CS 7641 CSE/ISYE 6740 Homework 4 Solutions

Le Song

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.

FALSE.

Symmetry: $K(u, v) = \alpha K_1(u, v) + \beta K_2(u, v) = \alpha K_1(v, u) + \beta K_2(v, u) = K(v, u)$

Positve semidefinite: $K \in R^{m \times m}$, $K = \alpha K_1 + \beta K_2$:

$\forall z \in R^m, z^T K z = \alpha z^T K_1 z + \beta z^T K_2 z$

Given that $z^T K_1 z \geq 0, z^T K_2 z \geq 0$, $\alpha \in R, \beta \in R$,

we cannot determine $sign\{\alpha z^T K_1 z + \beta z^T K_2 z\}$

Which violate Mercer's sufficient and necessary condition, so this is not a valid kernel function.

2.

TRUE.

Symmetry: $K(u, v) = K_1(f(u), f(v)) = K_1(f(v), f(u)) = K(v, u)$

Positve semidefinite: $\forall z \in R^m, z^T K z = \sum_{i=1}^m \sum_{j=1}^m z_i K(u_i, u_j) z_j = \sum_{i=1}^m \sum_{j=1}^m z_i K_1(f(u_i), f(u_j))$
$= \sum_{i=1}^m \sum_{j=1}^m z_i \left(\Phi(f(u_i))^T \Phi(f(u_j))\right) z_j = \sum_{i=1}^m \sum_{j=1}^m z_i \left(\sum_{k=1}^m \Phi_k(f(u_i)) \cdot \Phi_k(f(u_j))\right) z_j$
$= \sum_{k=1}^m \sum_{i=1}^m \sum_{j=1}^m z_i \Phi_k(f(u_i)) \cdot \Phi_k(f(u_j)) z_j = \sum_{k=1}^m (z_i \Phi_k(f(u_i)))^2 \geq 0$
$\Rightarrow$ Positve semidefinite!
So, this kernel function is a valid kernel function.

3.

FALSE.
Symmetry:

$$K(u,v) = \begin{cases} 1 & \text{if } \|u-v\|_2 \leqslant 1 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 1 & \text{if } \|v-u\|_2 \leqslant 1 \\ 0 & \text{otherwise} \end{cases} = K(v,u) \tag{2}$$

PSD: $u_1 = (0,0)$, $u_2 = (0,2)$, $u_3 = (0,1)$
$K(u_1,u_1) = 1$, $K(u_1,u_2) = 0$, $K(u_1,u_3) = 1$,
$K(u_2,u_1) = 0$, $K(u_2,u_2) = 1$, $K(u_2,u_3) = 1$,
$K(u_3,u_1) = 1$, $K(u_3,u_2) = 1$, $K(u_3,u_3) = 1$
$z^T K z = z_1^2 + z_1 z_2 + z_1 z_3 + z_2^2 + z_2 z_3 + z_3 z_1 + z_3 z_2 + z_3^2$, and we cannot determine its sign, so not PSD!
Since we find a contradiction, so that this kernel function is not valid.

4.

Symmetry:

$$K(u,v) = \frac{K'(u,v)}{\sqrt{K'(u,u)K'(v,v)}} = \frac{K'(v,u)}{\sqrt{K'(v,v)K'(u,u)}} = K(v,u)$$

PSD:

$$z^T K z = \sum_{i=1}^m \sum_{j=1}^m z_i \frac{K'(u_i,u_j)}{\sqrt{K'(u_i,u_i)K'(u_j,u_j)}} z_j = \sum_{i=1}^m \sum_{j=1}^m z_i \frac{\sum_{k=1}^m \Phi_k(u_i) \cdot \Phi_k(u_j)}{\sqrt{K'(u_i,u_i)K'(u_j,u_j)}} z_j$$

$$= \sum_{k=1}^m \sum_{i=1}^m \sum_{j=1}^m z_i \frac{\Phi_k(u_i)}{\sqrt{K'(u_i,u_i)}} z_j \frac{\Phi_k(u_j)}{\sqrt{K'(u_j,u_j)}} = \sum_{k=1}^m \left(z_i \frac{\Phi_k(u_i)}{\sqrt{K'(u_i,u_i)}} \right)^2 \geq 0$$

$\Rightarrow$ PSD! $\Rightarrow$ this kernel function is a valid kernel function.

