CS 7641 CSE/ISYE 6740 Homework 4

Deadline: 11/28 Monday, 11:55 pm

- Submit your answers as an electronic copy on T-square.
- No unapproved extension of deadline is allowed. Late submission will lead to 0 credit.
- Typing with Latex is highly recommended. Typing with MS Word is also okay. If you handwrite, try to be clear as much as possible. No credit may be given to unreadable handwriting.
- Explicitly mention your collaborators if any. For the programming problem, it is absolutely not allowed to share your source code with anyone in the class as well as to use code from the Internet without reference.
- Recommended reading: PRML Section 13.2

1 Kernels [20 points]

(a) Identify which of the followings is a valid kernel. If it is a kernel, please write your answer explicitly as 'True' and give mathematical proofs. If it is not a kernel, please write your answer explicitly as 'False' and give explanations. [8 pts]

Suppose K_1 and K_2 are valid kernels (symmetric and positive definite) defined on $R^m \times R^m$.

1. $K(u, v) = \alpha K_1(u, v) + \beta K_2(u, v), \alpha, \beta \in R$.
2. $K(u, v) = K_1(f(u), f(v))$ where $f : R^m \to R^m$. coefficients.
3. $$K(u, v) = \begin{cases} 1 & \text{if } \|u - v\|_2 \leqslant 1 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$
4. Suppose K' is a valid kernel.
$$K(u, v) = \frac{K'(u, v)}{\sqrt{K'(u, u)K'(v, v)}}. \tag{2}$$

(1)Answer: False. It is not a kernel.
$K(u, v)$ is a valid kernel, if and only if it is positive semidefinite, in other words:

$$\forall x \in R^m, \quad x^T K x \geq 0.$$

Since $K(u,v) = \alpha K_1(u,v) + \beta K_2(u,v)$, then we can obtain:

$$x^T(\alpha K_1 + \beta K_2)x = \alpha x^T K_1 x + \beta x^T K_2 x$$

Also K_1 and K_2 are valid kernel, then we must have $x^T K_1 x \geq 0, x^T K_2 x \geq 0$.

For instance, if $\alpha = \beta = -1$, then $x^T(\alpha K_1 + \beta K_2)x < 0$.

(2)Answer: True. It is a kernel.
Assume:

$$\alpha = f(u), \beta = f(v),$$

where $f : R^m \to R^m$.

Since K_1 is a valid kernel, then the Gram matrix is positive semidefinite, in other words:

$$\forall x \in R^m, \quad x^T K_1(\alpha, \beta)x = x^T K_1(f(u), f(v))x \geq 0.$$

Besides, we have:

$$K(u,v) = K_1(f(u), f(v)) = K_1(\alpha, \beta)$$

Thus:

$$\forall x \in R^m, \quad x^T K(u,v)x = x^T K_1(f(u), f(v))x = x^T K_1(\alpha, \beta)x \geq 0$$

Therefore, this kernel is a valid kernel.(Function f only transform m-dimension vectors u, v into another m-dimension vectors α, β, which won't affect validity of the kernel.)

(3)Answer: False. It is not a kernel.
According to the definition of K, we can notice that any element of $k_{ij} \in \{0, 1\}$, besides, since $||u - v||_2 = ||v - u||_2$, then $K(u,v)$ must be a symmetric matrix. Thus, for example:

Suppose $x = [-2, 1, 2]^T$ and K is:

$$K = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Then we can obtain:

$$x^T K x = -3 < 0$$

Thus we cannot guarantee that $x^T K x$ is greater or equal to 0.

Therefore, this kernel is not a valid kernel.

(4)Answer: True. It is a kernel.
Assume entry of matrix K' is K'_{ij}, so construct a diagonal matrix J whose entries are

$$J_{ii} = \frac{1}{\sqrt{K'_{ii}}}$$

Since K' is a valid kernel, then it is a positive semi-definite Gram matrix. M1atrix J's entries are all positive. Matrix K can be expressed as:

$$K = J^T K' J$$

According to the definition of positive semi-definite matrix, determinant of K is

$$|K| = |J^T||K'||J| > 0$$

Then $K(u, v)$ is a valid kernel.

(b) Write down kernelized version of Fisher's Linear Discriminant Analysis using kernel trick. Please provide full steps and all details of the method. [*Hint: Use kernel to replace inner products.*] [12 pts]

(b)Solution:
To construct a kenelized version of Fisher's Linear Discriminant Analysis, we are going to use the kernel trick. Given l data points x_i and map to a new feature space, F by function $\phi(x_i)$. l_i is the number of data points that belong to class C_i. c is the number of classes.

