SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Homework #8

Course: *Machine Learning (CS405)* – Professor: *Qi Hao*
Due date: *January 1th, 2020*

Homework Submission Instructions. Please write up your responses to the following problems clearly and concisely. We require you to write up your responses with A4 paper. You are allowed and encouraged to work together. You may discuss the homework to understand the problem and reach a solution in groups. However, *each student must write down the solution independently*. You must understand the solution well enough in order to reconstruct it by yourself. (This is for your own benefit: you have to take the exams alone.)

- **Written Homeworks.** All calculation problems **MUST** be written on single-sided A4 paper. You should bring and hand in it before class on the day of the deadline. Submitting the scan or photo version on Sakai will **NOT** be accepted.
- **Coding Homeworks.** All coding assignments will be done in Jupyter Notebooks. We will provide a `.ipynb` template for each assignment. Your final submission will be a `.ipynb` file with your answers and explanations (you should know how to write in Markdown or LaTeX). Make sure that all packages you need are imported at the beginning of the program, and your `.ipynb` file should **work step-by-step without any error**.

Question 1

Verify the M-step equations (13.18) and (13.19) for the initial state probabilities and transition probability parameters of the hidden Markov model by maximization of the expected complete-data log likelihood function (13.17), using appropriate Lagrange multipliers to enforce the summation constraints on the components of π and $\mathbf{A}$.

Answer. Consider first the maximization with respects to the components π_k of π. To do this we must take account of the summation constraint

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

We therefore first omit terms from $Q(\theta, \theta_{old})$ which are independent of π, and then add a Lagrange multiplier term to enforce the constraint, giving the following function to be maximized

$$\widetilde{Q} = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right).$$

Setting the derivative with respect to π_k equal to zero we obtain

$$0 = \gamma(z_{1k})\frac{1}{\pi_k} + \lambda. \qquad (1)$$

We now multiply through by π_k and then sum over k and make use of the summation constraint to give

$$\lambda = -\sum_{k=1}^{K}\gamma(z_{1k}).$$

Substituting back into (1) and solving for λ we obtain (13.18).
For the maximization with respect to $\mathbf{A}$ we follow the same steps and first omit terms from $Q(\theta, \theta_{old})$ which are independent of $\mathbf{A}$, and then add appropriate Lagrange multiplier terms to enforce the summation constraints. In this case there are K constraints to be satisfied since we must have

$$\sum_{k=1}^{K} A_{jk} = 1$$

for $j = 1, ..., K$. We introduce K Lagrange multipliers λ_j for $j = 1, ..., K$, and maximize the following function

$$\widehat{Q} = \sum_{n=2}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K}\xi(z_{n-1,j}, z_{nk})\ln A_{jk} + \sum_{j=1}^{K}\lambda_j(\sum_{k=1}^{K}A_{jk} - 1).$$

Setting the derivative of $\widehat{Q}$ with respect to A_{jk} to zero we obtain

$$0 = \sum_{n=2}^{N}\xi(z_{n-1,j}, z_{nk})\frac{1}{A_{jk}} + \lambda_j. \qquad (2)$$

Again we multiply through by A_{jk} and then sum over k and make use of the summation constraint to give

$$\lambda_j = -\sum_{n=2}^{N}\sum_{k=1}^{K}\xi(z_{n-1,j}, z_{nk}).$$

Substituting for λ_j in (2) and solving for A_{jk} we obtain (13.19).

Question 2

Suppose we wish to train a hidden Markov model by maximum likelihood using data that comprises R independent sequences of observations, which we denote by $\mathbf{X}^{(r)}$ where $r = 1, ..., R$. Show that in the E step of the EM algorithm, we simply evaluate posterior probabilities for the latent variables by running the α and β recursions independently for each of the sequences. Also show that in the M step, the initial probability and transition probability parameters are re-estimated using modified forms of (13.18) and (13.19) given by

$$\pi_k = \frac{\sum_{r=1}^{R} \gamma(z_{1k}^{(r)})}{\sum_{r=1}^{R} \sum_{j=1}^{K} \gamma(z_{1j}^{(r)})}$$

$$A_{jk} = \frac{\sum_{r=1}^{R} \sum_{n=2}^{N} \xi(z_{n-1,j}^{(r)}, z_{n,k}^{(r)})}{\sum_{r=1}^{R} \sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}^{(r)}, z_{n,l}^{(r)})}$$

where, for notational convenience, we have assumed that the sequences are of the same length (the generalization to sequences of different lengths is straightforward). Similarly, show that the M-step equation for re-estimation of the means of Gaussian emission models is given by

$$\mu_k = \frac{\sum_{r=1}^{R} \sum_{n=1}^{N} \gamma(z_{nk}^{(r)}) \mathbf{x}_n^{(r)}}{\sum_{r=1}^{R} \sum_{n=1}^{N} \gamma(z_{nk}^{(r)})}$$

Note that the M-step equations for other emission model parameters and distributions take an analogous form.