(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]

As the discriminant function of Fisher Discriminant Analysis is of the form $w^T x^*$ for new data point x^* and fixed w^T, we see that after the non-linear mapping using a function ϕ, w will be of the form:

$$w^K = \sum_{i=1}^{N} \alpha_i \phi(x_i)$$

To find the discriminant vectors, we apply the kernel trick to our data:

$$\mu_i^K = \frac{1}{n_i} \sum_{j=1}^{n_i} \phi(x_j)$$

While it is impractical to calculate $\phi(x_j)$ directly, we can avoid it by only looking at the inner products:

$$\begin{aligned}(w^K)^T \mu_i^K &= \frac{1}{n_i} \sum_{j=1}^{N} \sum_{k=1}^{n_i} \alpha_j k(x_j, x_k) \\ &= \alpha^{\mathbf{T}} \mathbf{M_i^K}\end{aligned}$$

The complete solution for both 2-class and multi-class kernelized version of LDA with all the steps can be found at `https://en.wikipedia.org/wiki/Kernel_Fisher_discriminant_analysis`
The original paper describing the method can be found at `http://courses.cs.tamu.edu/rgutier/csce666_f16/mika1999kernelLDA.pdf`

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]

$\psi(a, b = 1, c) = \phi_1(a, b = 1)\phi_2(b = 1, c)$

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

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

This is actually a marginal distribution. We first facterize it.

$$P(a=1,b=1) = \sum_c P(a=1,b=1,c) = \sum_c \frac{\phi_1(a=1,b=1)\cdot\phi_2(b=1,c)}{Z},$$

where $Z = \sum_{a,b,c}\prod_{i=1}^{2}\phi_i(x_i) = 74$. Therefore,

\psi (a, b=1, c) + \psi (a, b=0, c) = 74

$$\sum_c \frac{\phi_1(a=1,b=1)\cdot\phi_2(b=1,c)}{Z} = \frac{4+1+3}{74} = \frac{4}{37}.$$

(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

	Conditional Random Fields	Hidden Markov Model
Type of model	discriminative	generative
Objective function optimized	$P(Z\|X)$	$P(X, Z)$
Require a normalization constant	Yes	No

- Type of model: Because CRF descibes directly how to take a feature vector x and assign it a label y. However, HMM describes how a label vector y can probabilistically generate a feature vector x.
- Objective function optimized: CRF optimizes the conditional likelihood of the lables given the observed data. HMM optimizes a joint likelihood.
- Require a normalization constant: CRF: See as in (b). Z is the number, because the target is the probability so we need a normalizer; while HMM, the target is a set of parameters and when we doing optimization, it is irrelavent to the normalizer so we can get rid of it.

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]

$\vec{O} = H, T, H$, $\vec{Z} = Z_1, Z_2, Z_3$

$$P(\vec{O}; A, B, \pi) = \sum_{\vec{Z}} P(\vec{O}, \vec{Z}; A, B, \pi) = \sum_{\vec{Z}} P(Z_1)P(O_1|Z_1)\prod_{t=2}^{T} P(Z_t|Z_{t-1})\prod_{t=2}^{T} P(O_t|Z_t)$$

$$= \sum_{\vec{V}} P(Z_1)P(O_1|Z_1)P(Z_2|Z_1)P(Z_3|Z_2)P(O_2|Z_2)P(O_3|Z_3)$$

$$= \sum_{Z_3=1}^{3}\sum_{Z_2=1}^{3}\sum_{Z_1=1}^{3} P(Z_1)P(O_1|Z_1)P(Z_2|Z_1)P(Z_3|Z_2)P(O_2|Z_2)P(O_3|Z_3)$$

$$= \sum_{Z_3=1}^{3} P(O_3|Z_3)\sum_{Z_2=1}^{3} P(Z_3|Z_2)P(O_2|Z_2)\sum_{Z_1=1}^{3} P(Z_2|Z_1)P(O_1|Z_1)P(Z_1)$$

$$= \sum_{Z_3=1}^{3} B_{HZ_3}\sum_{Z_2=1}^{3} A_{Z_2Z_3}B_{TZ_2}\sum_{Z_1=1}^{3} A_{Z_1Z_2}B_{HZ_1}\pi(Z_1)$$