Consider Fisher discriminant to $c > 2$ classes(multi-classes) and also dimensionality D is greater than the number of classes c. And $D' > 1$ linear 'feature' $y_k = w_k^T x$, $k = 1, \cdots, D'$. Therefore, LDA can group the input into c classes by reducing the feature from vector x to vector y. Therefore, in new feature space ,we can have:

$$J(w) = \frac{w^T S'_B w}{w^T S'_W w}$$

where

$$S'_B = \sum_{i=1}^{c} l_i (m'_i - m')(m'_i - m')^T$$
$$S'_W = \sum_{i=1}^{c} \sum_{n=1}^{l_i} (\phi(x_n^i) - m'_i)(\phi(x_n^i) - m'_i)^T$$
$$m'_i = \frac{1}{l_i} \sum_{j=1}^{l_i} \phi(x_j^i)$$

m' is the mean of all the data in the new feature space.

$$m' = \frac{1}{l} \sum_{i=1}^{l} \phi(x_i) = \frac{1}{l} \sum_{i=1}^{c} l_i m'_i$$

w will have an expansion of the form:

$$w = \sum_{j=1}^{N} \alpha_j \phi(x_j)$$

Then note that

$$w^T m_i' = \frac{1}{l_i} \sum_{j=1}^{l} \sum_{k=1}^{l_i} \alpha_j k(x_j, x_k^i) = \alpha^T M_i$$

where

$$(M_i)_j = \frac{1}{l_i} \sum_{k=1}^{l_i} k(x_j, x_k^i)$$

Then the numerator of $J(w)$ is given by:

$$w^T S_B' w = w^T \sum_{i=1}^{c} l_i (m_i' - m')(m_i' - m')^T w = A^T M A$$

Similarly, the denominator can be written as

$$w^T S_W' w = w^T \sum_{i=1}^{c} \sum_{n=1}^{l_i} (\phi(x_n^i) - m_i')(\phi(x_n^i) - m_i')^T w = A^T N A$$

In order to maximize the $J(w)$

$$J(w) = \frac{|w^T S_B' w|}{|w^T S_W' w|}$$

The kernel trick can be used and thte multi-class KFD becomes:

$$A^* = argmax_A = \frac{A^T M A}{A^T N A}$$

where $A = [\alpha_1, \cdots \alpha_{c-1}]$ and

$$M = \sum_{j=1}^{c} l_j (M_j - M_*)(M_j - M_*)^T$$

$$N = \sum_{j=1}^{c} K_j (I - 1_{l_j}) K_j^T$$

$$(M_*)_j = \frac{1}{l} \sum_{k=1}^{l} k(x_j, x_k)$$

The projection of a new input, x_t, is given by:

$$y_t = (A^*)^T K_t$$

where the ith component of K_t is given by $k(x_i, x_t)$.

2 Markov Random Field, Conditional Random Field [20 pts]

[a-b] A probability distribution on 3 discrete variables a,b,c is defined by $P(a, b, c) = \frac{1}{Z}\psi(a, b, c) = \frac{1}{Z}\phi_1(a, b)\phi_2(b, c)$, where the table for the two factors are given below.

a	b	$\phi_1(a,b)$
0	0	4
0	1	3
1	0	3
1	1	1

b	c	$\phi_2(b,c)$
0	0	3
0	1	2
0	2	1
1	0	4
1	1	1
1	2	3

(a) Compute the slice of the joint factor $\psi(a,b,c)$ corresponding to $b=1$. This is the table $\psi(a,b=1,c)$. [5 pts]

(a)Solution:
For $\psi(a,b=1,c)$, here are the different combinations:

$$\begin{aligned}
\psi(a=0,b=1,c=0) &= \phi_1(a=0,b=1)\times\phi_2(b=1,c=0) = 3\times 4 = 12\\
\psi(a=0,b=1,c=1) &= \phi_1(a=0,b=1)\times\phi_2(b=1,c=1) = 3\times 1 = 3\\
\psi(a=1,b=1,c=0) &= \phi_1(a=1,b=1)\times\phi_2(b=1,c=0) = 1\times 4 = 4\\
\psi(a=1,b=1,c=1) &= \phi_1(a=1,b=1)\times\phi_2(b=1,c=1) = 1\times 1 = 1\\
\psi(a=0,b=1,c=2) &= \phi_1(a=0,b=1)\times\phi_2(b=1,c=2) = 3\times 3 = 9\\
\psi(a=1,b=1,c=2) &= \phi_1(a=1,b=1)\times\phi_2(b=1,c=2) = 1\times 3 = 3
\end{aligned}$$

Therefore the table for $\psi(a,b=1,c)$ is:

a	b	c	$\psi(a,b=1,c)$
0	1	0	12
0	1	1	3
1	1	0	4
1	1	1	1
0	1	2	9
1	1	2	3

(b) Compute $P(a=1,b=1)$. [5 pts]

(b)Solution:
Since all the combinations of a,b and c are the Z to normalize so that the total probability can be 1. Thus:

$$\begin{aligned}
\sum_{a,b,c}\psi(a,b,c) &= Z\\
&= 4\times(3+2+1)+3\times(4+1+3)+3\times(3+2+1)+1\times(4+1+3)\\
&= 74
\end{aligned}$$

$$
\begin{aligned}
P(a=1,b=1) &= \sum_{c\in\{0,1,2\}} P(a=1,b=1,c) \\
&= P(a=1,b=1,c=0) + P(a=1,b=1,c=1) + P(a=1,b=1,c=2) \\
&= \frac{1}{Z}\psi(1,1,0) + \frac{1}{Z}\psi(1,1,1) + \frac{1}{Z}\psi(1,1,2) \\
&= \frac{4}{74} + \frac{1}{74} + \frac{3}{74} = \frac{4}{37}
\end{aligned}
$$

