Homework 4

Chujie Chen
cchen641@gatech.edu

December 3, 2019

1 Kernels

(a)

1

Symmetry:

$$\begin{aligned} K(v,u) &= \alpha K_1(v,u) + \beta K_2(v,u) \\ &= \alpha K_1(u,v) + \beta K_2(u,v) \\ &= K(u,v) \end{aligned}$$

$K(u,v)$ is symmetric.

PSD:

Although K_1 and K_2 are valid kernels. For $\forall v \in \mathcal{R}^m$, we have

$$v^T K v = \alpha v^T K_1 v + \beta v^T K_2 v$$

So even we have $v^T K_1 v$ and $v^T K_2 v$, we still don't know the sign of $v^T K v$.

So the answer is **False**.

2

Symmetry:

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

So $K(u,v)$ is symmetric.

PSD:

For $\forall w \in \mathcal{R}^m$, we have

$$w^T K w = \sum_i^m \sum_j^m w_i K_1(f(u_i), f(v_j)) w_j$$

Since we know that K_1 is a valid kernel, we have

$$K_1(f(u_i), f(u_j)) = \Phi(f(u_i))^T \Phi(f(u_j))$$

for u can have indices from 1 to m and $\Phi(f(u_i)) = (\phi_1(f(u_i)), \phi_2(f(u_i)), ..., \phi_m(f(u_i)))$.

Thus,

$$\begin{aligned} w^T K w &= \sum_i^m \sum_j^m w_i \phi(f(u_i))^T \phi(f(u_j)) w_j \\ &= \sum_i^m \sum_j^m w_i w_j (\sum_k^m \phi(f(u_i))\phi(f(u_j))) \\ &= \sum_k^m (\sum_i^m w_i \phi(f(u_i)))^2 \end{aligned}$$

So K is PSD. Thus, the answer is **True**.

3

Symmetry:

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

PSD: If we look at

$$w^T K w = \sum_i \sum_j K_{u,v} w_i w_j$$

if there are u and v are no far than 1, there can be $z_i z_j$ term of which $i \neq j$. In these cases, we are not able to determine the sign. So it is not alwasy PSD. Then the answer is **False**.

4

Symmetry:

$$K(v,u) = K(v,u)$$

which is easy to prove. PSD:

$$\begin{aligned} w^T K w &= \sum_i \sum_j w_i \frac{K'(u_i,u_j)}{\sqrt{K'(u_i,u_i)K'(u_j,u_j)}} w_j \\ &= \sum_i \sum_j w_i \frac{\sum_k \phi_k(u_i)\phi_k(u_j)}{\sqrt{K'(u_i,u_i)K'(u_j,u_j)}} w_j \\ &= \sum_k (\sum_i w_i \frac{\phi_k(u_i)}{\sqrt{K'(u_i,u_i)}})^2 \end{aligned}$$

So it is PSD. The answer is **True**.

(b)

For LDA, we have

$$J(w) = \frac{w^T S_B^\phi w}{w^T S_W^\phi w}$$

where

$$S_B^\phi = (m_2^\phi - m_1^\phi)(m_2^\phi - m_1^\phi)^T$$

$$S_W^\phi = \sum_i \sum_n (\phi(x_n^i) - m_i^\phi)(\phi(x_n^i) - m_i^\phi)^T$$

and

$$m_i^\phi = \frac{1}{l_i} \sum_j \phi(x_j^i)$$

The data can be implicitly embedded by rewriting the algorithm in terms of dot products and using the kernel trick with $k(x, y) = \phi(x)\phi(y)$. Then we have

$$w = \sum_i \alpha_i \phi(x_i)$$

Then we have

$$w^T m_i^\phi = \frac{1}{l_i} \sum_k k(x_j, x_k^i) = \alpha^T M_i$$

We can rewrite the numerator and denominator of J. More details can be found here.

2 Markov Random Field, Conditional Random Field

(a)

Since

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

we have

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)

$$\begin{aligned} P(a = 1, b = 1) &= \sum_c P(a = 1, b = 1, c) \\ &= \sum_c \frac{\phi_1(a = 1, b = 1)\phi_2(b = 1, c)}{Z} \\ &= \sum_c \frac{\phi_1(a = 1, b = 1)\phi_2(b = 1, c)}{\psi(a, b = 1, c) + \psi(a, b = 0, c)} \\ &= \frac{4 + 1 + 3}{74} \\ &= \frac{4}{37} \end{aligned}$$

where $\psi(a, b = 0, c)$ can be get in a similar fashion as $\psi(a, b = 1, c)$.