$\alpha_1(Z_1) = B_{HZ_1}\pi(Z_1) = [1/6, 1/4, 1/12]$

$\alpha_2(Z_2) = B_{TZ_2}\sum_{Z_1=1}^{3} A_{Z_1Z_2} \cdot \alpha_1(Z_1) = [3/20, 17/960, 31/320]$

$\alpha_3(Z_3) = B_{HZ_3}\sum_{Z_2=1}^{3} A_{Z_2Z_3} \cdot \alpha_2(Z_2) = [0.093, 0.0396, 0.006289]$

$P(\vec{O}; A, B, \pi) = \sum_{Z_3=1}^{3} \alpha_3(Z_3) \approx 0.139$

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

If we have a sequence $\vec{X}$, like (a) we have $\vec{X} = \{H, T, H\}$. To estimate A, B, and π, we are going to use an adaption of EM algorithm.

Our optimization objective function:

$$Q(\theta) = \sum_{\vec{Z}} P(\vec{Z}|\vec{X};\theta) \ln P(\vec{X}, \vec{Z};\theta)$$

Because this objective function cannot be optimized under the fact that we do not know about θ (the first term conditional on θ, so that we cannot calculate it out until we have a guess for θ), so instead we use the EM approach.

$$Q(\theta, \theta^s) = \sum_{\vec{Z}} P(\vec{Z}|\vec{X};\theta^s) \ln P(\vec{X}, \vec{Z};\theta)$$

So the procedule is that:

Def 1: $\gamma(Z_t) = P(Z_t|\vec{X};\theta^s)$

Def 2: $\xi(Z_{t-1}, Z_t) = P(Z_{t-1}, Z_t|\vec{X};\theta^s)$

Def 3: $\gamma(\vec{Z}) = P(\vec{Z}|\vec{X};\theta^s)$

Corollary 1: $\gamma(Z_t^i) = E[Z_t^i] = \sum_{\vec{Z}} P(\vec{Z}|\vec{X};\theta^s) Z_t^i = \sum_{\vec{Z}} \gamma(\vec{Z}) Z_t^i$

Corollary 2: $\xi(Z_{t-1}^i, Z_t^j) = E[Z_{t-1}^i \cdot Z_t^j] = \sum_{\vec{Z}} P(\vec{Z}|\vec{X};\theta^s) Z_{t-1}^i \cdot Z_t^j = \sum_{\vec{Z}} \gamma(\vec{Z}) Z_{t-1}^i \cdot Z_t^j$

Step 1: E step For $Z_t \in \vec{Z}$, compute

$$\gamma(Z_t) = P(Z_t|\vec{X};\theta^s) = \frac{\alpha_t(Z_t)\beta_t(Z_t)}{P(\vec{X})}$$

$$\xi(Z_{t-1}, Z_t) = P(Z_{t-1}, Z_t|\vec{X};\theta^s) = \frac{\alpha_{t-1}(Z_{t-1})P(X_t|Z_t)P(Z_t|Z_{t-1})\beta_t(Z_t)}{P(\vec{X})}$$

Where $\alpha_t(Z_t) = P(X_{1:t}, Z_t)$, and it can be obtained by foward algorithm
$\beta_t(Z_t) = P(X_{t+1:T}|Z_t)$, and this can be obtained by backward algorithm.

Step 2: M Step Update $\theta = \text{argmax}_\theta Q(\theta, \theta^s) = \text{argmax}_\theta \sum_{\vec{Z}} P(\vec{Z}|\vec{X}; \theta^s) \ln P(\vec{X}, \vec{Z}; \theta)$

$$\hat{\pi}_i = \frac{\gamma(Z_1^i)}{\sum_{j=1}^m \gamma(Z_1^j)}$$

$$\hat{A}_{ij} = \frac{\sum_{t=2}^T \xi(Z_{t-1}^i, Z_t^j)}{\sum_{j=1}^m \sum_{t=2}^T \xi(Z_{t-1}^i, Z_t^j)}$$

$$\hat{B}_{jk} = \frac{\sum_{t=1}^T \gamma(Z_t^j)}{\sum_{l=1}^K \sum_{t=1}^T \gamma(Z_t^j)}$$

Step 3 Repeat Step 1, and Step 2 until converge.