(c) Explain the difference between Conditional Random Fields and Hidden Markov Models with respect to the following factors. Please give only a one-line explanation. [10 pts]

- Type of model - generative/discriminative
- Objective function optimized
- Require a normalization constant

(c)Answer:

- Type of model - generative/discriminative

  (1) Conditional Random Fields is discriminative model. It learns the conditional probability distribution $p(y|x)$ and it can produce samples, it can use to classify.

  (2) Hidden Markov Models is generative model. It learns the joint probability distribution $p(x, y)$ and it can produce samples according to joint probability distribution.

- Objective function optimized

  (1) The objective function of Conditional Random Fields is to maximize the probability of hidden state given observed sequence, $p(y|x)$.

  (2) The objective function of Hidden Markov Models is to maximize the joint probability distribution of hidden state and observed sequence, $p(x, y)$.

- Require a normalization constant

  (1) The Conditional Random Fields require a normalization constant. It need a normalization factor, constant Z, to ensures the distribution sums to 1.

  (2) The Hidden Markov Models doesn't require a normalization constant. Because the sum over all possible hypotheses is guaranteed to add up to 1.

3 Hidden Markov Model [50 pts]

This problem will let you get familiar with HMM algorithms by doing the calculations by hand. **[a-c]** There are three coins $(1, 2, 3)$, to throw them randomly, and record the result. $S = 1, 2, 3$; $V = H, T$ (Head or Tail); A, B, π is given as

A:

	1	2	3
1	0.9	0.05	0.05
2	0.45	0.1	0.45
3	0.45	0.45	0.1

π :

π	1/3	1/3	1/3

B:

	1	2	3
H	0.5	0.75	0.25
T	0.5	0.25	0.75

(a) Given the model above, what's the probability of observation $O = H, T, H$. [10 pts]

(a)Solution:
For the first throw, we can obtain:

$$P(H,1) = \frac{1}{3} \times 0.5 = 0.17, \quad P(H,2) = \frac{1}{3} \times 0.75 = 0.25, \quad P(H,3) = \frac{1}{3} \times 0.25 = 0.083$$

For the second throw, we can obtain:

$$\begin{aligned} P(HT,1) &= (0.17 \times 0.9 + 0.25 \times 0.45 + 0.083 \times 0.45) \times 0.50 = 0.1514 \\ P(HT,2) &= (0.17 \times 0.05 + 0.25 \times 0.1 + 0.083 \times 0.45) \times 0.25 = 0.0177 \\ P(HT,3) &= (0.17 \times 0.05 + 0.25 \times 0.45 + 0.083 \times 0.1) \times 0.75 = 0.0970 \end{aligned}$$

For the third throw, we can obtain:

$$\begin{aligned} P(HTH,1) &= (0.1514 \times 0.9 + 0.0177 \times 0.45 + 0.097 \times 0.45) \times 0.50 = 0.0939 \\ P(HTH,2) &= (0.1514 \times 0.05 + 0.0177 \times 0.1 + 0.097 \times 0.45) \times 0.75 = 0.0397 \\ P(HTH,3) &= (0.1514 \times 0.05 + 0.0177 \times 0.45 + 0.097 \times 0.1) \times 0.25 = 0.0063 \end{aligned}$$

Therefore, the probability of observation $O = H, T, H$ is

$$P(HTH) = \sum_{i=1}^{3} P(HTH, i) = 0.0939 + 0.0397 + 0.0063 = 0.1399$$

(b) Describe how to get the A, B, and π, when they are unknown. [10 pts]

(b)Solution:
To get the unknown parameters, A, B, and π, the problem becomes to an unsupervised learning problem. To solve it, we need to implement Baum-Welch (EM) algorithm.

EXPECTATION MAXIMIZATION
Given the result of coins, $V = v_1, v_2, \cdots v_T$, find parameters $\theta = (\pi, A, B)$ that maximize $p(V|\theta)$. Our state of knowledge of the values of the latent variables in S is given only by the posterior distribution $p(S|V, \theta)$. The current parameter values is denoted as θ^{old} to find the posterior distribution $p(S|V, \theta^{old})$.

- Starting with our best guess of the model, parameters $\theta^{old} = (\pi, A, B)$
- The joint probability over both latent and observed variables is then given by:

$$p(V, S|\theta) = p(s_1|\pi) \prod_{t=1}^{T} p(v_t|s_t, B) \prod_{t=2}^{T} p(s_t|s_{t-1}, A)$$

The complete log likelihood is

$$Q(\theta, \theta^{old}) = \sum_{S} p(S|V, \theta^{old}) \ln p(V, S|\theta)$$

For convenience, let me introduce some notation here.

