{k,m}^{\mathbb{I}\{y_{j,t}^n=m\}}=\prod_{m=1}^{I}\theta_{k,m}^{\sum_{j=1}^{Dt}\mathbb{I}\{y_{j,t}^n=m\}}=\prod_{m=1}^{I}\theta_{k,m}^{\mu_{t,m}^n}, \tag{7}$$

being $\mu_{t,m}^n$

$$\mu_{t,m}^n=\sum_{j=1}^{Dt}\mathbb{I}\{y_{j,t}^n=m\}. \tag{8}$$

The final expression of the complete data log-likelihood can be expressed in the following form

$$\begin{aligned}l_c\left(\boldsymbol{\theta}\right)=&\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{I}\{s_1^n=k|Y,\boldsymbol{\theta}\}\log(\pi_k)+\\&+\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{k'=1}^{K}\sum_{t=1}^{T_n}\mathbb{I}\{s_{t-1}^n=k,s_t^n=k'|Y,\boldsymbol{\theta}\}\log(a_{k,k'})+\\&+\sum_{n=1}^{N}\sum_{t=1}^{T_n}\mathbb{I}\{s_t^n=k|Y,\boldsymbol{\theta}\}\sum_{m=1}^{I}\mu_{t,m}^n\log(\theta_{k,m}).\end{aligned} \tag{9}$$

# 2 Expected Complete Data Logl-likelihood $Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{t-1}\right)$

The expected complete data log-likelihood has the following form

$$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{t-1}\right)=E\left\{l_c(\boldsymbol{\theta})|\mathcal{D},\boldsymbol{\theta}^{t-1}\right\} \tag{10}$$

As in the previous section, and considering the nature of $l_c$ and its three main components, this expectation calculation can be divided in three.

$$\mathbb{E}\left(\sum_{n=1}^{N}\mathbb{I}\left(s_1^n=k|Y,\boldsymbol{\theta}\right)\right)=\sum_{n=1}^{N}\gamma_{n,1}(k), \tag{11}$$

$$\mathbb{E}\left(\sum_{n=1}^{N}\sum_{t=2}^{T_n}\mathbb{I}\left(s_{t-1}^n=k,s_t^n=k'|Y,\boldsymbol{\theta}\right)\right)=\sum_{n=1}^{N}\sum_{t=2}^{T_n}\xi_{n,t}(k,k'), \tag{12}$$

$$\mathbb{E}\left(\sum_{n=1}^{N}\sum_{t=1}^{T_n}\mathbb{I}\left(s_t^n=k|Y,\boldsymbol{\theta}\right)\right)=\sum_{n=1}^{N}\sum_{t=1}^{T_n}\gamma_{n,t}(k). \tag{13}$$

Where

$$\sum_{k=1}^{K} \gamma_{n,t}(k) = 1. \tag{14}$$

Being $\xi_{n,t}(k,k')$

$$\xi_{n,t}(k,k') = \alpha_{t-1}^{n}(k) a_{k,k'} \prod_{m=1}^{I} \theta_{k,m}^{\mu_{t,m}^{n}} \beta_{t}^{n}(k'), \tag{15}$$

and $\gamma_{n,t}(k)$

$$\gamma_{n,t}(k) \propto \beta_{t}^{n}(k) \alpha_{t}^{n}(k) \tag{16}$$

The terms $\alpha$ and $\beta$ are computed by means of the **forward-backward algorithm** as follows

$$\alpha_{1}^{n}(k) = \pi_k \prod_{m=1}^{I} \theta_{k,m}^{\mu_{1,m}^{n}}, \tag{17}$$

$$\alpha_{t}^{n}(k) = \left( \sum_{k'=1}^{K} \alpha_{t-1}^{n}(k') a_{k',k} \right) \prod_{m=1}^{I} \theta_{k,m}^{\mu_{t,m}^{n}}, \tag{18}$$

$$\beta_{T_n}^{n}(k) = 1, \tag{19}$$

$$\beta_{t}^{n}(k) = \sum_{k'=1}^{K} a_{k,k'} \prod_{m=1}^{I} \theta_{k',m}^{\mu_{t+1,m}^{n}} \beta_{t+1}^{n}(k'). \tag{20}$$

So the complete expression for $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}\right)$ is now

$$\begin{aligned} Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}\right) = & \sum_{k=1}^{K} \sum_{n=1}^{N} \gamma_{n,1}(k) \log(\pi_k) + \\ & + \sum_{k=1}^{K} \sum_{k'=1}^{K} \sum_{n=1}^{N} \sum_{t=2}^{T_n} \xi_{n,t}(k,k') \log(a_{k,k'}) + \\ & + \sum_{k=1}^{K} \sum_{m=1}^{I} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma_{n,t}(k) \mu_{t,m}^{n} \log(\theta_{k,m}). \end{aligned} \tag{21}$$

# 3 ML Inference

There are three parameters for the model, which are $\mathbf{A}$, $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$, and they are computed by means of Lagrange multipliers.