(c)

1

Conditional random fields (CRF) are discriminative while HMM is generative: CRF uses feature x directly to get y while HMM has probabilities to determine the relation between x and y.

2

CRF optimizes the conditional likelihood function while the HMM uses joint likelihoods.

3

CRF needs a normalization constant as shown in (b) while HMM is using joint likelihood so we do not need the normalization term.

3 Hidden Markov Model

(a)

According to lecture 18 slide number 8, when we have the HMM and the sequences (H, T, H), clearly we need to use forward algorithm.

Initialization:
For the 1st trail, coin 1, 2, 3 separately,

$$\begin{aligned}\alpha_1 &= (\alpha_1^1, \alpha_1^2, \alpha_1^3) \\ &= P(H|coin1) * \pi(coin1) + P(H|coin2) * \pi(coin2) + P(H|coin3) * \pi(coin3) \\ &= 0.5 * \frac{1}{3} + 0.75 * \frac{1}{3} + 0.25 * \frac{1}{3} \\ &= (\frac{1}{6}, \frac{1}{4}, \frac{1}{12})\end{aligned}$$

Iteration:

$$\alpha_2^k = P(T|coink) \sum_i \alpha_1^i a_{i,k}$$

where $a_{i,k} = P(coink|coini)$ is the transition probability which can be got from A (i-th row, k-th column). Thus, for example, we can have

$$\begin{aligned}\alpha_2^1 &= P(T|coin1) \sum_i \alpha_1^i a_{i,1} \\ &= 0.5 * (\frac{1}{6} * 0.9 + \frac{1}{4} * 0.45 + \frac{1}{12} * 0.45) \\ &= \frac{3}{20}\end{aligned}$$

Similarly, we can have

$$\begin{aligned}\alpha_2^2 &= 0.25 * (\frac{1}{6} * 0.05 + \frac{1}{4} * 0.1 + \frac{1}{12} * 0.45) \\ &= \frac{17}{960} \\ \alpha_2^3 &= 0.75 * (\frac{1}{6} * 0.05 + \frac{1}{4} * 0.45 + \frac{1}{12} * 0.1) \\ &= \frac{31}{320}\end{aligned}$$

Thus, we have

$$\alpha_2 = (\frac{3}{20}, \frac{17}{960}, \frac{31}{320})$$

We can get α_3 in a similar fashion

$$\begin{aligned}\alpha_3 &= (P(H|coink) \sum_i \alpha_2^i a_{i,k})_{k=1,2,3} \\ &= (\frac{597}{6400}, \frac{203}{5120}, \frac{161}{25600})\end{aligned}$$

Termination:

$$P(HTH) = \sum_k \alpha_3^k = \frac{891}{6400} \approx 0.139$$

(b)

This is basically the **learning** question of HMM:
Given: an HMM M, with unspecified transition/emission probs., and a sequence x,
Find: parameters $\theta = (\pi_i, a_{ij}, \eta_{ik})$ that maximize $P(x|q)$
Algorithm: Baum-Welch (EM)

The complete log likelihood

$$\begin{aligned} l(\theta; x, y) &= logp(x, y) \\ &= log\Pi_n(p(y_{n,1})\Pi_{\dagger=2}^{T} p(y_{n,\dagger})|Pi_{\dagger=1}^{T} p(x_{n,\dagger}|x_{n,\dagger})) \end{aligned}$$

The expected complete log likelihood

$$\begin{aligned} < l(\theta; x, y) >= &\sum_n (< y_{n,1}^i >_{p(y_{n,1}|x_n)} log\pi_i) \\ &+ \sum_n \sum_{\dagger=2}^{T} (< y_{n,\dagger-1}^i y_{n,\dagger}^j >_{p(y_{n,\dagger-1}, y_{n,\dagger}|x_n) log a_{i,j})} \\ &+ \sum_n \sum_{\dagger=1}^{T} (x_{n,\dagger}^k < y_{n,\dagger}^i >_{p(y_{n,\dagger}|x_n} logb_{i,k}) \end{aligned}$$