– Marginal posterior distribution

$$\gamma(s_t) = p(s_t|V, \theta^{old})$$

– Joint posterior distribution

$$\zeta(s_{t-1}, s_t) = p(s_{t-1}, s_t|V, \theta^{old})$$

For each value of t, we can store $\gamma(s_t)$ using a set of K nonnegative numbers that sum to unity, and similarly to store $\zeta(s_{t-1}, s_t)$ using a $K \times K$ matrix. Since the expectation of a binary variable is probability that it takes the value 1, then we can obtain:

$$\gamma(s_{tk}) = \mathbb{E}[s_{tk}] = \sum_{s} \gamma(s) s_{tk}$$

$$\zeta(s_{t-1,j}, s_{t,k}) = \mathbb{E}[s_{t-1,j} s_{tk}] = \sum_{S} \gamma(s) s_{t-1,j} s_{tk}$$

Then the expected complete log likelihood is given by:

$$Q(\theta, \theta^{old}) = \sum_{k=1}^{K} \gamma(s_{1k}) \ln \pi_k + \sum_{t=2}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \zeta(s_{t-1,j} s_{tk}) \ln A_{jk} + \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma(s_{tk}) \ln B_{tk}$$

- **E step:** Implement efficient Forward and Backward algorithm to evaluate the quantity $\gamma(s_{tk})$ and $\zeta(s_{t-1,j} s_{tk})$.
  Using the conditional independence property and product rule of probability, we can obtain:

$$\gamma(s_{tk}) = \frac{p(v_1, \cdots, v_t, s_{tk}) p(v_t + 1, \cdots, v_T|s_{tk})}{p(V)}$$

Similarly,

$$\begin{aligned} \zeta(s_{t-1,j}, s_{nk}) &= p(s_{t-1,j}, s_t|V) \\ &= \frac{\alpha(s_{t-1,j}) p(v_t|s_{tk}) p(s_{tk}|s_{t-1,k}) \beta(s_{tk})}{p(V)} \end{aligned}$$

By using Forward-Backward algorithm, we can obtain:

$$\gamma(s_{tk}) = \frac{\alpha(s_{tk})\beta(s_{tk})}{\sum_{s_{tk}} \alpha(s_{tk})}$$
$$\zeta(s_{t-1,j}, s_{nk}) = \frac{\alpha(s_{t-1,j})p(v_t|s_{tk})p(s_{tk}|s_{t-1,k})\beta(s_{tk})}{\sum_{s_{tk}} \alpha(s_{tk})}$$

where we have defined:

$$\alpha(s_{tk}) = p(v_t|s_{tk}) \sum_{s_{t-1}} \alpha(s_{t-1,k})p(s_{tk}|s_{t-1,k})$$
$$\beta(s_{tk}) = \sum_{s_{t+1,k}} \beta(s_{t+1})p(v_{t+1}|s_{t+1,k})p(s_{t+1}|s_t)$$

- **M step:** To find the current parameters that maximize the log likelihood function, we should optimize analytically with Lagrange multipliers, because there are constraints on π, A and B.

$$L(\theta, \theta^{old}) = Q(\theta, \theta^{old}) + \lambda_\pi(\sum_{k=1}^{K} \pi_k - 1) + \sum_{j=1}^{K} \lambda_{A_j}(\sum_{k=1}^{K} A_{jk} - 1) + \sum_{k=1}^{K} \lambda_{B_k}(\sum_{t=1}^{T} B_{tk} - 1)$$

Take derivative with respect to π_k, A_{jk} and B_{tk}, let the equation equal to 0 and substitute the λ into the previous equations, then we can obtain:

$$\frac{\partial L}{\partial \pi_k} = \frac{\partial}{\partial \pi_k}(\sum_{k=1}^{K} \gamma(s_{1k}) \ln \pi_k) + \lambda_\pi = 0$$
$$\Rightarrow \lambda_\pi = -(\sum_{k=1}^{K} \gamma(s_{1k}))$$
$$\Rightarrow \pi_k = \frac{\gamma(s_{1k})}{\sum_{k=1}^{K} \gamma(s_{1k})}$$

$$\frac{\partial L}{\partial A_{jk}} = \frac{\partial}{\partial A_{jk}}(\sum_{t=2}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \zeta(s_{t-1,j}s_{tk}) \ln A_{jk}) + \lambda_{A_j}$$
$$= \frac{\sum_{t=2}^{T} \zeta(s_{t-1,j}s_{tk})}{A_{jk}} + \lambda_{A_j} = 0$$

