CS 7641 CSE/ISYE 6740 Homework 4

Prakash, Fall 2021

Deadline: 12/02, 11:59 pm

- Submit your answers as an electronic copy on Gradescope.
- No unapproved extension of deadline is allowed. Late submission will lead to 0 credit.
- For typed answers with LaTeX (recommended) or word processors, extra credits will be given (= 5 points). 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 (Collaborators - Kshitij Pisal). 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 (Sources - PRML, Wikipedia).

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. $K(u, v) = \alpha K_1(u, v) + \beta K_2(u, v), \alpha, \beta \in R$.
   - **False**. For $K(u, v)$ to be a valid kernel it has to be positive semi-definite and symmetric. (Mercer's necessary and sufficient condition).
   - $$\forall z \in R^m \; z^T K z = \alpha z^T K_1 z + \beta z^T K_2 z \longrightarrow (a)$$

     $$\text{Since } K_1 \text{ and } K_2 \text{ are valid kernels: } z^T K_1 z > 0; z^T K_2 z > 0 \longrightarrow (b)$$
   - if α and β are both negative then $z^T K z < 0$ (using (a) and (b)) which violates the required condition for a valid kernel function.
2. $K(u, v) = K_1(f(u), f(v))$ where $f : R^m \to R^m$.
   - **True**. It satisfies the required condition as below.

- $$K(u,v) = K_1(f(u), f(v)) = K_1(f(v), f(u)) = K(v,u) \longrightarrow (Symmetric)$$

  Since $f : R^m \to R^m$ is a transformation in the same domain K is simply a different kernel in that domain.

- $$K(u,v) = K_1(f(u), f(v)) = \langle \Phi(f(u)), \Phi(f(v)) \rangle = \langle \Phi_f(u), \Phi_f(v) \rangle \longrightarrow (PSD)$$

3.

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

- **False**.
- Let us consider $u1 = (0,0), u2 = (0,1)$ and $u3 = (1,-1)$, then K will be a 3 x 3 matrix given by: $K = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$

$$z^T K z = z_1^2 + 2z_1 z_2 + 2z_2 z_3 + z_2^2 + z_3^2$$

- For $z = (1,-1,1)$, $z^T K z = -1$ (not PSD). Since there is a contradiction K is not a valid kernel.

4. Suppose K' is a valid kernel.

$$K(u,v) = \frac{K'(u,v)}{\sqrt{K'(u,u)K'(v,v)}}. \tag{2}$$

- **True**.
- $$K(u,v) = \frac{K'(u,v)}{\sqrt{K'(u,u)K'(v,v)}} = \frac{K'(v,u)}{\sqrt{K'(u,u)K'(v,v)}} = K(v,u) \longrightarrow (Symmetric)$$

- Since K' is a valid kernel , it is positive semi definite. All diagonal elements of K' are positive. Let D be a diagonal matrix with all positive entries such that $D_{ii} = \frac{1}{\sqrt{K'_{ii}}}$

- Matrix K can be represented as : $K = D^T K' D$

  $|K| = |D^t||K||D| > 0$ PSD using K' is a valid kernel and D is a diagonal matrix with positive entries

(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]

Given N data points X_i and mapping Φ to feature space F. N_i is the number of points belonging to class C_i. To find the linear discriminant in F, we need to maximize:

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \text{ where:}$$

$$S_B = \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^T$$

$$S_W = \sum_{i=1}^{c} \sum_{n=1}^{N_i} (\phi(x_n^i) - \mu_i)(\phi(x_n^i) - \mu_i)^T$$

$$\mu_i = \frac{\sum_{n=1}^{N_i} \phi(x_n^i)}{N_i} \; ; \mu = \frac{\sum_{n=1}^{N} \phi(x_n)}{N}$$

w will have expansion of the form

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

We can avoid calculating $\phi(x_i)$ by using the kernel trick to replace inner product.