## 3.1 ML estimation of $\pi_k$

In order to compute $\pi_k$, we first need to take into account the following restrictions

$$0 \leq \pi_k \leq 1 \tag{22}$$

$$\sum_{k=1}^{K} \pi_k = 1. \tag{23}$$

Now, let us define the lagrangian as

$$L\left(Q(\pi_k),\lambda\right) = Q(\pi_k) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right), \tag{24}$$

which will be optimized in this way

$$\min_{\lambda}\max_{\pi_k}\{L\left(Q(\pi_k),\lambda\right)\}. \tag{25}$$

By first taking the derivative with respect to $\pi_k$ and equating it to 0

$$\begin{aligned}\frac{\partial L}{\partial \pi_k} &= \sum_{n=1}^{N}\frac{\gamma_{n,1}(k)}{\pi_k} - \lambda = 0,\\ \pi_k &= \frac{1}{\lambda}\sum_{n=1}^{N}\gamma_{n,1}(k).\end{aligned} \tag{26}$$

And later with respect to $\lambda$

$$\begin{aligned}\frac{\partial L}{\partial \lambda} &= \sum_{k=1}^{K}\pi_k - 1 = 0,\\ \sum_{k=1}^{K}\pi_k &= 1.\end{aligned} \tag{27}$$

Taking into account the previous equation, if in both sides of Equation 26 summatories all over $K$ are taken, then the value of $\lambda$ can be obtained

$$\begin{aligned}\sum_{k=1}^{K}\pi_k &= \sum_{k=1}^{K}\frac{1}{\lambda}\sum_{n=1}^{N}\gamma_{n,1}(k)\\ 1 &= \frac{1}{\lambda}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{n,1}(k).\end{aligned} \tag{28}$$

Now, considering the previously imposed restrictions and recalling that

$$\sum_{k=1}^{K}\gamma_{n,t}(k) = 1, \tag{29}$$

the value of $\lambda$ is

$$\begin{aligned}1 &= \frac{1}{\lambda}\sum_{n=1}^{N}1.\\ \lambda &= N.\end{aligned} \tag{30}$$

And with that value of $\lambda$ the estimated value of $\pi_k$ is

$$\widehat{\pi}_k = \frac{1}{N}\sum_{n=1}^{N}\gamma_{n,1}(k). \tag{31}$$

### 3.2 ML estimation of $\theta_{k,m}$

Now, the parameter to be estimated is $\theta_{k,m}$ and the constraint is now

$$\sum_{m=1}^{I} \theta_{k,m} = 1. \tag{32}$$

Being the whole expression

$$L\left(Q(\theta_{k,m}), \lambda\right) = Q(\theta_{k,m}) + \lambda \left( \sum_{m=1}^{I} \theta_{k,m} - 1 \right), \tag{33}$$

which will be optimized in this way

$$\min_{\lambda} \max_{\theta_{k,m}} \{ L\left(Q(\theta_{k,m}), \lambda\right) \}. \tag{34}$$

By first taking the derivative with respect to $\theta_{k,m}$ and equating it to 0

$$\begin{aligned} \frac{\partial L}{\partial \theta_{k,m}} &= \sum_{n=1}^{N} \sum_{t=1}^{T_n} \frac{\gamma_{n,t}(k) \mu_{t,m}^{n}}{\theta_{k,m}} - \lambda = 0, \\ \theta_{k,m} &= \frac{1}{\lambda} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma_{n,t}(k) \mu_{t,m}^{n}. \end{aligned} \tag{35}$$

And later with respect to $\lambda$

$$\begin{aligned} \frac{\partial L}{\partial \lambda} &= \sum_{m=1}^{I} \theta_{k,m} - 1 = 0, \\ \sum_{m=1}^{I} \theta_{k,m} &= 1. \end{aligned} \tag{36}$$

Taking into account the previous equation, if in both sides of Equations 35 summatories all over $I$ are taken, then the value of $\lambda$ can be obtained

$$\begin{aligned} \sum_{m=1}^{I} \theta_{k,m} &= \sum_{m=1}^{I} \frac{1}{\lambda} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma_{n,t}(k) \mu_{t,m}^{n} \\ 1 &= \frac{1}{\lambda} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{m=1}^{I} \gamma_{n,t}(k) \mu_{t,m}^{n}, \\ \lambda &= \sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{m=1}^{I} \gamma_{n,t}(k) \mu_{t,m}^{n}. \end{aligned} \tag{37}$$

And with that value of $\lambda$ the estimated value of $\theta_{k,m}$ is

$$\widehat{\theta}_{k,m} = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma_{n,t}(k) \mu_{t,m}^{n}}{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{m=1}^{I} \gamma_{n,t}(k) \mu_{t,m}^{n}}. \tag{38}$$