sum over k to eliminate A_{jk}

$$\Rightarrow \sum_{t=2}^{T} \sum_{k=1}^{K} \zeta(s_{t-1,j}s_{tk}) + \lambda_{A_j} \sum_{k=1}^{K} A_{jk} = 0$$
$$\Rightarrow \lambda_{A_j} = -\sum_{t=2}^{T} \sum_{k=1}^{K} \zeta(s_{t-1,j}s_{tk})$$
$$\Rightarrow A_{jk} = \frac{\sum_{t=2}^{T} \zeta(s_{t-1,j}s_{tk})}{\sum_{t=2}^{T} \sum_{k=1}^{K} \zeta(s_{t-1,j}s_{tk})}$$

$$\begin{aligned}\frac{\partial L}{\partial B_{tk}} &= \frac{\partial}{\partial B_{tk}}(\sum_{t=1}^{T}\sum_{k=1}^{K}\gamma(s_{tk})\ln B_{tk}) + \lambda_{B_k} \\ &= \frac{\gamma(s_{tk})}{B_{tk}} + \lambda_{B_k} = 0 \\ &\Rightarrow \lambda_{B_k} = -\frac{\gamma(s_{tk})}{B_{tk}} \\ &\Rightarrow B_{tk} = \frac{\gamma(s_{tk})}{\sum_{k=1}^{K}\gamma(s_{tk})}\end{aligned}$$

(c) In class, we studied discrete HMMs with discrete hidden states and observations. The following problem considers a continuous density HMM, which has discrete hidden states but continuous observations. Let $S_t \in 1, 2, ..., n$ denote the hidden state of the HMM at time t, and let $X_t \in R$ denote the real-valued scalar observation of the HMM at time t. In a continuous density HMM, the emission probability must be parameterized since the random variable X_t is no longer discrete. It is defined as $P(X_t = x|S_t = i) = \mathcal{N}(\mu_i, \sigma_i^2)$. Given m sequences of observations (each of length T), derive the EM algorithm for HMM with Gaussian observation model. [14 pts]

(c)Solution:
Given the m sequences of observations $X = x_1, x_2, \cdots x_T$, $p(x_t|s_t) = \mathcal{N}(\mu_k, \sigma_k^2)$, find parameters $\theta = (\pi, A, B)$ that maximize $p(X|\theta)$. Since the derivative of this problem is similar to problem 3(b), thus I am going to utilize same notation and some formulas derived from 3(b).

- Starting with our best guess of the model, parameters $\theta^{old} = (\pi, A, B)$
- From problem 3(b) we can know the following still hold:
  The joint probability over both latent and observed variables is then given by:
  $$p(X, S|\theta) = p(s_1|\pi)\prod_{t=1}^{T}p(x_t|s_t, B)\prod_{t=2}^{T}p(s_t|s_{t-1}, A)$$
  The complete log likelihood is
  $$Q(\theta, \theta^{old}) = \sum_{S}p(S|X, \theta^{old})\ln p(X, S|\theta)$$
  The expected complete log likelihood is given by:
  $$Q(\theta, \theta^{old}) = \sum_{k=1}^{K}\gamma(s_{1k})\ln\pi_k + \sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K}\zeta(s_{t-1,j}s_{tk})\ln A_{jk} + \sum_{t=1}^{T}\sum_{k=1}^{K}\gamma(s_{tk})\ln B_{tk}$$
- **E step:** Implement efficient Forward and Backward algorithm to evaluate the quantity $\gamma(s_{tk})$ and $\zeta(s_{t-1,j}s_{tk})$.
  By using Forward-Backward algorithm, we can obtain:
  $$\gamma(s_{tk}) = \frac{\alpha(s_{tk})\beta(s_{tk})}{\sum_{s_{tk}}\alpha(s_{tk})}$$
  $$\zeta(s_{t-1,j}, s_{nk}) = \frac{\alpha(s_{t-1,j})p(v_t|s_{tk})p(s_{tk}|s_{t-1,k})\beta(s_{tk})}{\sum_{s_{tk}}\alpha(s_{tk})}$$

where we have defined:

$$\alpha(s_{tk}) = p(x_t|s_{tk}) \sum_{s_{t-1}} \alpha(s_{t-1,k}) p(s_{tk}|s_{t-1,k})$$
$$\beta(s_{tk}) = \sum_{s_{t+1,k}} \beta(s_{t+1}) p(x_{t+1}|s_{t+1,k}) p(s_{t+1}|s_t)$$

- **M step:** To find the current parameters that maximize the log likelihood function, we should optimize analytically with Lagrange multipliers, because there are constraints on π, A and B. Similarly, we can use the same derivative from problem 3(b):

$$L(\theta, \theta^{old}) = Q(\theta, \theta^{old}) + \lambda_\pi (\sum_{k=1}^{K} \pi_k - 1) + \sum_{j=1}^{K} \lambda_{A_j} (\sum_{k=1}^{K} A_{jk} - 1) + \sum_{k=1}^{K} \lambda_{B_k} (\sum_{t=1}^{T} B_{tk} - 1)$$