$$w^T \mu_i = \frac{1}{N_i} \sum_{j=1}^{N} \sum_{k=1}^{N_i} \alpha_j k(x_j, x_k^i) = \alpha^T M_i$$

The numerator is simplified as:

$$w^T S_B w = w^T \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^T w = A^T M A$$

where $A = [\alpha_1, \alpha_2 .... \alpha_{c-1}]$; $M = \sum_{j=1}^{c} N_j(M_j - M_*)(M_j - M_*)^T$; $(M_*)j = \frac{1}{N} \sum_{k=1}^{N} k(x_j, x_k)$

The denominator can be simplified as:

$$w^T S_W w = w^T \sum_{i=1}^{c} \sum_{n=1}^{N_i} \phi(x_n^i) - \mu_i)(\phi(x_n^i) - \mu_i)^T w = A^T N A$$

where $N = \sum_{j=1}^{c} K_j(I - 1_{N_j})K_j^T$; K_j is $N \times N_j$ matrix with $(K_j)_{nm} = k(x_n, x_m^j)$ (kernel matrix for class j)

Using the Kernel trick, Fisher's LDA simplifies to:

$$A^* = \underset{A}{\text{argmax}} \frac{|A^T M A|}{|A^T N A|}$$

- A^* can be computed by finding the leading $c-1$ eigen vectors of $N^{-1}M$
- The projection for new input x^t is given by $y(x^t) = (A^*)^T K_t$ where i^{th} entry of $K_t = k(x_i, x_t)$
- The new point will be classified according to:

$$\underset{j}{\text{argmin}} D(y(x), \bar{y_j})$$

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]

Ans:

- On fixing $b=1$, a can take 2 values 0 and 1 and c can take 3 values 0, 1 and 2. So, the table for $\psi(a,b=1,c)$ will contain 6 values.
- The different combinations of values that a and c can take are as follows:

$$\psi(a=0,b=1,c=0) = \phi_1(a=0,b=1) \times \phi_2(b=1,c=0) = 3 \times 4 = 12$$

$$\psi(a=0,b=1,c=1) = \phi_1(a=0,b=1) \times \phi_2(b=1,c=1) = 3 \times 1 = 3$$

$$\psi(a=0,b=1,c=2) = \phi_1(a=0,b=1) \times \phi_2(b=1,c=2) = 3 \times 3 = 9$$

$$\psi(a=1,b=1,c=0) = \phi_1(a=1,b=1) \times \phi_2(b=1,c=0) = 1 \times 4 = 4$$

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

$$\psi(a=1,b=1,c=2) = \phi_1(a=1,b=1) \times \phi_2(b=1,c=2) = 1 \times 3 = 3$$

- 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
0	1	2	9
1	1	0	4
1	1	1	1
1	1	2	3

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

- $$P(a = 1, b = 1) = \sum_c P(a = 1, b = 1, c)$$

↪ $$P = \frac{1}{Z}\psi(1, 1, 0) + \frac{1}{Z}\psi(1, 1, 1) + \frac{1}{Z}\psi(1, 1, 2)$$

↪ $$P = \frac{1}{Z}(4 + 1 + 3) = \frac{8}{Z}$$

- Now, we calculate the normalizing factor Z as follows:

↪ $$Z = \sum_{a,b,c} \psi(a, b, c) = \sum_{a,b,c} \phi_1(a, b)\phi_2(b, c)$$

↪ $$Z = \phi_1(0, 0)\sum_c \phi_2(0, c) + \phi_1(0, 1)\sum_c \phi_2(1, c) + \phi_1(1, 0)\sum_c \phi_2(0, c) + \phi_1(1, 1)\sum_c \phi_2(1, c)$$

↪ $$Z = 4(3 + 2 + 1) + 3(4 + 1 + 3) + 3(3 + 2 + 1) + 1(4 + 1 + 3) = 24 + 24 + 18 + 8 = 74$$

↪ Therefore, we have $P(a = 1, b = 1) = \frac{8}{74} = \frac{4}{37}$.