First, we can calculate the E-step:

$$\begin{aligned} \gamma_{n,\dagger}^i &=< y_{n,\dagger}^i >= p(y_{n,\dagger}^i = 1|x_n) \\ \xi_{n,\dagger}^{i,j} &=< y_{n,\dagger-1}^i, y_{n,\dagger}^j >= p(y_{n,\dagger-1}^i = 1, y_{n,\dagger}^j = 1|x_n) \end{aligned}$$

Then we can calculate the M step:

$$\begin{aligned} \pi_i &= \frac{\sum_n \gamma_{n,1}^i}{N} \\ A_{ij} &= \frac{\sum_n \sum_{\dagger=2}^{T} \xi_{n,\dagger}^{i,j}}{\sum_n \sum_{\dagger=1}^{T-1} \gamma_{n,\dagger}^i} \\ B_{ik} &= \frac{\sum_n \sum_{\dagger=1}^{T} \gamma_{n,\dagger}^i x_{n,\dagger}^k}{\sum_n \sum_{\dagger=1}^{T-1} \gamma_{n,\dagger}^i} \end{aligned}$$

where $N = \sum_i \sum_n \gamma_{n,1}^i$. In above equations, trail number † is between $1 \sim T$, sequence result index is between $1 \sim N$. Note that these are the same to equations from PRML 13.2:

$$\begin{aligned} \pi_k &= \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})} \\ A_{jk} &= \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j} z_{nl})} \\ B_{jk} &= \frac{\sum_n^N \gamma(Zn, jX_{nk}}{\sum_{l=1}^{K} \sum_{n=1}^{N} \gamma(Z_{n,j} X_{n,l})} \end{aligned}$$

but with different indices and definitions. Then, we just need to repeat E and M steps until converge.

(c)

Comparing to the previous questions, $S_t = i$ is similar to the coin state i at t trail, $X_t = x$ is similar to the observation x (head/tail). So we can get the objective function for EM. The original objective function (PRML 13.12):

$$Q(\theta, \theta^{old}) = \sum_Z p(Z|X, \theta^{old}) lnp(X, Z|\theta)$$

From PRML we have, for E steps:

$$\gamma(z_{nk}) = p(z_{nk}|X_k, \theta^{old})$$
$$\xi(z_{n-1,j}, z_{nk}) = p(z_{n-1,k}, z_{nk}|X_k, \theta^{old})$$

And M steps: (note that according to the question, k can be 1,2,...m) similar to what I got before

$$\pi_i = \frac{\sum_{n=1}^{m} \gamma_{n,1}^{i}}{\sum_i \sum_n \gamma_{n,1}^{i}}$$
$$A_{ij} = \frac{\sum_{n=1}^{m} \sum_{\dagger=2}^{T} \xi_{n,\dagger}^{i,j}}{\sum_{n=1}^{m} \sum_{\dagger=1}^{T-1} \gamma_{n,\dagger}^{i}}$$

From PRML 13.20 and 13.21, we can also have

$$\mu_i = \frac{\sum_{n=1}^{m} \sum_{\dagger=1}^{T} \gamma_{n,\dagger}^{i} x_{n,\dagger}^{k}}{\sum_{n=1}^{m} \sum_{\dagger=1}^{T} \gamma_{n,\dagger}^{i}}$$
$$\sigma_i = \frac{\sum_{n=1}^{m} \sum_{\dagger=1}^{T} \gamma_{n,\dagger}^{i} (x_{n,\dagger}^{k} - \mu_i)^2}{\sum_{n=1}^{m} \sum_{\dagger=1}^{T} \gamma_{n,\dagger}^{i}}$$

Then we just repeat E and M steps until converge.

(d)

1

False. Although $\sum_i p(Z_t^i|Z_{t-1}) = 1$, we can not say $\sum_i p(Z_t|Z_{t-1}^i) = 1$.

2

False. s **to** t. It should be given state s, the conditional probability of going to t.

3

True. Yes, Markov can only use information from the last state not the very previous ones.

4

False. It can only guarantee the local optimum.

4 Programming

Basically, this is a decoding problem with HMM. We can use forward-backward algorithm to get the posterior probabilities.