Proof:

$$P(\vec{X}, \vec{Z}; \theta) = P(Z_1)P(X_1|Z_1) \prod_{t=2}^T P(Z_t|Z_{t-1})P(X_t|Z_t)$$

$$= \prod_{i=1}^m \pi_i^{Z_1^i} \cdot \prod_{t=2}^T \prod_{i=1}^m \prod_{j=1}^m A_{ij}^{Z_{t-1}^i Z_t^j} \cdot \prod_{t=1}^T \prod_{j=1}^m \prod_{k=1}^K B_{jk}^{Z_t^j X_t^k}$$

$$\ln P(\vec{X}, \vec{Z}; \theta) = \sum_{i=1}^m Z_1^i \ln \pi_i + \sum_{t=2}^T \sum_{i=1}^m \sum_{j=1}^m Z_{t-1}^i Z_t^j \ln A_{ij} + \sum_{t=1}^T \sum_{j=1}^m \sum_{k=1}^K Z_t^j X_t^k \ln B_{jk}$$

s.t. $\sum_{i=1}^m \pi_i = 1$, $\sum_{j=1}^m A_{ij} = 1$, $\sum_{k=1}^K B_{jk} = 1$

The Largrangian is

$$L(\theta, \theta^s) = \sum_{\vec{Z}} \gamma(\vec{Z}) \ln P(\vec{X}, \vec{Z}; \theta) + \lambda_\pi (1 - \sum_{i=1}^m \pi_i) + \sum_{i=1}^m \lambda_{A_i} (1 - \sum_{j=1}^m A_{ij}) + \sum_{j=1}^m \lambda_{B_j} (1 - \sum_{k=1}^K B_{jk})$$

$$\frac{\partial L(\theta, \theta^s)}{\partial \pi_i} = \frac{\partial}{\partial \pi_i} \left(\sum_{i=1}^m \sum_{\vec{Z}} \gamma(\vec{Z}) Z_1^i \ln \pi_i + \lambda_\pi (1 - \sum_{i=1}^m \pi_i) \right) = \frac{\gamma(Z_1^i)}{\pi_i} - \lambda_\pi = 0$$

$$\sum_{i=1}^m \pi_i = \sum_{i=1}^m \frac{\gamma(Z_1^i)}{\lambda_\pi} = 1 \Rightarrow \lambda_\pi = \sum_{i=1}^m \gamma(Z_1^i)$$

$$\hat{\pi}_i = \frac{\gamma(Z_1^i)}{\sum_{j=1}^m \gamma(Z_1^j)}$$

$$\frac{\partial L(\theta, \theta^s)}{\partial A_{ij}} = \frac{\partial}{\partial A_{ij}} \left(\sum_{t=2}^{T} \sum_{i=1}^{m} \sum_{j=1}^{m} \xi(Z_{t-1}^i, Z_t^j) \ln A_{ij} + \sum_{i=1}^{m} \lambda_{A_i} (1 - \sum_{j=1}^{m} A_{ij}) \right)$$

$$= \frac{\sum_{t=2}^{T} \xi(Z_{t-1}^i, Z_t^j)}{A_{ij}} - \lambda_{A_i} = 0$$

$$\sum_{j=1} A_{ij} = 1 \Rightarrow \lambda_{A_i} = \sum_{j=1}^{m} \sum_{t=2}^{T} \xi(Z_{t-1}^i, Z_t^j)$$

$$\hat{A}_{ij} = \frac{\sum_{t=2}^{T} \xi(Z_{t-1}^i, Z_t^j)}{\sum_{j=1}^{m} \sum_{t=2}^{T} \xi(Z_{t-1}^i, Z_t^j)}$$