(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.

↪ Conditional random fields (CRF) is a discriminative model, it learns the conditional distribution $p(y|x)$ whereas HMM is a generative model, it learns the joint probability $p(x, y)$.

- Objective function optimized.

↪ For conditional random fields, the objective is to maximize $p(y|x)$ and for HMM it is to maximize $p(x, y)$.

- Require a normalization constant.

↪ CRF has a normalizing factor since sum overall possible hypotheses need not be 1, whereas in HMM it is required.

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]

- From the given table, we can write:

$$P(H, S1) = \frac{1}{3} \times 0.5 = 0.17$$

$$P(H, S2) = \frac{1}{3} \times 0.75 = 0.25$$

$$P(H, S3) = \frac{1}{3} \times 0.25 = 0.083$$

↪ Now, moving to the next state, we need Tails in the sequence:

$$P(HT, S1) = (0.17 \times 0.9 + 0.25 \times 0.45 + 0.083 \times 0.45) \times 0.5 = 0.1514$$

$$P(HT, S2) = (0.17 \times 0.05 + 0.25 \times 0.1 + 0.083 \times 0.45) \times 0.5 = 0.0177$$

$$P(HT, S3) = (0.17 \times 0.05 + 0.25 \times 0.45 + 0.083 \times 0.1) \times 0.5 = 0.0970$$

↪ For the next throw, we need a Heads again:

$$P(HTH, S1) = (0.1514 \times 0.9 + 0.0177 \times 0.45 + 0.097 \times 0.45) \times 0.5 = 0.1514$$

$$P(HTH, S2) = (0.1514 \times 0.05 + 0.0177 \times 0.1 + 0.097 \times 0.45) \times 0.5 = 0.0177$$

$$P(HTH, S3) = (0.1514 \times 0.05 + 0.0177 \times 0.45 + 0.097 \times 0.1) \times 0.5 = 0.0970$$

↪ Finally, we can calculate the probability as:

$$P(HTH) = \sum_{S_i=1}^{3} P(HTH, S_i) = 0.1399$$

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

- Let X_t and Y_t be the hidden states of HMM. The observed variable X can be either H or T, and the state can be either 1, 2 or 3. The joint probability is given as:

$$P(X_1, X_2..X_T, Y_1, Y_2, ..Y_T) = P(Y_1)\prod_{t=2}^{T} P(Y_t|Y_{t-1})\prod_{t=1}^{T} P(X_t|Y_t)$$

- Log. likelihood of the above probability is given as:

$$L(X, Y, \theta) = \sum_i \gamma_1^i \log \pi_i + \sum_{t=2}^{T}\sum_i\sum_j \tau_{t,t-1}^{i,j} \log P(Y_t^i = 1|Y_{t-1}^j = 1) + \sum_{t=1}^{T}\sum_i\sum_k \gamma_t^i X_t^k \log P(X_t^k|Y_t^i = 1)$$

- γ_t^i as the conditional probability of Y being in state i at time t, and $\tau_{t,t-1}^{i,j}$ as the conditional probability that Y is in state i at t and j at $t-1$. The **E**-step can be derived using forward backward algorithms:

$$\gamma_t^i = \frac{P(X_1, ..X_T, Y_t^i = 1)}{P(X_1, ..X_T)} = \frac{P(X_1, ..X_t, Y_t^i = 1)P(X_{t+1}..X_T|Y_t^i = 1)}{P(X_1, ..X_T)}$$

$$\Rightarrow \gamma_t^i = \frac{\alpha_t^i \beta_t^i}{\sum_i \alpha_T^i}$$

And,

$$\tau_{t,t-1}^{i,j} = \frac{P(X_1, ..X_T, Y_t^i = 1, Y_{t-1}^j = 1)}{P(X_1, ..X_T)}$$

$$\Rightarrow \tau_{t,t-1}^{i,j} = \frac{P(X_1, ..X_{t-1}, Y_{t-1}^j = 1)P(Y_t^i = 1|Y_{t-1}^j = 1)P(X_t|Y_t^i = 1)P(X_{t+1}, ..X_T|Y_t^i = 1)}{P(X)}$$

$$\Rightarrow \tau_{t,t-1}^{i,j} = \frac{\sum_k \alpha_{t-1}^j X_t^k B_{i,k} A_{i,j} \beta_t^i}{\sum_i \alpha_T^i}$$