Take derivative with respect to π_k, A_{jk} and B_{tk}, let the equation equal to 0 and substitute the λ into the previous equations, then we can obtain:

$$\frac{\partial L}{\partial \pi_k} = \frac{\partial}{\partial \pi_k} (\sum_{k=1}^{K} \gamma(s_{1k}) \ln \pi_k) + \lambda_\pi = 0$$
$$\Rightarrow \lambda_\pi = -(\sum_{k=1}^{K} \gamma(s_{1k}))$$
$$\Rightarrow \pi_k = \frac{\gamma(s_{1k})}{\sum_{k=1}^{K} \gamma(s_{1k})}$$

$$\frac{\partial L}{\partial A_{jk}} = \frac{\partial}{\partial A_{jk}} (\sum_{t=2}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \zeta(s_{t-1,j} s_{tk}) \ln A_{jk}) + \lambda_{A_j}$$
$$= \frac{\sum_{t=2}^{T} \zeta(s_{t-1,j} s_{tk})}{A_{jk}} + \lambda_{A_j} = 0$$

sum over k to eliminate A_{jk}

$$\Rightarrow \sum_{t=2}^{T} \sum_{k=1}^{K} \zeta(s_{t-1,j} s_{tk}) + \lambda_{A_j} \sum_{k=1}^{K} A_{jk} = 0$$
$$\Rightarrow \lambda_{A_j} = -\sum_{t=2}^{T} \sum_{k=1}^{K} \zeta(s_{t-1,j} s_{tk})$$
$$\Rightarrow A_{jk} = \frac{\sum_{t=2}^{T} \zeta(s_{t-1,j} s_{tk})}{\sum_{t=2}^{T} \sum_{k=1}^{K} \zeta(s_{t-1,j} s_{tk})}$$

Because the observed sequences is in Gaussian distribution and $B_{tk} = p(x_t|s_t) = \mathcal{N}(\mu_k, \sigma_k^2)$

$$L = \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma(s_{tk}) \ln \mathcal{N}(\mu_k, \sigma_k^2) = \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma(s_{tk}) (-\frac{\ln 2\pi}{2} - \ln \sigma_k - \frac{(x_t - \mu_k)^2}{2\sigma_k^2})$$

Take derivative with respect to μ_k:

$$\frac{\partial L}{\partial \mu_k} = \sum_{t=1}^{T} \gamma(s_{tk}) \frac{x_t - \mu_k}{\sigma_k^2} = 0$$
$$\Rightarrow \mu_k = \frac{\sum_{t=1}^{T} \gamma(s_{tk}) x_t}{\sum_{t=1}^{T} \gamma(s_{tk})}$$

Take derivative with respect to σ_k:

$$\frac{\partial L}{\partial \sigma_k} = \sum_{t=1}^{T} \gamma(s_{tk})(-\frac{1}{\sigma_k} + \frac{(x_t - \mu_k)^2}{\sigma_k^3}) = 0$$
$$\Rightarrow \sum_{t=1}^{T} \gamma(s_{tk})(x_t - \mu_k)^2 = \sum_{t=1}^{T} \gamma(s_{tk}) \sigma_k^2$$
$$\Rightarrow \sigma_k = \sqrt{\frac{\sum_{t=1}^{T} \gamma(s_{tk})(x_t - \mu_k)^2}{\sum_{t=1}^{T} \gamma(s_{tk})}}$$

(d) For each of the following sentences, say whether it is true or false and provide a short explanation (one sentence or so). [16 pts]

- The weights of all incoming edges to a state of an HMM must sum to 1.
- An edge from state s to state t in an HMM denotes the conditional probability of going to state s given that we are currently at state t.
- The "Markov" property of an HMM implies that we cannot use an HMM to model a process that depends on several time-steps in the past.
- The Baum-Welch algorithm is a type of an Expectation Maximization algorithm and as such it is guaranteed to converge to the (globally) optimal solution.

(d)Answer:

- False: The weights of all incoming edges to a state of an HMM doesn't need to sum to 1.
- False: An edge from state s to state t denotes as $p(t|s)$ in Graphical Hidden Markov Model.
- True: The Markov property is that the conditional probability distribution of future states of the process only depends on the present state, not on the sequence of events that happened before it.
- False: The Baum-Welch algorithm is a type of an Expectation Maximization algorithm, but it can only guarantee to converge to local optimal solution instead of global optimal solution.

4 Programming [30 pts]

In this problem, you will implement algorithm to analyze the behavior of *SP500* index over a period of time. For each week, we measure the price movement relative to the previous week and denote it using a binary variable (+1 indicates up and 1 indicates down). The price movements from week 1 (the week of January 5) to week 39 (the week of September 28) are plotted below.

