

Assignment 2: HMM for Categorical Data Sequences

Miguel Fernández Díaz

January 16, 2019

1 Expression for the complete data log-likelihood for the N sequences

According to this expression for the complete data log-likelihood

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

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, \mathbf{B} the observatoin emission probabilites and $\boldsymbol{\pi}$ is the initial state probability distribution.

The parameters of the model are

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

Equation (1) is composed by three main terms: the $\boldsymbol{\pi}$, \mathbf{A} and \mathbf{B} ones which can be rewritten as

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

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

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

In this case, \mathbb{I} represents an indicator function, $k = \{1, \dots, K\}$ 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} , $t = \{1, \dots, T_n\}$ 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

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

With a further development we can get

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

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

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

So the complete data log-likelihood is written in this way

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

2 Expression for the expected complete data log likelihood

The expected complete data log-likelihood has the following form

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

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.

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

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

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

Where

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

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

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

The terms α and β are computed by means of the forward-backward algorithm as follows

$$\alpha_1^n(k) = \pi_k \prod_{m=1}^I \theta_{k,m}^{\mu_{1,m}^n}, \quad (16)$$

$$\alpha_t^n(k) = \left(\sum_{k'=1}^K \alpha_{t-1}^n(k') a_{k',k} \right) \prod_{m=1}^I \theta_{k,m}^{\mu_{t,m}^n}, \quad (17)$$

$$\beta_{T_n}^n(k) = 1, \quad (18)$$

$$\beta_t^n(k) = \sum_{k'=1}^K a_{k,k'} \prod_{m=1}^I \theta_{k',m}^{\mu_{t+1,m}^n} \beta_{t+1}^n(k'). \quad (19)$$

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

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}) &= \sum_{k=1}^K \sum_{n=1}^N \gamma_{n,1}(k) \log(\pi_k) + \\ &+ \sum_{k=1}^K \sum_{k'=1}^K \sum_{n=1}^N \sum_{t=2}^{T_n} \xi_{n,t}(k, k') \log(a_{k,k'}) + \\ &+ \sum_{k=1}^K \sum_{m=1}^I \sum_{n=1}^N \sum_{t=1}^{T_n} \gamma_{n,t}(k) \mu_{t,m}^n \log(\theta_{k,m}). \end{aligned} \quad (20)$$

3 Maximum Likelihood estimation of the parameters of the model

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 π_k

In order to compute π_k , we first need to take into account the following restrictions

$$0 \leq \pi_k \leq 1 \quad (21)$$

$$\sum_{k=1}^K \pi_k = 1. \quad (22)$$

Now, let us define the lagrangian as

$$L(Q(\pi_k), \lambda) = Q(\pi_k) + \lambda \left(\sum_{k=1}^K \pi_k - 1 \right), \quad (23)$$

which will be optimized in this way

$$\min_{\lambda} \max_{\pi_k} \{L(Q(\pi_k), \lambda)\}. \quad (24)$$

By first taking the derivative with respect to π_k and equating it to 0

$$\begin{aligned} \frac{\partial L}{\partial \pi_k} &= 0 = \sum_{n=1}^N \frac{\gamma_{n,1}(k)}{\pi_k} - \lambda, \\ \pi_k &= \frac{1}{\lambda} \sum_{n=1}^N \gamma_{n,1}(k). \end{aligned} \quad (25)$$

And later with respect to λ

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

Taking into account the previous equation, if in both sides of (25) summatories all over K are taken, then the value of λ can be obtained

$$\begin{aligned}\sum_{k=1}^K \pi_k &= \sum_{k=1}^K \frac{1}{\lambda} \sum_{n=1}^N \gamma_{n,1}(k) \\ 1 &= \frac{1}{\lambda} \sum_{n=1}^N \sum_{k=1}^K \gamma_{n,1}(k).\end{aligned}\tag{27}$$

Now, considering the previously imposed restrictions and recalling that

$$\sum_{k=1}^K \gamma_{n,t}(k) = 1,\tag{28}$$

the value of λ is

$$\begin{aligned}1 &= \frac{1}{\lambda} \sum_{n=1}^N 1. \\ \lambda &= N.\end{aligned}\tag{29}$$

