521 Homework 5

Evan Hofmeister

May 5, 2018

Question 1a)

(a) (2pts) Show that if an HMM transition matrix A and emission matrix B are initialized to uniformly constant values, then the EM algorithm fails to update the parameters meaningfully.

Solution:

Note define the below distributions:

$$\begin{aligned} Transition&:\\ &A_{i',i} = p(h_{t+1} = i'|h_t = i)\\ Initial&:\\ &a_i = p(h_1 = i)\\ Emission&:\\ &B_{i,j} = p(v_t = i|h_t = j) \end{aligned}$$

M-step

Assuming i.i.d. data, the M-step is given by maximising the ‘energy’ component:

$$\sum_{n=1}^{N}\langle \log p(v_1^n,\ v_2^n \ldots,\ v_{T^n}^n,\ h_1^n,\ h_2^n,\ \ldots,\ h_{T^n}^n)\rangle_{p^{old}(\mathrm{h}^n|\mathrm{v}^n)}$$

with respect to the parameters $\mathrm{A}, \mathrm{B}, \mathrm{a}$; h^n denotes $h_{1:T_n}$ (to write formulas more easily). We can write the HMM as follows:

$$\sum_{n=1}^{N}\{\langle\log p(h_1)\rangle_{p^{old}(h_1|\mathrm{v}^n)} + \sum_{t=1}^{T_n-1}\langle\log p(h_{t+1}|h_t)\rangle_{p^{old}(h_t,h_{t+1}|\mathrm{v}^n)} + \sum_{t=1}^{T_n}\langle\log p(v_t^n|h_t)\rangle_{p^{old}(h_t|\mathrm{v}^n)}\}$$

Note that $p^{new}(h_1 = i)$ denotes the (new) table entry for the probability that the initial hidden variable is in state i.

optimising the previous equation with respect to $p(h_1)$,($p(h_1)$ is a distribution) we find:

$$a_i^{new} \equiv p^{new}(h_1 = i) = \frac{1}{N}\sum_{n=1}^{N} p^{old}(h_1 = i|\mathrm{v}^n)$$

The M-step for the transition is:

$$A^{new}_{i,i} \equiv p^{new}(h_{t+1} = i'|h_t = i) \propto \sum_{n=1}^{N} \sum_{t=1}^{T_n-1} p^{old}(h_t = i,\ h_{t+1} = i'|\mathrm{v}^n)$$

which is the number of times that a transition from hidden state i to hidden state i' occurs, averaged over all times (since we assumed stationarity) and training sequences. Now if we normalize this result:

$$A^{new}_{i,i} = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n-1} p^{old}(h_t = i, h_{t+1} = i'|\mathrm{v}^n)}{\sum_{i'} \sum_{n=1}^{N} \sum_{t=1}^{T_n-1} p^{old}(h_t = i, h_{t+1} = i|\mathrm{v}^n)}$$

If we assume transition matrix A, and emission matrix B are initialized to iniformly constant values then, we know that the marginal joint conditional distributions from the previous, $p^{old}(h_t,\ h_{t+1}|v_{1:T})$ are uniform distribution. From this we can find for any $p^{new}(h_{t+1} = i'|h_t = i)$,

from this we can see that the M-step calculation of $A_{i,j}$ that because the conditional above, $A_{i,j}$ will not be calculated meaningfully. It turns out to only equate to the fractional of the number of hidden variables (H). Basically because the conditional distributions from the previous iteration turn out to have no influence on updating or calculation in the M-step, so our algorithm fails to move towards convergence.

The M-step update for the emission component is:

$$B^{new}_{j,i} \equiv p^{new}(v_t = j|h_t = i) \propto \sum_{n=1}^{N} \sum_{t=1}^{T_n} \mathrm{I}[v^n_t = j]p^{old}(h_t = i|\mathrm{v}^n)$$

which is the expected number of times that, for the observation being in state j, the hidden state is i. The proportionality constant is determined by the normalisation requirement.