Consider a Hidden Markov Model in which x_t denotes the economic state (good or bad) of week t and y_t denotes the price movement (up or down) of the *SP500* index. We assume that $x_{(t+1)} = x_t$ with probability 0.8, and $P_{(Y_t|X_t)}(y_t = +1|x_t = \text{good}) = P_{(Y_t|X_t)}(y_t = -1|x_t = \text{bad}) = q$. In addition, assume that $P_{(X_1)}(x_1 = \text{bad}) = 0.8$. Load the `sp500.mat`, implement the algorithm, briefly describe how you implement this and report the following :

(a) Assuming $q = 0.7$, plot $P_{(X_t|Y)}(x_t = \textbf{good}|y)$ for $t = 1, 2, ..., 39$. What is the probability that the economy is in a good state in the week of week 39. [15 pts]

(a)Answer:
For the implement detail of the programming part, the code is followed the notation and procedure from our lecture slides:

(1)Since we need to plot the probability of good state in 39 weeks, then we need to calculate for 39 individual states instead of a 39-state sequence. Thus this algorithm implements forward and backward algorithm to solve this problem. It includes two parts, the forward algorithm and backward algorithm.

(2)According to the price_move of sp500.mat, we can notice that there are two possible outcome of y which are +1 and -1, and two possible outcome of x which are good and bad during 39 weeks. Thus i(+1/-1) can be 1 or 2, k(good/bad) can be 1 or 2 and t can take from 1 to 39.

Figure 1 shows the probability that the economy is in a good state during 39 weeks. And the probability that the economy is in a good state in the 39th week is

$$P_{(X_{39}|Y)}(x_{39} = \text{good}|y) = 0.683.$$

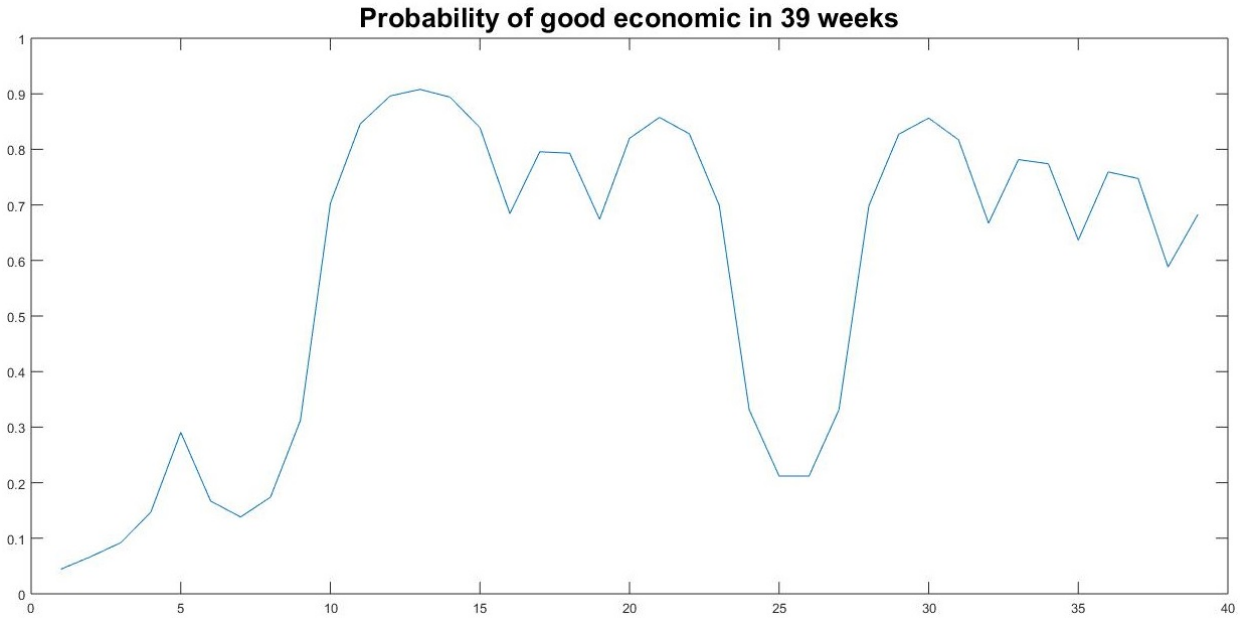

Figure 1: Probability of good economic state in 39 weeks with q=0.7

(b) Repeat (a) for $q = 0.9$, and compare the result to that of (a). Explain your comparison in one or two sentences. [15 pts]

(b)Answer:
Figure 2 shows the probability that the economy is in a good state during 39 weeks. And the probability that the economy is in a good state in the 39th week is

$$P_{(X_{39}|Y)}(x_{39} = \text{good}|y) = 0.838.$$

Figure 3 shows the comparison of the probability with $q = 0.7$ and $q = 0.9$ respectively.

Result of comparison:
According to the figure 3, we can easily notice that the uncertainty of orange plot($q = 0.9$) is less than the blue plot($q = 0.7$), which means if we have a higher probability of q, then we are more confident to determine the if economy is in good state or not along with the price movement.