Answer. First of all, note that for every observed variable there is a corresponding latent variable, and so for every sequence $\mathbf{X}^{(r)}$ of observed variables there is a corresponding sequence $\mathbf{Z}^{(r)}$ of latent variables. The sequences are assumed to be independent given the model parameters, and so the joint distribution of all latent and observed variables will be given by

$$p(\mathbf{X}, \mathbf{Z}|\theta) = \prod_{r=1}^{R} p(\mathbf{X}^{(r)}, \mathbf{Z}^{(r)}|\theta)$$

where $\mathbf{X}$ denotes $\{\mathbf{X}^{(r)}\}$ and $\mathbf{Z}$ denotes $\{\mathbf{Z}^{(r)}\}$. Using the sum and product rules of probability we then see that posterior distribution for the latent sequences then factorizes with respect to those sequences, so that

$$
\begin{aligned}
p(\mathbf{Z}|\mathbf{X},\theta) &= \frac{p(\mathbf{X},\mathbf{Z}|\theta)}{\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\theta)} \\
&= \frac{\prod_{r=1}^{R} p(\mathbf{X}^{(r)},\mathbf{Z}^{(r)}|\theta)}{\sum_{\mathbf{Z}^{(1)}}\cdots\sum_{\mathbf{Z}^{(R)}}\prod_{r=1}^{R} p(\mathbf{X}^{(r)},\mathbf{Z}^{(r)}|\theta)} \\
&= \prod_{r=1}^{R}\{\frac{p(\mathbf{X}^{(r)},\mathbf{Z}^{(r)}|\theta)}{\sum_{\mathbf{Z}^{(r)}} p(\mathbf{X}^{(r)},\mathbf{Z}^{(r)}|\theta)}\} \\
&= \prod_{r=1}^{R} p(\mathbf{Z}^{(r)},\mathbf{X}^{(r)}|\theta)
\end{aligned}
$$

Thus the evaluation of the posterior distribution of the latent variables, corresponding to the E-step of the EM algorithm, can be done independently for each of the sequences (using the standard alpha-beta recursions).

Now consider the M-step. We use the posterior distribution computed in the E-step using θ_{old} to evaluate the expectation of the complete-data log likelihood. From our expression for the joint distribution we see that this is given by

$$
\begin{aligned}
Q(\theta,\theta_{old}) &= \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X},\mathbf{Z}|\theta)] \\
&= \mathbb{E}_{\mathbf{Z}}[\sum_{r=1}^{R}\ln p(\mathbf{X}^{(r)},\mathbf{Z}^{(r)}|\theta)] \\
&= \sum_{r=1}^{R} p(\mathbf{Z}^{(r)},\mathbf{X}^{(r)}|\theta_{old})\ln p(\mathbf{X},\mathbf{Z}|\theta) \\
&= \sum_{r=1}^{R}\sum_{k=1}^{K}\gamma(z_{1k}^{(r)})\ln\pi_k + \sum_{r=1}^{R}\sum_{n=2}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K}\xi(z_{n-1,j}^{(r)},z_{nk}^{(r)})\ln A_{jk} \\
&\quad + \sum_{r=1}^{R}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk}^{(r)})\ln p(\mathbf{x}_n^{(r)}|\phi_k).
\end{aligned}
$$

We now maximize this quantity with respect to π and $\mathbf{A}$ in the usual way, with Lagrange multipliers to take account of the summation constraints (see Answer 1), yielding (13.124) and (13.125). The M-step results for the mean of the Gaussian follow in the usual way also.