- For the **M**-step, we maximize the log likelihood while keeping the summation of π_i, A_{ij}, B_{ik} equal to 1. Using the Lagrange multiplier, we can write the log likelihood as:

$$L(X, Y, \theta) = \sum_i \gamma_1^i \log \pi_i + \sum_{t=1}^{T}\sum_i\sum_j \tau_{t,t-1}^{ij} \log A_i j + \sum_{t=1}^{T}\sum_i\sum_k \gamma_t^i X_t^k \log B_{ik}$$

$$+\lambda_1(1 - \sum_i \pi_i) + \sum_j \lambda_{aj}(1 - \sum_i A_{ij}) + \sum_i \lambda_{bi}(1 - \sum_k B_{ik})$$

- Now we equate the partial differential with respect to the desired variable as zero to get the desired expressions:

$$\frac{\partial L}{\partial \pi_i} = \frac{\gamma_1^i}{\pi_i} - \lambda_1 = 0$$

$$\Rightarrow \pi_i = \frac{\gamma_1^i}{\lambda_1} = 0$$

- We can find the value of λ_1 as:

$\sum_i \frac{\gamma_1^i}{\lambda_1} = 1$. Which gives us $\lambda_1 = \sum_i \gamma_1^i$

$$\Rightarrow \pi_i = \frac{\gamma_1^i}{\sum_i \gamma_1^i}$$

- For A_{ij} we get:

$$\frac{\partial L}{\partial A_{ij}} = \frac{\sum_{t=2}^{T} \tau_{t,t-1}^{ij}}{A_{ij}} - \lambda_{aj} = 0$$

$$\Rightarrow A_{ij} = \frac{\sum_{t=2}^{T} \tau_{t,t-1}^{ij}}{\lambda_{aj}}$$

- Again, we use $\sum_i A_{ij} = 1$ to find out λ_{aj}, we get:

$$A_{ij} = \frac{\sum_{t=2}^{T} \tau_{t,t-1}^{ij}}{\sum_i \sum_{t=2}^{T} \tau_{t,t-1}^{ij}}$$

- For B_{ik} we get:

$$\frac{\partial L}{\partial B_{ik}} = \frac{\sum_{t=1}^{T} \gamma_t^i X_t^k}{B_{ik}} - \lambda_{bi} = 0$$

$$\Rightarrow B_{ik} = \frac{\sum_{t=1}^{T} \gamma_t^i X_t^k}{\lambda_{bi}}$$

- Again, we use $\sum_k B_{ik} = 1$ and $\sum_k X_t^k = 1$ to find out λ_{bi}, we get:

$$B_{ik} = \frac{\sum_{t=1}^{T} \gamma_t^i X_t^k}{\sum_{t=1}^{T} \gamma_t^i}$$

(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]

In the previous question, the change in indexing of the variables is as follows:

$$X_t \rightarrow X_{p,t}$$

Where p is taken from them sequences. It is given that $P(X_t = x|S_t = i) = \mathcal{N}(\mu_i, \sigma_i^2)$, hence $P(X_{p,t}|Y_t^i = 1)$ is given by $B_i = \mathcal{N}(\mu_i, \sigma_i^2)$ where $B_i(X)$ is the probability density function of the given normal distribution at the point X.
Following the steps mentioned in the previous question, we take the Log. Likelihood and modify it as required.

