

Assignment 2: HMM for Categorical Data Sequences

Daniel Barrejón Moreno

January 28, 2019

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

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

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

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 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

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

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

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

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} and $t = \{1, \dots, T_n\}$ 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 $\mu_{t,m}^n$

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

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

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

2 Expected Complete Data Log-likelihood $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1})$

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.

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

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

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

Where

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

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

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

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

$$\gamma_{n,t}(k) \propto \beta_t^n(k) \alpha_t^n(k) \quad (16)$$

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

$$\alpha_1^n(k) = \pi_k \prod_{m=1}^I \theta_{k,m}^{\mu_{1,m}^n}, \quad (17)$$

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

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

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

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

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

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 π_k

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

$$0 \leq \pi_k \leq 1 \quad (22)$$

$$\sum_{k=1}^K \pi_k = 1. \quad (23)$$

Now, let us define the lagrangian as

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

which will be optimized in this way

$$\min_{\lambda} \max_{\pi_k} \{L(Q(\pi_k), \lambda)\}. \quad (25)$$

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

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

And later with respect to λ

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

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

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

Now, considering the previously imposed restrictions and recalling that

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

the value of λ is

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

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

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

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

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

$$\sum_{m=1}^I \theta_{k,m} = 1. \quad (32)$$

Being the whole expression

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

which will be optimized in this way

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

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

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

And later with respect to λ

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

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

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

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

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

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

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

$$\sum_{k'=1}^K a_{k,k'} = 1. \quad (39)$$

Being the whole expression

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

which will be optimized in this way

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

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

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

And later with respect to λ

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

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

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

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

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

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

$$\operatorname{argmax}_S p(S, Y) = \operatorname{argmax}_i \left\{ \operatorname{argmax}_{s_{1:T-1}} p(s_T = i, s_{1:T-1}, Y) \right\}. \quad (46)$$

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

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

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

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

$$(50)$$

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

$$\hat{s}_T = \operatorname{argmax}_k \delta_T(k) \quad (51)$$

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

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

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

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