Question 3

In this exercise, we show that when the Kalman filter equations are applied to independent observations, they reduce to the results given in Section 2.3 for the maximum likelihood solution for a single Gaussian distribution. Consider the problem of finding the mean μ of a single Gaussian random variable x, in which we are given a set of independent observations $\{x_1,...,x_N\}$. To model this we can use a linear dynamical system governed by (13.75) and (13.76), with latent variables $\{z_1,...,z_N\}$ in which $\mathbf{C}=1, \mathbf{A}=1$ and $\Gamma=0$. Let the parameters μ_0 and $\mathbf{P}_0$ of the initial state be denoted by μ_0 and σ_0^2, respectively, and suppose that $\mathbf{\Sigma}$ becomes σ^2. Write down the corresponding Kalman filter equations starting from the general results (13.89) and (13.90), together with (13.94) and (13.95). Show that these are equivalent to the results (2.141) and (2.142) obtained directly by considering independent data.

Answer. Since $\mathbf{C} = 1$, $\mathbf{P}_0 = \sigma_0^2$ and $\Sigma = \sigma^2$, (13.97) gives

$$\mathbf{K}_1 = \frac{\sigma_0^2}{\sigma_0^2 + \sigma}.$$

Substituting this into (13.94) and (13.95), we get

$$\begin{aligned}
\mu_1 &= \mu_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma}(x_1 - \mu_0) \\
&= \frac{1}{\sigma_0^2 + \sigma}(\sigma_0^2 x_1 - \sigma\mu_0) \\
\sigma_1^2 &= (1 - \frac{\sigma_0^2}{\sigma_0^2 + \sigma})\sigma_0^2 \\
&= \frac{\sigma_0^2\sigma^2}{\sigma_0^2 + \sigma}
\end{aligned}$$

where σ_1^2 replaces $\mathbf{V}_1$. We note that these agree with (2.141) and (2.142), respectively. We now assume that (2.141) and (2.142) hold for N, and we rewrite them as

$$\mu_N = \sigma_N^2(\frac{1}{\sigma_0^2}\mu_0 + \frac{N}{\sigma^2}\mu_{ML}^{(N)}) \qquad (3)$$

$$\sigma_N^2 = \frac{\sigma_0^2\sigma^2}{N\sigma_0^2 + \sigma^2} \qquad (4)$$

where, analogous to (2.143),

$$\mu_{ML}^{(N)} = \frac{1}{N}\sum_{n=1}^{N} x_n. \qquad (5)$$

Since $\mathbf{A} = 1$ and $\Gamma = 0$, (13.88) gives

$$\mathbf{P}_N = \sigma_N^2 \qquad (6)$$

substituting this into (13.92), we get

$$\mathbf{K}_{K+1} = \frac{\sigma_N^2}{N\sigma_N^2 + \sigma}. \qquad (7)$$

Using (4), (6), (7) and (13.90), we get

$$\begin{aligned}
\sigma_{N+1}^2 &= (1 - \frac{\sigma_N^2}{\sigma_N^2 + \sigma})\sigma_N^2 \\
&= \frac{\sigma_N^2\sigma^2}{\sigma_N^2 + \sigma} \qquad (8) \\
&= \frac{\sigma_0^2\sigma^4/(\sigma_0^2 + \sigma)}{(\sigma^2\sigma_0^2\sigma^4 + \sigma^2 N\sigma_0^2)/(\sigma_0^2 + \sigma)} \\
&= \frac{\sigma_0^2\sigma^2}{(N+1)\sigma_0^2 + \sigma}.
\end{aligned}$$

Using (3), (5), (7), (8) and (13.89), we get

$$
\begin{aligned}
\mu_{N+1} &= \mu_N + \frac{\sigma_N^2}{\sigma_N^2 + \sigma}(x_{N+1} - \mu_N) \\
&= \frac{1}{\sigma_N^2 + \sigma}(\sigma_N^2 x_{N+1} + \sigma \mu_N) \\
&= \frac{\sigma_N^2}{\sigma_0^2 + \sigma} x_{N+1} + \frac{\sigma_N^2 \sigma^2}{\sigma_N^2 + \sigma}\left(\frac{1}{\sigma_0^2}\mu_0 + \frac{1}{\sigma^2}\sum_{n=1}^{N} x_n\right) \\
&= \sigma_{N+1}^2\left(\frac{1}{\sigma_0^2}\mu_0 + \frac{N+1}{\sigma^2}\mu_{ML}^{(N+1)}\right).
\end{aligned}
$$

Thus (3) and (4) must hold for all $N \geq 1$.

