COMS E6998 Homework 1

Mengqi Zong $< mz2326@columbia.edu >$

February 23, 2012

Question 1

a. The decoding problem: find

$$\begin{aligned}
arg \max_{s \in S^m} p(\underline{s}|\underline{x};\underline{w}) &= arg \max_{s \in S^m} \frac{\exp\left(\underline{w}\cdot\underline{\Phi}(\underline{x},\underline{s})\right)}{\sum_{\underline{s}' \in S^m}\exp\left(\underline{w}\cdot\underline{\Phi}(\underline{x},\underline{s}')\right)} \\
&= arg \max_{s \in S^m} \exp\left(\underline{w}\cdot\underline{\Phi}(\underline{x},\underline{s}')\right) \\
&= arg \max_{s \in S^m} \underline{w}\cdot\underline{\Phi}(\underline{x},\underline{s}') \\
&= arg \max_{s \in S^m} \underline{w}\cdot\sum_{j=1}^{m}\underline{\phi}(\underline{x},j,s_{j-2},s_{j-1},s_j) \\
&= arg \max_{s \in S^m} \sum_{j=1}^{m}\underline{w}\cdot\underline{\phi}(\underline{x},j,s_{j-2},s_{j-1},s_j)
\end{aligned}$$

Here comes the dynamic programming algorithm...

- Initialization-1: for $s \in S$

$$\pi[1,s] = \underline{w}\cdot\underline{\phi}(\underline{x},1,s_{00},s_0,s)$$

  where s_{00}, s_0 is a special "initial" state.

- Initialization-2: for $s \in S$

$$\pi[2,s] = \max_{s' \in S}\left[\pi[1,s'] + \underline{w}\cdot\underline{\phi}(\underline{x},2,s_0,s',s)\right]$$

- For $j = 3...m, s = 1...k$

$$\pi[j, s] = \max_{s', s'' \in S} [\pi[j-1, s''] + \underline{w} \cdot \underline{\phi}(\underline{x}, j, s'', s', s)]$$

- We have

$$\max_{s_1 \ldots s_m} \sum_{j=1}^{m} \underline{w} \cdot \underline{\phi}(\underline{x}, j, s_{j-2}, s_{j-1}, s_j) = \max_s \pi[m, s]$$

- This algorithm runs in $O(mk^3) time$.

b. To estimate the parameters, we assume we have a set of n labeled examples, $(\underline{x}^i, \underline{s}^i)_{i=1}^n$. Each $\underline{x}^i$ is an input sequence $x_1^i...x_m^i$, each $\underline{s}^i$ is a state sequence $s_1^i...s_m^i$.

We then prooceed in exactly the same way as for regular log-linear models. The regularized log-likelihood function is

$$L(\underline{w}) = \sum_{i=1}^{n} \log p(\underline{s}^i | \underline{x}^i; \underline{w}) - \frac{\lambda}{2} ||\underline{w}||^2$$

Our parameter estimates are

$$\underline{w}^* = arg \max_{\underline{w} \in \mathbb{R}^d} \sum_{i=1}^{n} \log p(\underline{s}^i | \underline{x}^i; \underline{w}) - \frac{\lambda}{2} ||\underline{w}||^2$$

Now we use gradient-based optimization methods to find $\underline{w}^*$. Let's compute the derivatives:

$$\frac{\partial}{\partial w_k} L(\underline{w}) = \sum_i \Phi_k(\underline{x}^i, \underline{s}^i) - \sum_i \sum_{\underline{s} \in S^m} p(\underline{s} | \underline{x}^i; \underline{w}) \Phi_k \underline{x}^i, \underline{s} - \lambda w_k$$

The first term is easily computed, because

$$\sum_i \Phi_k(\underline{x}^i, \underline{s}^i) = \sum_i \sum_{j=1}^{m} \phi_k(\underline{x}^i, j, s_{j-2}^i, s_{j-1}^i, s_j^i)$$