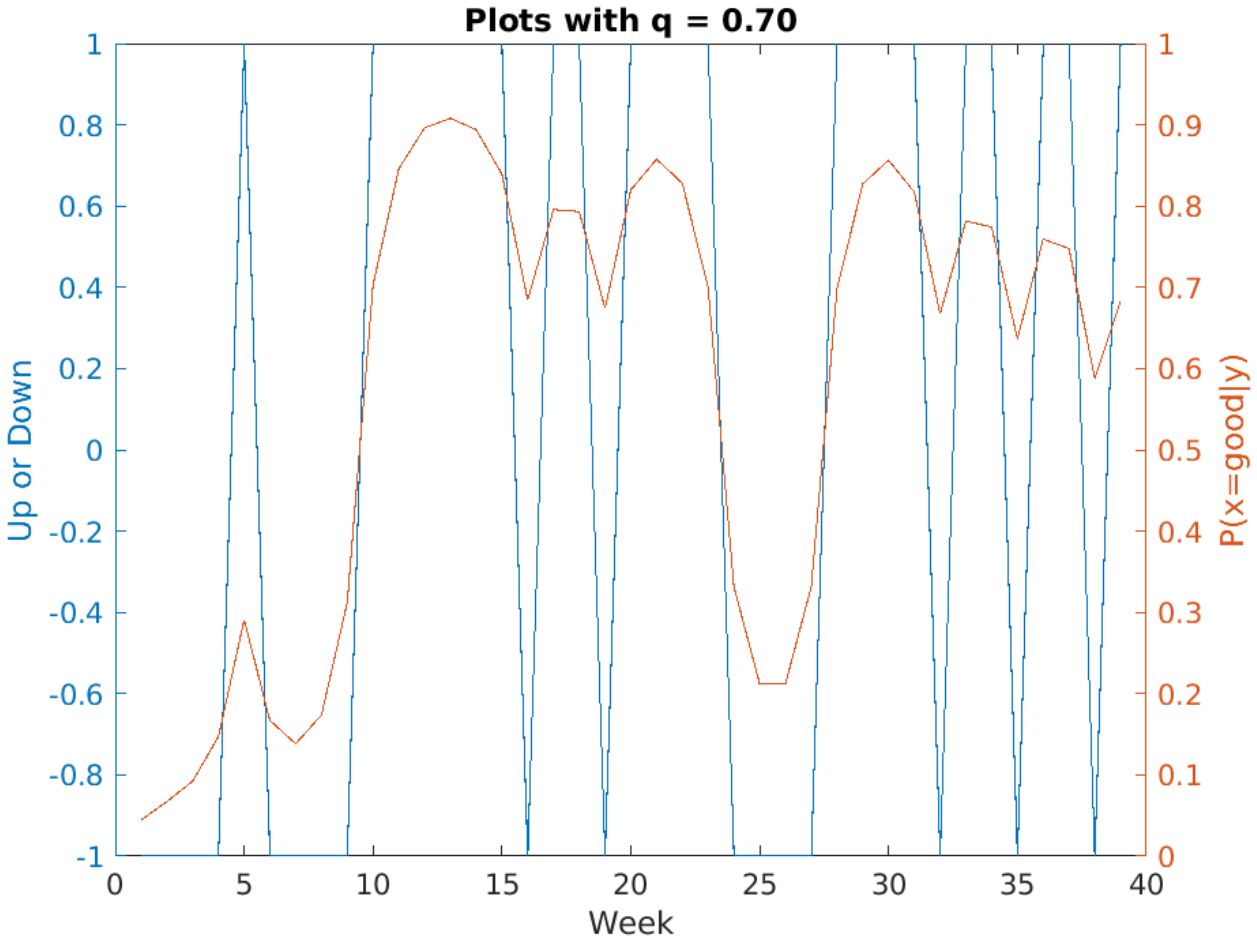
The forward process will give us α while the backward can give us β, then we can get the posterior probabilities. Note that the x and y in this problem have different definitions from those in lecture slides. We should regard x as y, y as x here.