$$\frac{\partial L(\theta, \theta^s)}{\partial B_{jk}} = \frac{\partial}{\partial B_{jk}} \left(\sum_{t=1}^{T} \sum_{j=1}^{m} \sum_{k=1}^{K} \gamma(Z_t^j) X_t^k \ln B_{jk} + \sum_{j=1}^{m} \lambda_{B_j} (1 - \sum_{k=1}^{K} B_{jk}) \right)$$

$$= \frac{\sum_{t=1}^{T} \gamma(Z_t^j) X_t^k}{B_{jk}} - \lambda_{B_j} = 0$$

$$\sum_{k=1}^{K} B_{jk} = \sum_{k=1}^{K} \frac{\sum_{t=1}^{T} \gamma(Z_t^j) X_t^k}{\lambda_{B_j}} = 1$$

$$\Rightarrow \lambda_{B_j} = \sum_{k=1}^{K} \sum_{t=1}^{T} \gamma(Z_t^j) X_t^k$$

$$\hat{B}_{jk} = \frac{\sum_{t=1}^{T} \gamma(Z_t^j) X_t^k}{\sum_{l=1}^{K} \sum_{t=1}^{T} \gamma(Z_t^j) X_t^l}$$

(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]

For convenience, I use Z_t to mean for S_t

$$Q(\theta, \theta^s) = \sum_{\vec{Z}^{(1)},...\vec{Z}^{(M)}} P(\vec{Z}^{(1)}, ...\vec{Z}^{(M)}|\vec{X}^{(1)}, ...\vec{X}^{(M)}; \theta^s) \ln P(\vec{X}^{(1)}, ...\vec{X}^{(M)}, \vec{Z}^{(1)}, ...\vec{Z}^{(M)}; \theta)$$

$$= \sum_{n=1}^{M} \sum_{\vec{Z}^{(n)}} P(\vec{Z}^{(n)}|\vec{X}^{(n)}; \theta^s) \ln P(\vec{X}^{(n)}, \vec{Z}^{(n)}; \theta)$$

We see that this objective function look almost identical to the previous one, except now it is a summation of m expectation of log-likelihood. As well as the emission probability now is Gussian distribution.

$$P(\vec{X}^{(n)}, \vec{Z}^{(n)}; \theta) = P(Z_1^{(i)}) P(X_1^{(n)}|Z_1^{(n)}) \prod_{t=2}^{T} P(Z_t^{(n)}|Z_{t-1}^{(n)}) P(X_t^{(n)}|Z_t^{(n)})$$

$$= \prod_{i=1}^{m} \pi_i^{Z_1^{i\,(n)}} \cdot \prod_{t=2}^{T} \prod_{i=1}^{m} \prod_{j=1}^{m} A_{ij}^{Z_{t-1}^{i\,(n)} Z_t^{j\,(n)}} \cdot \prod_{t=1}^{T} P(X_t^{(n)}|Z_t^{(n)})$$

$$\ln P(\vec{X}^{(n)}, \vec{Z}^{(n)}; \theta) = \sum_{i=1}^{m} Z_1^{i\,(n)} \ln \pi_i + \sum_{t=2}^{T} \sum_{i=1}^{m} \sum_{j=1}^{m} Z_{t-1}^{i\,(n)} Z_t^{j\,(n)} \ln A_{ij} + \sum_{t=1}^{T} \ln P(X_t^{(n)} | Z_t^{(n)})$$

s.t. $\sum_{i=1}^{m} \pi_i = 1$, $\sum_{j=1}^{m} A_{ij} = 1$, $\sum_{X_t^{(n)}} P(X_t^{(n)} | Z_t^{(n)}) = 1$

The Largrangian is

$$L(\theta, \theta^s) = \sum_{n=1}^{M} \sum_{\vec{Z}^{(n)}} \ln P(\vec{X}^{(n)}, \vec{Z}^{(n)}; \theta) + \lambda_\pi (1 - \sum_{i=1}^{m} \pi_i) + \sum_{i=1}^{m} \lambda_{A_i} (1 - \sum_{j=1}^{m} A_{ij}) + \sum_{j=1}^{m} \lambda_X (1 - \sum_{X_t^{(n)}} P(X_t^{(n)} | Z_t^{(n)}))$$