We now consider how to compute the second term:

$$
\begin{aligned}
\sum_i \sum_{\underline{s} \in S^m} p(\underline{s}|\underline{x}^i;\underline{w}) \Phi_k(\underline{x}^i, \underline{s}) &= \sum_{\underline{s} \in S^m} p(\underline{s}|\underline{x}^i;\underline{w}) \sum_{j=1}^{m} \phi_k(\underline{x}^i, j, s_{j-2}, s_{j-1}, s_j) \\
&= \sum_{j=1}^{m} \sum_{a \in S, b \in S} q_j^i(a,b,c) \phi_k(\underline{x}^i, j, a, b, c)
\end{aligned}
$$

where

$$
q_j^i(a,b,c) = \sum_{\underline{s} \in S^m : s_{j-2}=a, s_{j-1}=b, s_j=c} p(\underline{s}|\underline{x}^i;\underline{w})
$$

For a given i, all q_j^i terms can be computed simultaneously in $O(mk^3)$ time using the forward-backward algorithm, a dynamic programming algorithm that is closely reltaed to Viterbi.

c. Here is the structured perceptron algorithm for parameter estimation:

- Input: labeled examples, $(\underline{x}^i, \underline{s}^i)_{i=1}^n$.
- Initialization: $\underline{w} = \underline{0}$.
- For $t = 1...T$, for $i = 1...n$:
  - Use the algorithm in part a to calculate

$$
\begin{aligned}
\underline{s}^* &= arg \max_{\underline{s} \in \mathrm{Y}} \ \underline{w} \cdot \underline{\Phi}(\underline{x}^i, \underline{s}) \\
&= arg \max_{\underline{s} \in \mathrm{Y}} \sum_{j=1}^{m} \underline{w} \cdot \underline{\phi}(\underline{x}, j, s_{j-2}, s_{j-1}, s_j)
\end{aligned}
$$

  - Update:

$$
\begin{aligned}
\underline{w} &= \underline{w} + \underline{\Phi}(\underline{x}^i, \underline{s}^i) - \underline{\Phi}(\underline{x}^i, \underline{s}^*) \\
&= \underline{w} + \sum_{j=1}^{m} \underline{\phi}(\underline{x}, j, s_{j-2}, s_{j-1}^i, s_j^i) - \sum_{j=1}^{m} \phi(\underline{x}, j, s_{j-2}^*, s_{j-1}^*, s_j^*)
\end{aligned}
$$

- Return $\underline{w}$.

Question 2

a. Here are all the non-zero parameters for the HMM:

$$\begin{aligned}
e(x = a|s = A) &= 1 \\
e(x = c|s = B) &= \frac{8}{9} \\
e(x = b|s = B) &= \frac{1}{9} \\
e(x = c|s = C) &= 1 \\
e(x = d|s = D) &= 1 \\
t(s' = B|s = A) &= 0.5 \\
t(s' = C|s = A) &= 0.5 \\
t(s' = D|s = B) &= 1 \\
t(A) &= 0.2 \\
t(B) &= 0.8
\end{aligned}$$

For input sequence $x_1\ x_2$, the output is :

$$P(x_1x_2, s_1s_2; \theta) = \max\ t(s_1)t(s_2|s_1)e(x_1|s_1)e(x_2|s_2)$$

Here is the output from HMM on three input sequences:

- For input sequence $a\ b$, the output is $A\ B$.
- For input sequence $a\ c$, the output is $A\ C$.
- For input sequence $c\ d$, the output is $B\ D$.

b. Here are the features for a bigram CRF for the sequence modeling problem:

$$\phi_1(\underline{x}, j, s_{j-1}, s_j) = \begin{cases} 1 & \text{if } x_j = b, s_{j-1} = A, s_j = B \\ 0 & \text{o.w.} \end{cases}$$

$$\phi_2(\underline{x}, j, s_{j-1}, s_j) = \begin{cases} 1 & \text{if } x_j = c, s_{j-1} = A, s_j = C \\ 0 & \text{o.w.} \end{cases}$$

$$\phi_3(\underline{x}, j, s_{j-1}, s_j) = \begin{cases} 1 & \text{if } j = 2, x_j = d, s_j = D \\ 0 & \text{o.w.} \end{cases}$$

$$\phi_4(\underline{x}, j, s_{j-1}, s_j) = \begin{cases} 1 & \text{if } j = 1, x_j = c, s_j = B \\ 0 & \text{o.w.} \end{cases}$$

$$\phi_5(\underline{x}, j, s_{j-1}, s_j) = \begin{cases} 1 & \text{if } j = 1, x_j = a, s_j = A \\ 0 & \text{o.w.} \end{cases}$$

Question 3

a. $v_1^* = 0$.