Question 4

The Kalman filter and smoother equations allow the posterior distributions over individual latent variables, conditioned on all of the observed variables, to be found efficiently for linear dynamical systems. Show that the sequence of latent variable values obtained by maximizing each of these posterior distributions individually is the same as the most probable sequence of latent values. To do this, simply note that the joint distribution of all latent and observed variables in a linear dynamical system is Gaussian, and hence all conditionals and marginals will also be Gaussian, and then make use of the result (2.98) in textbook.

The result (2.98) in textbook:
Given a joint Gaussian distribution $\mathcal{N}(\mathrm{x}|\mu, \Sigma)$ with $\Lambda \equiv \Sigma^{-1}$ and

$$
\mathrm{x} = \begin{pmatrix} \mathrm{x}_a \\ \mathrm{x}_b \end{pmatrix} \quad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}
$$

$$
\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}
$$

Marginal distribution:

$$
p(\mathrm{x}_a) = \mathcal{N}(\mathrm{x}_a|\mu_a, \Sigma_{aa}) \tag{2.98}
$$

Answer. Since the joint distribution over all variables, latent and observed, is Gaussian, we can maximize w.r.t. any chosen set of variables. In particular, we can maximize w.r.t. all the latent variables jointly or maximize each of the marginal distributions separately. However, from (2.98), we see that the resulting means will be the same in both cases and since the mean and the mode coincide for the Gaussian, maximizing w.r.t. to latent variables jointly and individually will yield the same result.

Question 5

Andrew lives a simple life. Some days he is Angry and some days he is Happy. But he hides his emotional state, and so all you can observe is whether he smiles, frowns, laughs, or yells. We start on day 1 in the Happy state and there is one transition per day.

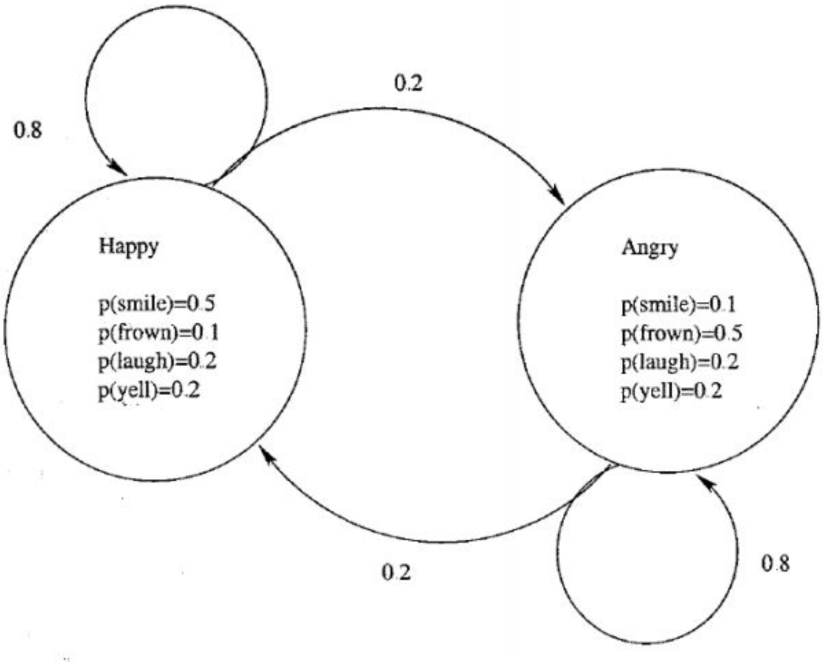


Figure 1: Transition Model for Andrew

We define:
q_t: state on day t
O_t: observation on day t
(a) What is $P(q_2 = Happy)$?
(b) What is $P(O_2 = frown)$?
(c) What is $P(q_2 = Happy|O_2 = frown)$?
(d) What is $P(O_{100} = yell)$?
(e) Assume that $O_1 = O_2 = O_3 = O_4 = O_5 = frown$.What is the most likely sequence of the states?