Def 1: $\gamma^{(n)}(Z_t^{(n)}) = P(Z_t^{(n)} | \vec{X}^{(n)}; \theta^s)$

Def 2: $\xi^{(n)}(Z_{t-1}^{(n)}, Z_t^{(n)}) = P(Z_{t-1}^{(n)}, Z_t^{(n)} | \vec{X}^{(n)}; \theta^s)$

Take derivatives the same as in the proof of (b).

$$\frac{\partial L(\theta, \theta^s)}{\partial \pi_i} = \frac{\partial}{\partial \pi_i} \left(\sum_{i=1}^{m} \sum_{n=1}^{M} \sum_{\vec{Z}^{(n)}} \gamma(\vec{Z}^{(n)}) Z_1^{i\,(n)} \ln \pi_i + \lambda_\pi (1 - \sum_{i=1}^{m} \pi_i) \right) = \frac{\sum_{n=1}^{M} \gamma(Z_1^{i\,(n)})}{\pi_i} - \lambda_\pi = 0$$

$$\sum_{i=1}^{m} \pi_i = \sum_{i=1}^{m} \frac{\sum_{n=1}^{M} \gamma(Z_1^{i\,(n)})}{\lambda_\pi} = 1 \Rightarrow \lambda_\pi = \sum_{i=1}^{m} \sum_{n=1}^{M} \gamma(Z_1^{i\,(n)})$$

$$\hat{\pi}_i = \frac{\sum_{n=1}^{M} \gamma(Z_1^{i\,(n)})}{\sum_{n=1}^{M} \sum_{j=1}^{m} \gamma(Z_1^{j\,(n)})}$$

$$\frac{\partial L(\theta, \theta^s)}{\partial A_{ij}} = \frac{\partial}{\partial A_{ij}} \left(\sum_{t=2}^{T} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{n=1}^{M} \xi^{(n)}(Z_{t-1}^{i\,(n)}, Z_t^{j\,(n)}) \ln A_{ij} + \sum_{i=1}^{m} \lambda_{A_i} (1 - \sum_{j=1}^{m} A_{ij}) \right)$$

$$= \frac{\sum_{t=2}^{T} \sum_{n=1}^{M} \xi^{(n)}(Z_{t-1}^{i\,(n)}, Z_t^{j\,(n)})}{A_{ij}} - \lambda_{A_i} = 0$$

$$\sum_{j=1} A_{ij} = 1 \Rightarrow \lambda_{A_i} = \sum_{n=1}^{M} \sum_{j=1}^{m} \sum_{t=2}^{T} \xi^{(n)}(Z_{t-1}^{i\,(n)}, Z_t^{j\,(n)})$$

$$\hat{A}_{ij} = \frac{\sum_{t=2}^{T} \sum_{n=1}^{M} \xi^{(n)}(Z_{t-1}^{i\,(n)}, Z_t^{j\,(n)})}{\sum_{n=1}^{M} \sum_{j=1}^{m} \sum_{t=2}^{T} \xi^{(n)}(Z_{t-1}^{i\,(n)}, Z_t^{j\,(n)})}$$

$$\frac{\partial L(\theta, \theta^s)}{\partial \mu_i} = \sum_{n=1}^{M}\sum_{t=1}^{T} \gamma^{(n)}(Z_t^{i(n)})(-\frac{1}{2\sigma_i^2})2(X_t^{(n)} - \mu_i)(-1) = 0$$

$$\hat{\mu}_i = \frac{\sum_{n=1}^{M}\sum_{t=1}^{T} \gamma^{(n)}(Z_t^{i(n)}) X_t^{(n)}}{\sum_{n=1}^{M}\sum_{t=1}^{T} \gamma^{(n)}(Z_t^{i(n)})}$$

$$\frac{\partial L(\theta, \theta^s)}{\partial \sigma_i^2} = \sum_{n=1}^{M}\sum_{t=1}^{T} \gamma^{(n)}(Z_t^{i(n)})(-\frac{-\frac{1}{2}(2\pi\sigma_i^2)^{-3/2}\cdot 2\pi}{(2\pi\sigma_i^2)^{-1/2}}) - \frac{1}{2}(X_t^{(n)} - \mu_i)^2(-1)(-\sigma_i^2)^{-2} = 0$$

$$\hat{\sigma}_i = \frac{\sum_{n=1}^{M}\sum_{t=1}^{T} \gamma^{(n)}(Z_t^{i(n)})(X_t^{(n)} - \mu_i)^2}{\sum_{n=1}^{M}\sum_{t=1}^{T} \gamma^{(n)}(Z_t^{i(n)})}$$