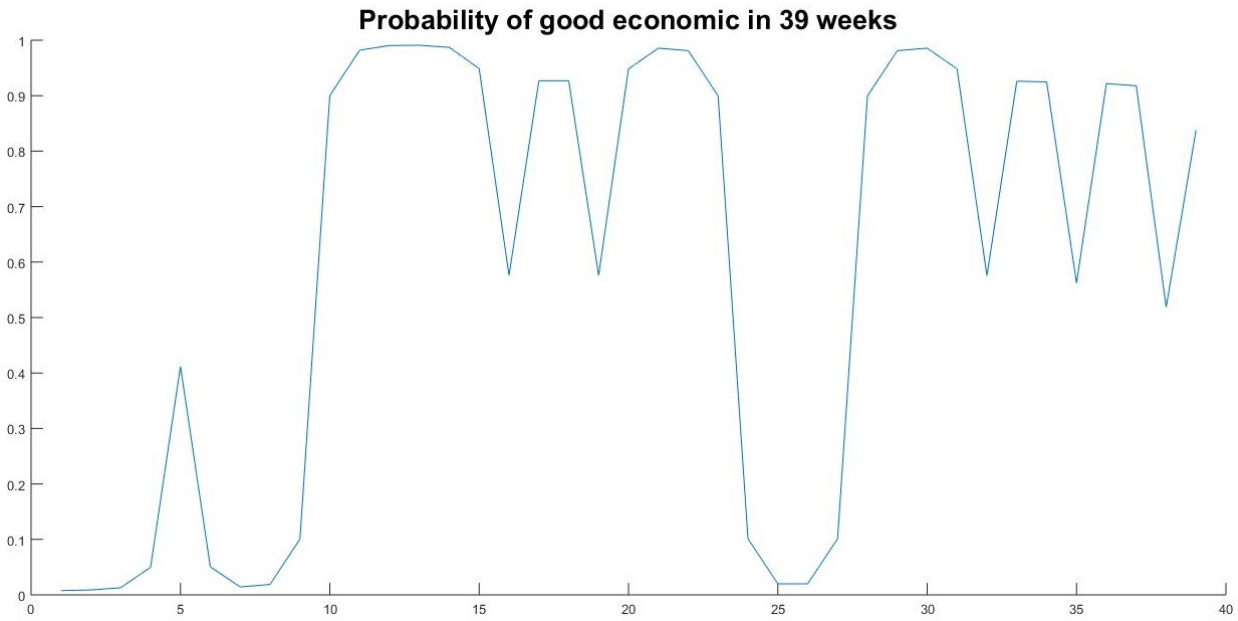

Figure 2: Probability of good economic state in 39 weeks with q=0.9

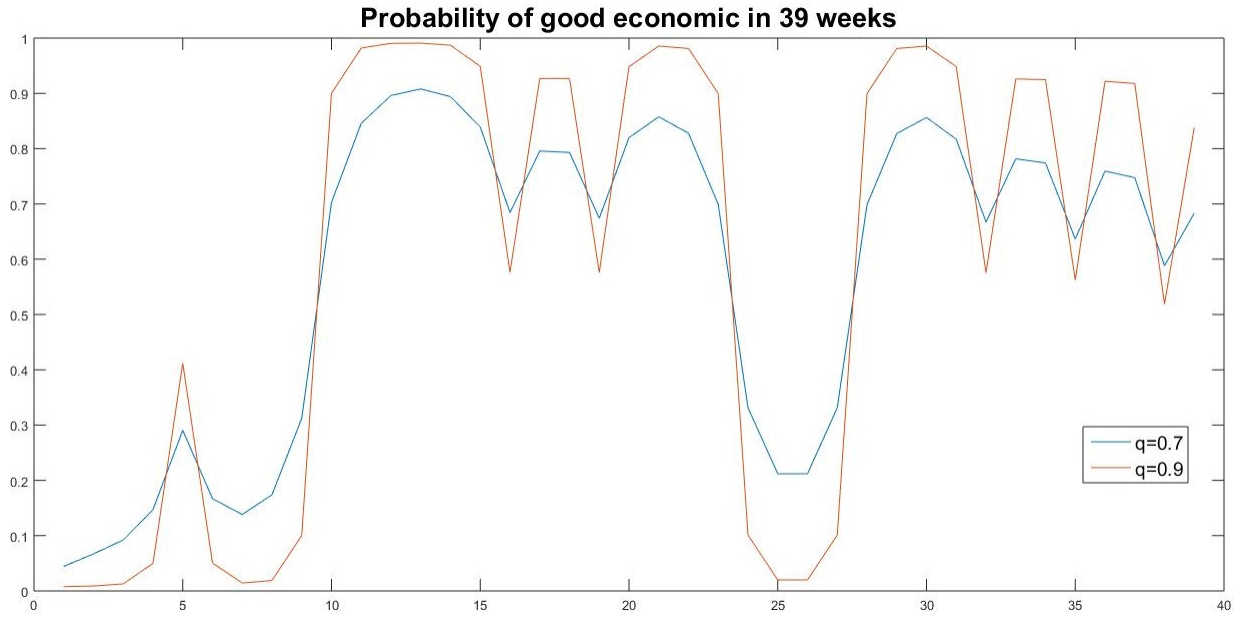

Figure 3: Comparison of probability of good economic state in 39 between q=0.7 and q=0.9 weeks