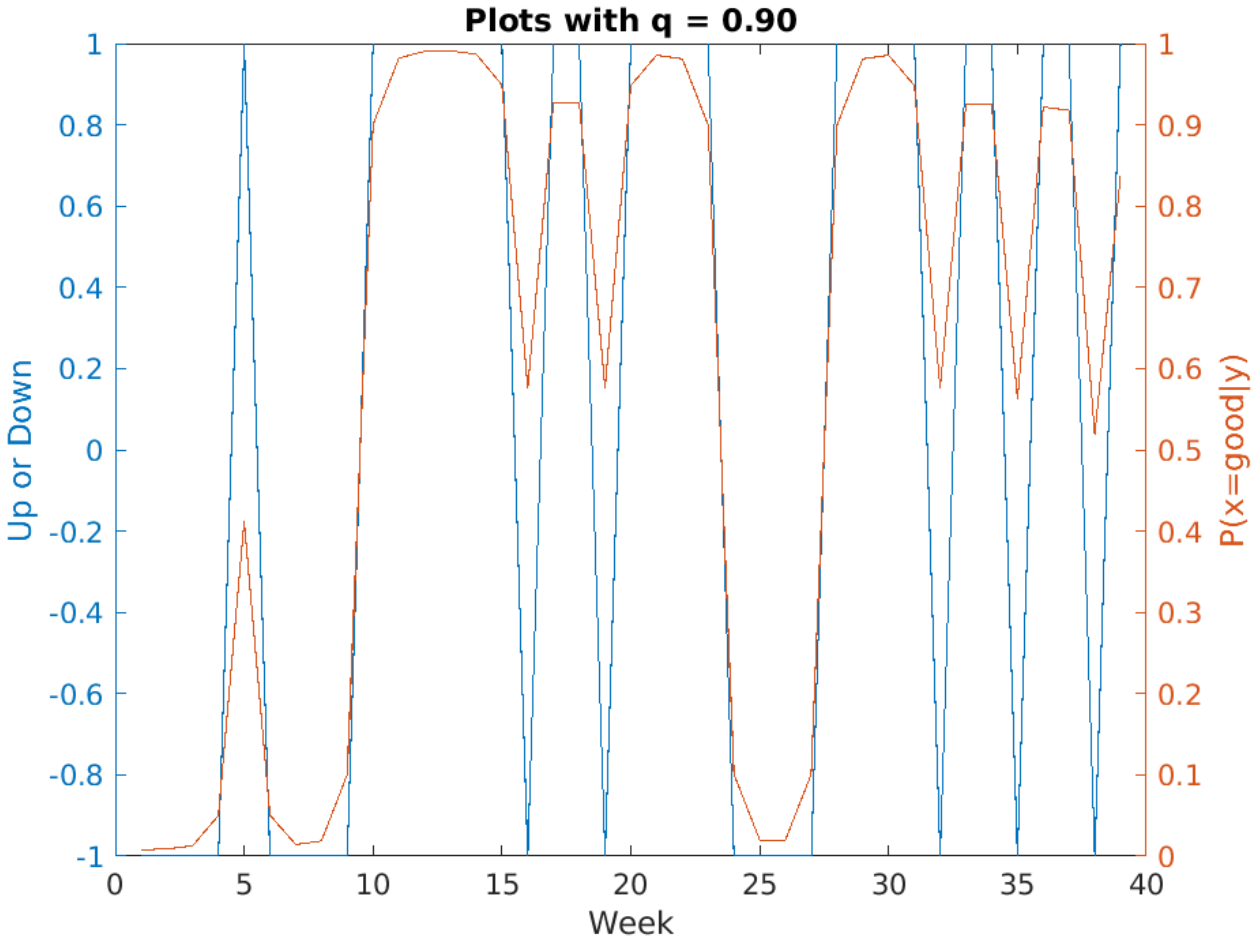
(a)

From figure below, it is 0.6830

(b)

From figure below, it is 0.8379. The higher the q is, the more correlated the label (up / down) and the hidden state (good / bad). When q is 0.9, good states tend to have a higher result and vice verse.


Plots with q = 0.70
Up or Down
P(x=good|y)
Week
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0.9
0.7
0.5
0.3
0.1
0
5
10
15
20
25
30
35
40


Plots with q = 0.90
Up or Down
P(x=good|y)
Week
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0.9
0.7
0.5
0.3
0.1
0
5
10
15
20
25
30
35
40