Since $f_1(x_i, y) = 0$ for all $i \in \{1...n\}, y \in Y$, we have

$$v_1 \cdot f_1(x_i, y) = 0$$

Then we get

$$\begin{aligned} p(y|x; v) &= \frac{\exp\left(v \cdot f(x, y)\right)}{\sum_{y \in \mathrm{Y}} \exp\left(v \cdot f(x, y)\right)} \\ &= \frac{\exp\left(0 + \sum_{j=2}^{d} v_j \cdot f_j(x, y)\right)}{\sum_{y \in \mathrm{Y}} \exp\left(0 + \sum_{j=2}^{d} v_j \cdot f_j(x, y)\right)} \\ &= \frac{\exp\left(\sum_{j=2}^{d} v_j \cdot f_j(x, y)\right)}{\sum_{y \in \mathrm{Y}} \exp\left(\sum_{j=2}^{d} v_j \cdot f_j(x, y)\right)} \end{aligned}$$

Since we only want to know the v_1^*, so we can assume that $v_2...v_d$ are fixed. Then $p(y|x; v)$ is a constant regarding of v_1. From regularized log-likelihood function

$$L(v) = \sum_{i=1}^{n} \log p(y_i | x_i; v) - \frac{\lambda}{2} \sum_j |v_j|$$

we can see that the only part we can optimize is

$$-\frac{\lambda}{2} \sum_j |v_j|$$

Obviously, when $v_1^* = 0$, the maximum value is reached.

b. $v_2^* = 0$.

Since $f_2(x_i, y) = 10$ for all $i \in \{1...n\}, y \in Y$, we have

$$v_2 \cdot f_2(x_i, y) = 10v_2$$

Then we get

$$\begin{aligned} p(y|x; v) &= \frac{\exp\left(v \cdot f(x, y)\right)}{\sum_{y \in \mathrm{Y}} \exp\left(v \cdot f(x, y)\right)} \\ &= \frac{\exp\left(10v_2 + \sum_{j \neq 2} v_j \cdot f_j(x, y)\right)}{\sum_{y \in \mathrm{Y}} \exp\left(10v_2 + \sum_{j \neq 2} v_j \cdot f_j(x, y)\right)} \end{aligned}$$

Then the regularized log-likelihood function

$$\begin{aligned} L(v) &= \sum_{i=1}^{n} \log p(y_i|x_i; v) - \frac{\lambda}{2} \sum_j |v_j| \\ &= \sum_{i=1}^{n} \log \frac{\exp\left(10v_2 + \sum_{j \neq 2} v_j \cdot f_j(x_i, y_i)\right)}{\sum_{y \in \mathrm{Y}} \exp\left(10v_2 + \sum_{j \neq 2} v_j \cdot f_j(x_i, y)\right)} - \frac{\lambda}{2} \sum_j |v_j| \\ &= \sum_{i=1}^{n} \log \frac{\exp\left(10v_2\right) \cdot \exp\left(\sum_{j \neq 2} v_j \cdot f_j(x_i, y_i)\right)}{\sum_{y \in \mathrm{Y}} \exp\left(10v_2\right) \cdot \exp\left(\sum_{j \neq 2} v_j \cdot f_j(x_i, y)\right)} - \frac{\lambda}{2} \sum_j |v_j| \\ &= \sum_{i=1}^{n} \log \frac{\exp\left(\sum_{j \neq 2} v_j \cdot f_j(x_i, y_i)\right)}{\sum_{y \in \mathrm{Y}} \exp\left(\sum_{j \neq 2} v_j \cdot f_j(x_i, y)\right)} - \frac{\lambda}{2} \sum_j |v_j| \end{aligned}$$

Since we only want to know the v_2^*, so we can assume that $v_1, v_3, ..., v_d$ are fixed. Then we can see that the only part we can optimize is

$$-\frac{\lambda}{2}|v_2|$$

Obviously, when $v_2^* = 0$, the maximum value is reached.

c. $v_3^* = 0$.