From the Estep, we also know that with this condition, v_t is indipendent of h_t so B^{new} will update without being conditioned on any v_t meaning that the Mstep will not update anything and the EM algorithm will be broken. Esentially, The M-step in the EM algorithm will not contribute to finding the updated meaningful perameters, and so, we will not be able to find a solution.

Question 1b)

(b) (2pts) Exercise 23.11 of [B] textbook (A second-order HMM

Second order HMM defined as:

$$p(h_{1:T}, v_{1:T}) = p(h_1)p(v_1|h_1)p(v_2|h_2)p(h_2|h_1)\prod_{t=3}^{T} p(h_t|h_{t-1}, h_{t-2})p(v_t|h_t)$$

now maximize with respect to h_T:

$$
\begin{aligned}
& max_{h_T} p(h_1)p(v_1|h_1)p(v_2|h_2)p(h_2|h_1)\prod_{t=3}^{T} p(h_t|h_{t-1},h_{t-2})p(v_t|h_t) \\
& = p(h_1)p(v_1|h_1)p(v_2|h_2)p(h_2|h_1)\prod_{t=3}^{T-1} p(h_t|h_{t-1},h_{t-2})p(v_t|h_t) max_{h_T} p(h_T|h_{T-1},h_{T-2})p(v_T|h_T) \\
& \mu_{T-1}(h_{T-1}) = max_{h_T} p(h_T|h_{T-1},h_{T-2})p(v_T|h_T) \\
& \mu_{t-1}(h_{t-1}) = max_{h_t} p(h_t|h_{t-1},h_{t-2})p(v_t|h_t)\mu(h_t)
\end{aligned}
$$

if we de???ne $\mu T(h_T) := 1$. Thus, we obtain the following recursion to calculate ţ$_t(h_t)$:

$$
\begin{aligned}
& \mu_t(h_t) = max_{h_t+1} p(h_{t+1}|h_t,h_{t-1})p(v_{t+1}|h_{t+1})\mu_{t+1}(h_{t+1}) \\
& \mu T(h_T) := 1
\end{aligned}
$$

with $\mu(hT) = 1$. This means that the effect of maximising over $h_2, ..., h_T$ is compressed into a message $\mu(h1)$ so that the most likely state h_1^* is given by:

$$
\begin{aligned}
h_1^* &= p(h_1)p(v_1|h_1)\mu(h_1) \\
h_2^* &= p(h_2|h_1)p(v_2|h_2)\mu(h_2)
\end{aligned}
$$

In this second order example the message we are passing is now in the form of a matrix instead of a vector like in the previous one dimmentional examples we've covered.

Once we find h_{t-1}^* and h_{t-2}^* we can find h_t^* by:

$$
h_t^* = argmax_{h_t} p(h_t|h_{t-1}^*, h_{t-2}^*)p(v_t|h_t)\mu(h_t)
$$

h_{t-1}^* and h_{t-2}^* are found similarly to before, and backtracking is used. Then we have defined the optimal for h_t above. Basically just a system of recursion is used.

Question 1c) ———————————————————-

(c) (2pts) Exercise 23.13 of [B] textbook (Consider the HMM defined on hidden variables

We need to to esentially find the transition matrix from H to V with the given information. The problem gives us the form of $P(H|V)$.

We can define $P(H|V)$ as follows using the pairwise marginal formula defined in the textbook and in class.

We will also need the likelihood $p(v_{1:T})$ of a sequence of observations can be computed from:

$$
p(v_{1:T}) = \sum_{h_T} p(h_T,\ v_{1:T}) = \sum_{h_T} \alpha(h_T)\ (23.2.26)
$$

an explicit recursion is as follows:

$$
p(h_t,\ h_{t+1}|v_{1:T}) \propto p(v_{1:t},\ v_{t+1},\ v_{t+2:T},\ h_{t+1},\ h_t)/p(v_{1:T})
$$

$$p(h_t,\ h_{t+1}|v_{1:T}) \propto \alpha(h_t)p(v_{t+1}|h_{t+1})p(h_{t+1}|h_t)\beta(h_{t+1})/p(v_{1:T})$$
$$= p(h_t,\ h_{t+1}|v_{1:T}) \propto \alpha(h_t)p(v_{t+1}|h_{t+1})p(h_{t+1}|h_t)\beta(h_{t+1})/\sum_{h_T}\alpha(h_T))$$