Procedule:

Step 1: E step For $Z_t^{(n)} \in \vec{Z}$, compute

$$\gamma^{(n)}(Z_t^{(n)}) = P(Z_t^{(n)}|\vec{X}^{(n)}; \theta^s) = \frac{\alpha_t(Z_t^{(n)})\beta_t(Z_t^{(n)})}{P(\vec{X}^{(n)})}$$

$$\xi(Z_{t-1}^{(n)}, Z_t^{(n)}) = P(Z_{t-1}^{(n)}, Z_t^{(n)}|\vec{X}^{(n)}; \theta^s) = \frac{\alpha_{t-1}(Z_{t-1}^{(n)})P(X_t^{(n)}|Z_t^{(n)})P(Z_t^{(n)}|Z_{t-1}^{(n)})\beta_t(Z_t^{(n)})}{P(\vec{X}^{(n)})}$$

Where $\alpha_t(Z_t^{(n)}) = P(X_{1:t}^{(n)}, Z_t^{(n)})$, and it can be obtained by foward algorithm
$\beta_t(Z_t^{(n)}) = P(X_{t+1:T}^{(n)}|Z_t^{(n)})$, and this can be obtained by backward algorithm.

Step 2: M Step Update $\theta = \text{argmax}_\theta Q(\theta, \theta^s) = \text{argmax}_\theta \sum_{\vec{Z}^{(n)}} P(\vec{Z}^{(n)}|\vec{X}^{(n)}; \theta^s) \ln P(\vec{X}^{(n)}, \vec{Z}^{(n)}; \theta)$

$$\hat{\pi}_i = \frac{\sum_{n=1}^{M} \gamma(Z_1^{i(n)})}{\sum_{n=1}^{M} \sum_{j=1}^{m} \gamma(Z_1^{j(n)})}$$

$$\hat{A}_{ij} = \frac{\sum_{t=2}^{T} \sum_{n=1}^{M} \xi^{(n)}(Z_{t-1}^{i\ (n)}, Z_t^{j(n)})}{\sum_{n=1}^{M} \sum_{j=1}^{m} \sum_{t=2}^{T} \xi^{(n)}(Z_{t-1}^{i\ (n)}, Z_t^{j(n)})}$$

$$\hat{\mu}_i = \frac{\sum_{n=1}^{M} \sum_{t=1}^{T} \gamma^{(n)}(Z_t^{i(n)}) X_t^{(n)}}{\sum_{n=1}^{M} \sum_{t=1}^{T} \gamma^{(n)}(Z_t^{i(n)})}$$

$$\hat{\sigma}_i = \frac{\sum_{n=1}^{M} \sum_{t=1}^{T} \gamma^{(n)}(Z_t^{i(n)})(X_t^{(n)} - \mu_i)^2}{\sum_{n=1}^{M} \sum_{t=1}^{T} \gamma^{(n)}(Z_t^{i(n)})}$$

Step 3 Repeat Step 1, and Step 2 until converge.

(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.

  FALSE.

  $\sum_{i=1}^{m} P(Z_t|Z_{t-1}^i) \neq 1$

- 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.

  FALSE.

  An edge from state s to state t denotes the conditional probability of going to t given that we are currently at state s.

- 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.

  TRUE.

  The "Markov" property has limited horizon, which means the probability current state is only depend on the last state, for the previous states, it is independent.

- 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.

  FALSE.

  The Baum-Welch algorithm is guaranteed to increase the log likelihood of the model however it is not guaranteed to find globally best parameters. It may converges to local optimum, depending on initial conditions

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]

0.6830

(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]

0.8379

The conditional probability of good economy gets more volatility when q becomes larger. This is because when q becomes larger the correlation between stock price and economy state increases (emission probability)