Since $f_3(x_i, y) = i$ for all $i \in \{1...n\}, y \in Y$, we have

$$v_3 \cdot f_3(x_i, y) = i \cdot v_3$$

Then we get

$$\begin{aligned}
p(y|x;v) &= \frac{\exp\left(v \cdot f(x,y)\right)}{\sum_{y \in \mathrm{Y}} \exp\left(v \cdot f(x,y)\right)} \\
&= \frac{\exp\left(i \cdot v_3 + \sum_{j \neq 3} v_j \cdot f_j(x,y)\right)}{\sum_{y \in \mathrm{Y}} \exp\left(i \cdot v_3 + \sum_{j \neq 3} v_j \cdot f_j(x,y)\right)}
\end{aligned}$$

Then the regularized log-likelhood function

$$\begin{aligned}
L(v) &= \sum_{i=1}^{n} \log p(y_i|x_i;v) - \frac{\lambda}{2} \sum_j |v_j| \\
&= \sum_{i=1}^{n} \log \frac{\exp\left(i \cdot v_3 + \sum_{j \neq 3} v_j \cdot f_j(x_i,y_i)\right)}{\sum_{y \in \mathrm{Y}} \exp\left(i \cdot v_3 + \sum_{j \neq 3} v_j \cdot f_j(x_i,y)\right)} - \frac{\lambda}{2} \sum_j |v_j| \\
&= \sum_{i=1}^{n} \log \frac{\exp\left(i \cdot v_3\right) \cdot \exp\left(\sum_{j \neq 3} v_j \cdot f_j(x_i,y_i)\right)}{\sum_{y \in \mathrm{Y}} \exp\left(i \cdot v_3\right) \cdot \exp\left(\sum_{j \neq 3} v_j \cdot f_j(x_i,y)\right)} - \frac{\lambda}{2} \sum_j |v_j| \\
&= \sum_{i=1}^{n} \log \frac{\exp\left(\sum_{j \neq 3} v_j \cdot f_j(x_i,y_i)\right)}{\sum_{y \in \mathrm{Y}} \exp\left(\sum_{j \neq 3} v_j \cdot f_j(x_i,y)\right)} - \frac{\lambda}{2} \sum_j |v_j|
\end{aligned}$$

Since we only want to know the v_3^*, so we can assume that $v_1, v_2, v_4, ..., v_d$ are fixed. Then we can see that the only part we can optimize is

$$-\frac{\lambda}{2}|v_3|$$

Obviously, when $v_3^* = 0$, the maximum value is reached.

Question 4

When $t = 1, y_i(\underline{\theta}_1 \cdot \underline{x}_i) = 0, M_1 = 0$, we have

$$\begin{aligned}\underline{\theta}_{1+1} &= (1 - \frac{1}{1})\underline{\theta}_1 + \frac{1}{\lambda} y_1 \underline{x}_1 \\ &= \frac{1}{\lambda(2-1)} \sum_{i=1}^{M_2} y'_i \underline{x}'_i\end{aligned}$$

This satisfies the condition from the question.

Suppose for $t = k - 1$, we have Suppose for $t = k - 1$, we have

$$\underline{\theta}_t = \frac{1}{\lambda(t-1)} \sum_{i=1}^{M_t} y'_i \underline{x}'_i$$

Then for $t = k$, we have two situations:

- If $y_i(\underline{\theta}_{k-1} \cdot \underline{x}_i) < 1$, then

$$\begin{aligned}\underline{\theta}_{(k-1)+1} &= (1 - \frac{1}{k-1})\, \underline{\theta}_{k-1} + \frac{1}{\lambda(k-1)} y_i \underline{x}_i \\ &= (1 - \frac{1}{k-1})(\frac{1}{\lambda((k-1)-1)} \sum_{i=1}^{k-1} y'_i \underline{x}'_i) + \frac{1}{\lambda(k-1)} y_i \underline{x}_i \\ &= \frac{k-2}{k-1} \cdot \frac{1}{\lambda(k-2)} \sum_{i=1}^{k-1} y'_i \underline{x}'_i + \frac{1}{\lambda(k-1)} y_i \underline{x}_i \\ &= \frac{1}{\lambda(k-1)} \sum_{i=1}^{k-1} y'_i \underline{x}'_i + \frac{1}{\lambda(k-1)} y_i \underline{x}_i \\ &= \frac{1}{\lambda(k-1)} \sum_{i=1}^{k} y'_i \underline{x}'_i\end{aligned}$$