Now given the form $P(V,H)$, we can wewrite the form of $P(H|V)$ is proportional to:

$$\prod_{t=1}^{T}\zeta(h_{t-1},h_t)\prod_{t=1}^{T}p(h_t|h_{t-1})p(v_t|h_t)$$

def: $\kappa = h_T$

We also then can infer:

$$\tilde{p}(h_T|h_{T-1}) = \zeta(h_{T-1},h_T)/\sum_{\kappa}\zeta(h_{T-1},h_T)$$

Now we can expand the previous where $\omega = \epsilon(1,T-1)$:

$$[\sum_{h_T}\zeta(h_{T-1},h_T)](\tilde{p}(h_T|h_{T-1}))\prod_{\omega}\zeta(h_{t-1},h_t)$$

Now given the form of $\tilde{p}$, we have enentially proved the posterior is a markov chain given that we have satisfied the given form.

Where as previous for the end case, we can define:

$$\tilde{p}(h_t|h_{t-1}) = \zeta(h_{t-1},h_t)\frac{\zeta(h_t,h_T)}{\sum_{\kappa}\zeta(h_{T-1},h_T)}$$

Question 2) ——————————————————————-

2. (4pts) Exercise 24.5 of the [B] textbook (Consider a supervised learning In part 2 of the exercise, you do not need to write a routine.

We can use MLE, more specifically Ordinary Least Sqeares to estimate W_t^T.

To do this, we just reference the basic OLS method:

Our regression is in the form $\hat{y} = W_i^T X_i$

$$S(W) = \sum_{i=1}^{n}(y - x_i^T W)^2 = (y - XW)^T(y - XW)$$

$$[\hat{W}_t] = argmin_{b\epsilon(R)}S(W) = \sum_{i=1}^{n}(y - x_i^T W)^2 = (y - XW)^T(y - XW)$$

and to find η_t^y we can simply:

$$\hat{\eta} = y - \hat{y} = y - W^T X$$

Next we can use reduced chi-squared to estimate σ_y^2:

$$S^2 = \hat{\eta}^T \hat{\eta}/(n-p)$$

This is the OLS estimate of σ^2 which is quite similar to the MLE estimator but is always unbiased regardless of sample size.

part ii)

We can think of this model as a typical latent linear dynamical system (LDS) and think of it as a state space model:

$$\begin{aligned} &Wt = W_{t-1} + \eta_t^w \\ &y_t = W_t^T x_t + \eta^y \\ &\eta_t^w | W_{t-1} \sim N(\tilde{w_t}, \Sigma_t^w) \\ &\eta_t^y | W_{t-1} \sim N(\tilde{y_t}, \Sigma_t^y) \end{aligned}$$

Now to find $E[W_t|D]$ we find f_t as defined in lecture and the textbook. We need to write a recursive relationship between f_t and f_{t-1}

• As in the case of HMMs, the filtering probability $p(\mathrm{w}_t|\mathrm{y}_{1:t})$ can be found by recursion:

$$p(\mathrm{w}_t|\mathrm{y}_{1:t}) = \frac{p(\mathrm{y}_{1:t}, \mathrm{w}_t)}{p(\mathrm{y}_{1:t})} = \frac{1}{p(\mathrm{y}_{1:t})} \int_{-\infty}^{\infty} p(\mathrm{y}_{1:t-1},\ \mathrm{w}_{t-1},\ \mathrm{y}_t,\ \mathrm{w}_t) \mathrm{dw}_{t-1}$$

$$= \frac{p(\mathrm{y}_{1:(t-1)})}{p(\mathrm{y}_{1:t})} \int_{-\infty}^{\infty} p(\mathrm{w}_{t-1}|\mathrm{y}_{1:t-1}) p(\mathrm{w}_t|\mathrm{w}_{t-1}) p(\mathrm{y}_t|\mathrm{w}_t) \mathrm{dw}_{t-1}.$$

• This can be seen as a recursive formula for $p(\mathrm{w}_t|\mathrm{y}_{1:t})$. However, it is not possible to directly use this relationship, as there are infinite values for w_t.