$$L(X, Y, \theta) = \sum_p \sum_i \gamma_{p,1}^i \log \pi_i + \sum_p \sum_{t=2}^{T} \sum_i \sum_j \tau_{p,t,t-1}^{i,j} \log P(Y_t^i = 1|Y_{t-1}^j = 1)$$

$$+ \sum_p \sum_{t=1}^{T} \sum_i \gamma_{p,t}^i X_t^k \log P(X_{p,t}|Y_t^i = 1)$$

For textbfE-step we calculate γ and τ as follows:

$$\gamma_{p,t}^i = \frac{P(X_{p,1}, ..X_{p,T}, Y_t^i = 1)}{P(X_{p,1}, ..X_{p,T})} = \frac{P(X_{p,1}, ..X_{p,t}, Y_t^i = 1)P(X_{p,t+1}..X_{p,T}|Y_t^i = 1)}{P(X_{p,1}, ..X_{p,T})}$$

$$\Rightarrow \gamma_{p,t}^i = \frac{\alpha_{p,t}^i \beta_{p,t}^i}{\sum_i \alpha_{p,T}^i}$$

And,

$$\tau_{p,t,t-1}^{i,j} = \frac{P(X_{p,1}, ..X_{p,T}, Y_t^i = 1, Y_{t-1}^j = 1)}{P(X_{p,1}, ..X_{p,T})}$$

$$\Rightarrow \tau_{p,t,t-1}^{i,j} = \frac{P(X_{p,1}, ..X_{p,t-1}, Y_{t-1}^j = 1)P(Y_t^i = 1|Y_{t-1}^j = 1)P(X_{p,t}|Y_t^i = 1)P(X_{p,t+1}, ..X_{p,T}|Y_t^i = 1)}{P(X)}$$

$$\Rightarrow \tau_{p,t,t-1}^{i,j} = \frac{\alpha_{p,t-1}^j B_i(X_{p,t}) A_{i,j} \beta_{p,t}^i}{\sum_i \alpha_{p,T}^i}$$

For **M**-step, we again use the Lagrangian multiplier to modify the log likelihood and get:

$$L(X, Y, \theta) = \sum_p \sum_i \gamma_{p,1}^i \log \pi_i + \sum_p \sum_{t=1}^{T} \sum_i \sum_j \tau_{p,t,t-1}^{ij} \log A_i j + \sum_p \sum_{t=1}^{T} \sum_i \gamma_{p,t}^i \log B_i(X_{p,t})$$

$$+ \lambda_1 (1 - \sum_i \pi_i) + \sum_j \lambda_{aj} (1 - \sum_i A_{ij})$$

Differentiating the log likelihood with respect to the desired variables, we get:

$$\frac{\partial L}{\partial \pi_i} = \frac{\sum_p \gamma_{p,1}^i}{\pi_i} - \lambda_1 = 0$$

$$\Rightarrow \pi_i = \frac{\sum_p \gamma_{p,1}^i}{\lambda_1} = 0$$

By finding the value of λ_1 as $\sum_i \frac{\sum_p \gamma_{p,1}^i}{\lambda_1} = 1$. Which gives us $\lambda_1 = \sum_p \sum_i \gamma_{p,1}^i$

$$\Rightarrow \pi_i = \frac{\sum_p \gamma_{p,1}^i}{\sum_p \sum_i \gamma_{p,1}^i}$$

For A_{ij} we get:

$$\frac{\partial L}{\partial A_{ij}} = \frac{\sum_p \sum_{t=2}^{T} \tau_{p,t,t-1}^{ij}}{A_{ij}} - \lambda_{aj} = 0$$

$$\Rightarrow A_{ij} = \frac{\sum_p \sum_{t=2}^{T} \tau_{p,t,t-1}^{ij}}{\lambda_{aj}}$$