- If $y_i(\underline{\theta}_{k-1} \cdot \underline{x}_i) \geq 1$, then

$$
\begin{aligned}
\underline{\theta}_{(k-1)+1} &= (1 - \frac{1}{k-1})\,\underline{\theta}_{k-1} \\
&= (1 - \frac{1}{k-1})(\frac{1}{\lambda((k-1)-1)}\sum_{i=1}^{k-1} y'_i \underline{x}'_i) \\
&= \frac{k-2}{k-1} \cdot \frac{1}{\lambda(k-2)}\sum_{i=1}^{k-1} y'_i \underline{x}'_i \\
&= \frac{1}{\lambda(k-1)}\sum_{i=1}^{k-1} y'_i \underline{x}'_i
\end{aligned}
$$

As a result, for any $t \geq 2$ we have

$$
\underline{\theta}_t = \frac{1}{\lambda(t-1)}\sum_{i=1}^{M_t} y'_i \underline{x}'_i
$$

Question 5

Proof: First, define $\underline{\omega}^k$ to be the parameter vector when the algorithm amkes its kth error. Note that we have

$$
\underline{\omega}^1 = \underline{0}
$$

Next, assuming the kth error is made on example t, we have

$$
\begin{aligned}
\underline{\omega}^{k+1} \cdot \underline{\omega}^* &= (\underline{\omega}^k + (\underline{\Phi}(\underline{x}^i, \underline{s}^i) - \underline{\Phi}(\underline{x}^i, \underline{s}^*))) \cdot \underline{\omega}^* \\
&= \underline{\omega}^k \cdot \underline{\omega}^* + \underline{\Phi}(\underline{x}^i, \underline{s}^i) \cdot \underline{\omega}^* - \underline{\Phi}(\underline{x}^i, \underline{s}^*) \cdot \underline{\omega}^* \\
&\geq \underline{\omega}^k \cdot \underline{\omega}^* + \delta
\end{aligned}
$$

It follows by induction on k (recall that $||\underline{\omega}^1|| = 0$), that

$$
\underline{\omega}^{k+1} \cdot \underline{\omega}^* \geq k\delta
$$

In addition, because $||\underline{\omega}^{k+1}|| \times ||\underline{\omega}^*|| \geq \underline{\omega}^{k+1} \cdot \underline{\omega}^*$, and $||\underline{\omega}^*|| = 1$, we have

$$
||\underline{\omega}^{k+1}|| \geq k\delta
$$

In the second part of the proof, we will derive an upper bound on $||\underline{\omega}^{k+1}||$. We have

$$
\begin{aligned}
||\underline{\omega}^{k+1}||^2 &= ||\underline{\omega}^k + \underline{\Phi}(\underline{x}^i, \underline{s}^i) - \underline{\Phi}(\underline{x}^i, \underline{s}^*)||^2 \\
&= ||\underline{\omega}^k||^2 + ||\underline{\Phi}(\underline{x}^i, \underline{s}^i) - \underline{\Phi}(\underline{x}^i, \underline{s}^*)||^2 + 2\underline{\omega}^k \cdot \underline{\Phi}(\underline{x}^i, \underline{s}^i) - 2\underline{\omega}^k \cdot \underline{\Phi}(\underline{x}^i, \underline{s}^*) \\
&\leq ||\underline{\omega}^k||^2 + R^2
\end{aligned}
$$

It follows by induction on k (recall that $||\underline{\omega}^1||^2$), that

$$||\underline{\omega}^{k+1}||^2 \leq kR^2$$

Combining the two bounds gives

$$k^2\delta^2 \leq ||\underline{\omega}^{k+1}||^2 \leq kR^2$$

from which it follows that

$$k \leq \frac{R^2}{\delta^2}$$