• We need to store the distribution $p(\mathrm{w}_t|\mathrm{y}_{1:t})$ using finite number of parameters. Then, we may recursively update those finite number of parameters. Indeed, as we will show next, $p(\mathrm{w}_t|\mathrm{y}_{1:t})$ has a normal distribution, that is

$$p(\mathrm{w}_t|\mathrm{y}_{1:t}) = N(\mathrm{w}_t|\mathrm{f}_t,\ F_t)$$

where $\mathrm{f}_t := \mathrm{E}(\mathrm{W}_t|\mathrm{Y}_{1:t} = \mathrm{y}_{1:t})$ and $F_t := \mathrm{Coy}(\mathrm{W}_t,\ \mathrm{W}_t|\mathrm{Y}_{1:t} = \mathrm{y}_{1:t})$

To simplify (LDS is time-homogenous and have zero bias):

$$\begin{cases} \mathrm{W}_t = A\mathrm{W}_{t-1} + \eta_t^w; & \eta_t^w|\mathrm{W}_{t-1} \sim N(0,\ \Sigma^w)\ , \\ \mathrm{Y}_t = B\mathrm{W}_t + \eta_t^y; & \eta_t^y|\mathrm{W}_t \sim N(0,\ \Sigma^y)\ , \end{cases}$$

• Since $(\mathrm{w}_t,\ \mathrm{y}_{1:t})$ is jointly normal $p(\mathrm{w}_t,\ \mathrm{y}_t|\mathrm{y}_{1:(t-1)})$ is also normal by property (4) of the normal density. Therefore, we have

$$p(\mathrm{w}_t,\ \mathrm{y}_t|\mathrm{y}_{1:(t-1)}) = N(\begin{pmatrix} \mathrm{w}_t \\ \mathrm{y}_t \end{pmatrix} \begin{pmatrix} \mu_\mathrm{w} \\ \mu_\mathrm{y} \end{pmatrix},\ \begin{pmatrix} \sum_\mathrm{ww} & \sum_\mathrm{wy} \\ \sum_\mathrm{wy}^\mathrm{T} & \sum_\mathrm{yy} \end{pmatrix})$$

We can find: $\mu_\mathrm{w} := \mathrm{E}(\mathrm{W}_t|\mathrm{Y}_{1:(t-1)} = \mathrm{y}_{1:(t-1)}) = \mathrm{E}(A\mathrm{W}_{t-1} + \eta_t^w|\mathrm{Y}_{1:(t-1)} = \mathrm{y}_{1:(t-1)}) = A\mathrm{f}_{t-1}$,

Property 4) ——————————————————————-

Note, we have referenced property 4, as stated below: $p(\mathrm{h}_t|\mathrm{v}_{1:t}) \sim N(\mathrm{h}_t|\mathrm{f}_t,\ F_t)$ where

$$\mathrm{f}_t = \mu_\mathrm{h} + \Sigma_\mathrm{hv}\Sigma_\mathrm{vv}^{-1}(\mathrm{v}_t - \mu_\mathrm{v}) = A\mathrm{f}_{t-1} + P_{t-1}B^\mathrm{T}(BP_{t-1}B^\mathrm{T})^{-1}(\mathrm{v}_t - B\ A\ \mathrm{f}_{t-1})\ ,$$

and

$$F_t = \Sigma_\mathrm{hh} - \Sigma_\mathrm{hv}\Sigma_\mathrm{vv}^{-1}\Sigma_\mathrm{hv}^\mathrm{T} = [I - K_{t-1}B]P_{t-1}.$$

Here, we have defined

$$P_{t-1} := AF_{t-1}A^\mathrm{T} + \Sigma^h,$$

and

$$K_{t-1}\ := P_{t-1}B^\mathrm{T}(BP_{t-1}B^\mathrm{T})^{-1}$$