### 3.3 ML estimation of $a_{k,k'}$

At last, the parameter to be estimated is $a_{k,k'}$ and the constraint is now

$$\sum_{k'=1}^{K} a_{k,k'} = 1. \tag{39}$$

Being the whole expression

$$L\left(Q(a_{k,k'}), \lambda\right) = Q(a_{k,k'}) + \lambda \left( \sum_{k'=1}^{K} a_{k,k'} - 1 \right), \tag{40}$$

which will be optimized in this way

$$\min_{\lambda} \max_{a_{k,k'}} \{ L\left(Q(a_{k,k'}), \lambda\right) \}. \tag{41}$$

By first taking the derivative with respect to $a_{k,k'}$ and equating it to 0

$$\begin{aligned} \frac{\partial L}{\partial a_{k,k'}} &= \sum_{n=1}^{N} \sum_{t=2}^{T_n} \frac{\xi_{n,t}(kk')}{a_{k,k'}} - \lambda = 0, \\ a_{k,k'} &= \frac{1}{\lambda} \sum_{n=1}^{N} \sum_{t=2}^{T_n} \xi_{n,t}(kk'). \end{aligned} \tag{42}$$

And later with respect to $\lambda$

$$\begin{aligned} \frac{\partial L}{\partial \lambda} &= \sum_{m=1}^{I} a_{k,k'} - 1 = 0, \\ \sum_{k'=1}^{K} a_{k,k'} &= 1. \end{aligned} \tag{43}$$

Taking into account the previous equation, if in both sides of (42) summatories all over $I$ are taken, then the value of $\lambda$ can be obtained

$$\begin{aligned} \sum_{k'=1}^{K} a_{k,k'} &= \sum_{k'=1}^{K} \frac{1}{\lambda} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \xi_{n,t}(k,k') \\ 1 &= \frac{1}{\lambda} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{k'=1}^{K} \xi_{n,t}(k,k'), \\ \lambda &= \sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{k'=1}^{K} \xi_{n,t}(k,k'). \end{aligned} \tag{44}$$

And with that value of $\lambda$ the estimated value of $a_{k,k'}$ is

$$\widehat{a}_{k,k'} = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \xi_{n,t}(k,k')}{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{k'=1}^{K} \xi_{n,t}(k,k')}. \tag{45}$$

### 3.4 MAP decoding based on Forward-Backwards

In order to determine the state of each document for the different sequences we obtain the most probable state for each document at each sequence from the parameter $\gamma_{n,t}(k)$ which is computed in the E-step of the Baum-Welch EM using Equation 16.

### 3.5 ML Viterbi decoding

As a comparison for the MAP decoding comment above, we have also implemented a Viterbi decoder. The Viterbi decoder is a dynamic programming algorithm for finding the most likely sequence of hidden states result of a sequence of observed events. Since we are dealing with hidden and observed states with HMM, the algorithm suits perfectly to our problem. The Viterbi decoding algorithm looks as follows

$$\underset{S}{\text{argmax}} p(S,Y) = \underset{i}{\text{argmax}} \left\{ \underset{s_{1:T-1}}{\text{argmax}} p\left(s_T = i, s_{1:T-1}, Y\right) \right\}. \tag{46}$$

In particular, we will use an iterative implementation of the Viterbi Algorithm. For such implementation we define two iterative steps described below.

- **Forward step**: Computed with the following equations

$$\delta_1(k) = \pi_k \prod_{m=1}^{I} \theta_{k,m}^{\mu_{1,m}^n} \quad 1 \leq k \leq K \tag{47}$$

$$\delta_t(k) = \prod_{m=1}^{I} \theta_{k,m}^{\mu_{t,m}^n} \max_{k'} a_{k',k} \delta_{t-1}(k') \quad 1 \leq k \leq K, 1 \leq t \leq T \tag{48}$$

$$\varphi_t(k) = \underset{k'}{\text{argmax}}\, a_{k',k} \delta_{t-1}(k') \quad 1 \leq k \leq K, 1 \leq t \leq T \tag{49}$$

$$\tag{50}$$

- **Backwards step**: The state estimation is computed using the following Equations

$$\hat{\mathbf{s}}_T = \arg\max_{\boldsymbol{k}} \delta_T(k) \tag{51}$$

$$\hat{s}_t = \varphi_{t+1}\left(\hat{s}_{t+1}\right) 1 \leq t \leq T. \tag{52}$$

## 4 Experiments

For the experiments...

One important aspect about the implementation of the algorithm is that instead of working with the actual $\xi_{n,t}(k,k')$ we work with a similar implementation used in **murphy2012machine** which consider working with the expected sufficient statistics for the transition matrix, for a given observation sequence. This may be rewritten as follows

$$\xi_{\Sigma}(k,k') = \sum_{t=2}^{T} p(S(t) = k, S(t+1) = k' | y(1:T)), \tag{53}$$

where the subscript $\Sigma$ indicates the sum over $t$. Notice that, for a given sequence, this matrix is no longer a tensor but a matrix of dimension $K \times K$.