We $\sum_i A_{ij} = 1$ to find out λ_{aj}, we get:

$$A_{ij} = \frac{\sum_p \sum_{t=2}^{T} \tau_{pt,t-1}^{ij}}{\sum_p \sum_i \sum_{t=2}^{T} \tau_{p,t,t-1}^{ij}}$$

Since $B_i(X_{p,t})$ is a normal distribution, we can write the log as:

$$\log B_i(X_{p,t}) = -\log \sigma_i - \frac{1}{2} \log 2\pi - \frac{(X_{p,t} - \mu_i)^2}{2\sigma_i^2}$$

Partially with respect to μ_i and σ_i and equate to 0:

$$\frac{\partial L}{\partial \mu_i} = \sum_p \sum_t \gamma_{p,t}^i \frac{(X_{p,t} - \mu_i)}{\sigma_i^2} = 0$$

Which gives:

$$\mu_i = \frac{\sum_p \sum_t \gamma_{p,t}^i X_{p,t}}{\sum_p \sum_t \gamma_{p,t}^i}$$

Now, repeating the same procedure for σ_i, we get:

$$\frac{\partial L}{\partial \sigma_i} = \sum_p \sum_t -\frac{\gamma_{p,t}^i}{\sigma_i} + \frac{\gamma_{p,t}^i (X_{p,t} - \mu_i)^2}{\sigma_i^3} = 0$$

Which gives:

$$\sigma_i = \sqrt{\frac{\sum_p \sum_t \gamma_{p,t}^i (X_{p,t} - \mu_i)^2}{\sum_p \sum_t \gamma_{p,t}^i}}$$

(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.

↪ For validity, we require the summation of the transition weights to be 1, and not necessarily the same for the incoming weights.

- 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.

↪ This gives the conditional probability of going to state t given the current 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.

↪ In HMM, the next state only depends on the current state and not the entire prefix sequence.

- 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.

↪ This algorithm does not guarantee a global optimum convergence, Although it can converge to a local optimum.

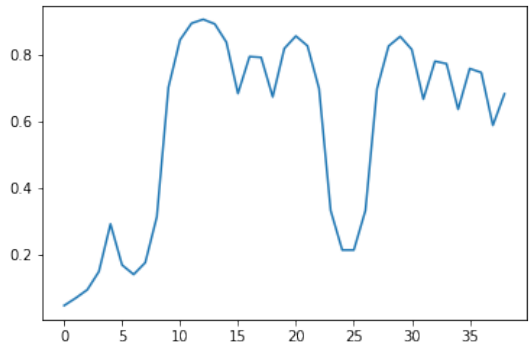
4 Programming [30 pts]

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 included in `sp500.mat`. 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 in `algorithm.m/algorithm.py` and submit this file. In your report, briefly describe how you implement your algorithm 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]

- For $q = 0.7$, we have probability that the economy is in a good state at week 39 as 0.6830. The plot is seen as in Figure 1.

Figure 1: Probability that there is 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]

- For $q = 0.9$, we have probability that the economy is in a good state at week 39 as 0.8380. The plot is seen as in Figure 1.
- Looking at figure 3, we see that at week 39, the uncertainty of the $q = 0.9$ plot is less than the $q = 0.7$ plot. This shows us that we are more confident in case of higher probability of q in our predictions.

Figure 2: Probability that there is good economic state in 39 weeks with q=0.9

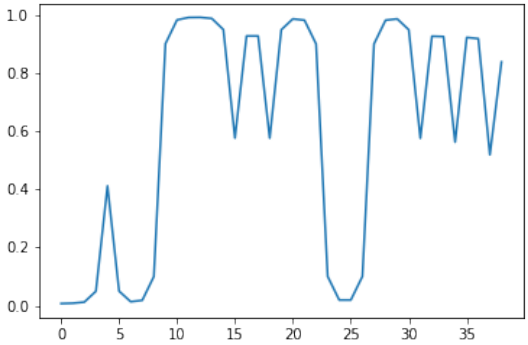


Figure 3: Comparison plot with q=0.7 and q=0.9