And with that value of λ the estimated value of π_k is

$$\hat{\pi}_k = \frac{1}{N} \sum_{n=1}^N \gamma_{n,1}(k).\tag{30}$$

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

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

$$\sum_{m=1}^I \theta_{k,m} = 1.\tag{31}$$

Being the whole expression

$$L(Q(\theta_{k,m}), \lambda) = Q(\theta_{k,m}) + \lambda \left(\sum_{m=1}^I \theta_{k,m} - 1 \right),\tag{32}$$

which will be optimized in this way

$$\min_{\lambda} \max_{\theta_{k,m}} \{L(Q(\theta_{k,m}), \lambda)\}.\tag{33}$$

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

$$\begin{aligned}\frac{\partial L}{\partial \theta_{k,m}} &= 0 = \sum_{n=1}^N \sum_{t=1}^{T_n} \frac{\gamma_{n,t}(k) \mu_{t,m}^n}{\theta_{k,m}} - \lambda, \\ \theta_{k,m} &= \frac{1}{\lambda} \sum_{n=1}^N \sum_{t=1}^{T_n} \gamma_{n,t}(k) \mu_{t,m}^n.\end{aligned}\tag{34}$$

And later with respect to λ

$$\begin{aligned}\frac{\partial L}{\partial \lambda} &= 0 = \sum_{m=1}^I \theta_{k,m} - 1, \\ \sum_{m=1}^I \theta_{k,m} &= 1.\end{aligned}\tag{35}$$

Taking into account the previous equation, if in both sides of (34) summatories all over I are taken, then the value of λ can be obtained

$$\begin{aligned}\sum_{m=1}^I \theta_{k,m} &= \sum_{m=1}^I \frac{1}{\lambda} \sum_{n=1}^N \sum_{t=1}^{T_n} \gamma_{n,t}(k) \mu_{t,m}^n \\ 1 &= \frac{1}{\lambda} \sum_{n=1}^N \sum_{t=1}^{T_n} \sum_{m=1}^I \gamma_{n,t}(k) \mu_{t,m}^n, \\ \lambda &= \sum_{n=1}^N \sum_{t=1}^{T_n} \sum_{m=1}^I \gamma_{n,t}(k) \mu_{t,m}^n.\end{aligned}\tag{36}$$

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

$$\hat{\theta}_{k,m} = \frac{\sum_{n=1}^N \sum_{t=1}^{T_n} \gamma_{n,t}(k) \mu_{t,m}^n}{\sum_{n=1}^N \sum_{t=1}^{T_n} \sum_{m=1}^I \gamma_{n,t}(k) \mu_{t,m}^n}.\tag{37}$$

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

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

$$\sum_{k'=1}^K a_{k,k'} = 1.\tag{38}$$

Being the whole expression

$$L(Q(a_{k,k'}), \lambda) = Q(a_{k,k'}) + \lambda \left(\sum_{k'=1}^K a_{k,k'} - 1 \right),\tag{39}$$

which will be optimized in this way

$$\min_{\lambda} \max_{a_{k,k'}} \{L(Q(a_{k,k'}), \lambda)\}.\tag{40}$$

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

$$\begin{aligned}\frac{\partial L}{\partial a_{k,k'}} &= 0 = \sum_{n=1}^N \sum_{t=2}^{T_n} \frac{\xi_{n,t}(kk')}{a_{k,k'}} - \lambda, \\ a_{k,k'} &= \frac{1}{\lambda} \sum_{n=1}^N \sum_{t=2}^{T_n} \xi_{n,t}(kk').\end{aligned}\tag{41}$$

And later with respect to λ

$$\begin{aligned}\frac{\partial L}{\partial \lambda} &= 0 = \sum_{m=1}^I a_{k,k'} - 1, \\ \sum_{k'=1}^K a_{k,k'} &= 1.\end{aligned}\tag{42}$$

Taking into account the previous equation, if in both sides of (41) summatories all over I are taken, then the value of λ can be obtained

$$\begin{aligned}\sum_{k'=1}^K a_{k,k'} &= \sum_{k'=1}^K \frac{1}{\lambda} \sum_{n=1}^N \sum_{t=1}^{T_n} \xi_{n,t}(k, k') \\ 1 &= \frac{1}{\lambda} \sum_{n=1}^N \sum_{t=1}^{T_n} \sum_{k'=1}^K \xi_{n,t}(k, k'), \\ \lambda &= \sum_{n=1}^N \sum_{t=1}^{T_n} \sum_{k'=1}^K \xi_{n,t}(k, k').\end{aligned}\tag{43}$$

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

$$\hat{a}_{k,k'} = \frac{\sum_{n=1}^N \sum_{t=1}^{T_n} \xi_{n,t}(k, k')}{\sum_{n=1}^N \sum_{t=1}^{T_n} \sum_{k'=1}^K \xi_{n,t}(k, k')}.\tag{44}$$