Answer. (a)

$$P(q_2 = Happy) = 0.8$$

(b)

$$P(O_2 = frown) = 0.8 * 0.1 + 0.2 * 0.5 = 0.18$$

(c)

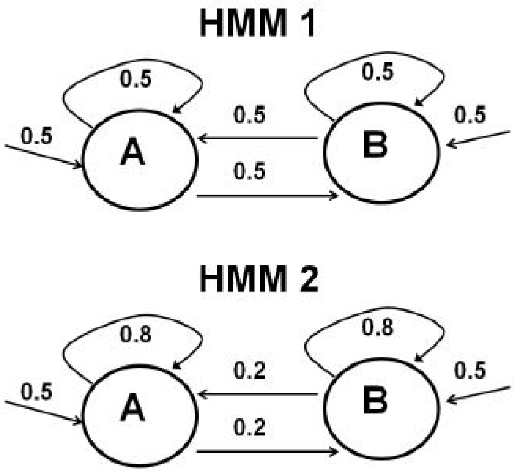
$$P(q_2 = Happy|O_2 = frown) = \frac{P(O_2 = frown|q_2 = Happy)P(O_2 = Happy)}{P(q_2 = frown)} = \frac{0.1 * 0.8}{0.18} = \frac{4}{9}$$

(d)

$$\begin{aligned} & P(O_{100} = yell) \\ = & P(O_{100} = yell|q_{100} = Happy)P(q_{100} = Happy) + P(O_{100} = yell|q_{100} = Angry)P(q_{100} = Angry) \\ = & 0.2 \end{aligned}$$

(e) Happy, Angry, Angry, Angry, Angry

Question 6



The figure above presents two HMMs. States are represented by circles and transitions by edges. In both, emissions are deterministic and listed inside the states. Transition probabilities and starting probabilities are listed next to the relevant edges. For example, in HMM 1 we have a probability of 0.5 to start with the state that emits A and a probability of 0.5 to transition to the state that emits B if we are now in the state that emits A.
In the question below, $O_{100} = \text{A}$ means that the 100th symbol emitted by the HMM is A.

(a)What is $P(O_{100} = \text{A}, O_{101} = \text{A}, O_{102} = \text{A})$ for HMM1?

(b)What is $P(O_{100} = \text{A}, O_{101} = \text{A}, O_{102} = \text{A})$ for HMM2?

(c)Let P_1 be: $P_1 = P(O_{100} = \text{A}, O_{101} = \text{B}, O_{102} = \text{A}, O_{103} = \text{B})$ for HMM1 and let $P_2 = P(O_{100} = \text{A}, O_{101} = \text{B}, O_{102} = \text{A}, O_{103} = \text{B})$ for HMM2.Choose the correct answer from the choices below and briefly explain.
1.$P_1 > P_2$
2.$P_1 = P_2$
3.$P_1 < P_2$
4.Impossible to tell the relationship between the two probabilities

Answer. (a)Note that

$$P(O_{100} = \text{A}, O_{101} = \text{A}, O_{102} = \text{A}) = P(O_{100} = \text{A}, O_{101} = \text{A}, O_{102} = \text{A}, S_{100} = \text{A}, S_{101} = \text{A}, S_{102} = \text{A})$$

since if we are not always in state A we will not be able to emit A. Given the Markov property this can be written as:

$$\begin{aligned} &P(O_{100} = \mathrm{A}, O_{101} = \mathrm{A}, O_{102} = \mathrm{A}, S_{100} = \mathrm{A}, S_{101} = \mathrm{A}, S_{102} = \mathrm{A}) \\ = &P(O_{100} = \mathrm{A}|S_{100} = A)P(S_{100} = A)P(O_{101} = \mathrm{A}|S_{101} = \mathrm{A})P(S_{101} = \mathrm{A}|S_{100} = \mathrm{A}) \\ &\cdot P(O_{102} = \mathrm{A}|S_{102} = \mathrm{A})P(S_{102} = \mathrm{A}|S_{101} = \mathrm{A}) \end{aligned}$$

The emission probabilities in the above equation are all 1. The transitions are all 0.5. So the only question is: What is $P(S_{100} = \mathrm{A})$? Since the model is fully symmetric, the answer to this is 0.5 and so the total equation evaluates to: 0.5^3

(b) For HMM2, $P(O_{100} = \mathrm{A}, O_{101} = \mathrm{A}, O_{102} = \mathrm{A}) = 0.5 * 0.8^2$

(c) P_1 evaluates to 0.5^4 while P_2 is $0.5 * 0.2^4$ so clearly $P_1 > P_2$