1 Q: List probabilities HMM, MEMM, and CRF need to estimate during training. (how these probabilities sum to constant 1). (Point out sample space of each probability distribution).

Ans:

1 MEMM:

- to compute maximum entropy we need to define a window first. we can compute $q_t = c$ using last state q_{t-1} or last two states q_{t-1}, q_{t-2} and so on. I assume we are using bigram model that predict current q_t based on previous state q_{t-1}. So we have:

$$pr(q_t = c \mid q_{t-1}, O_t ; \theta^c) = \frac{1}{Z(q_{t-1}, O_t)} \exp\left\{ \sum_{i=1}^{d} \theta_i^c f_i(q_{t-1}, O_t, q_t = c) \right\}$$

and we have optimize log-likelihood to find θ^*.

- So with bigram MEMM or window we have to compute $\sum_{c=\{\text{all tags}\}} pr(q_t = c \mid q_{t-1}, O_t ; \theta^c) = 1$ T times (T = number of sequences)

Amin Hosseiny Marani

(1) cont'd

CRF:

for CRF we have $\frac{1}{Z(o)} \alpha_t(q_1)\, \beta_t(q_2) \times \exp(S(q_1, q_2, o))$

for three sequences as an example. so we need

to compute TN probabilities for α, same number

for β, and $(T-1)$ possible triangle cliques for each

o and since we have T sequences, it's $(T-1) \times T$.

total number would be $T(T-1) \times 2TN \simeq T^2 + 2TN$

probabilities.

Amin Hosseiny Marani

1 Con'td

HMM parameters:

We can use Expectation maximization to learn A, B, and π.

in E step:

each $T \times N$ computation
$$\begin{cases} \alpha_t(j) & \text{for all tags } (j\text{s}) \text{ and all sequences } t. \\ \beta_t(j) & \text{for all tags } (j\text{s}) \text{ and all sequences } t. \end{cases}$$

M-step:

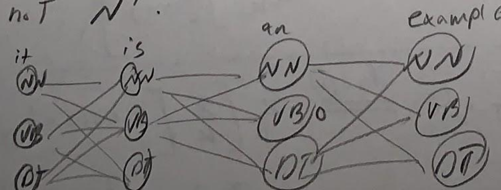
$$\overbrace{Pr(q_t = i, q_{t+1} = j \mid O, \lambda)}^{T^2 \text{ computation}} \longrightarrow \text{sum to } 1$$

$$\underbrace{Pr(q_t = i \mid O, \lambda)}_{T \text{ computation}} \longrightarrow \text{sum to } 1$$

then $T^2 + T + 2TN$ probabilities should be computed.

---

* if we look at a lattice of HMM with 3 tags like below we see for $T = 4$ we have to compute 3^4 probabilities, but since lots of these are zero, and we avoid those paths, the real number is not N^T.



Amin Hosseiny Marani

(2) I can think of <u>starting with capital letter for</u> a feature.

So <u>f_1: c = NNP And $o_t = X$ And starts with capital letter.</u>

In this case if term X did not appear in our traing set, the model cannot generalize and correctly classify it, because of $o_t = x$ in f_1. removing that, however makes the accuracy falls below 100%, due to tagging any word starting with a capital letter.

* maybe similar features such as capital letter can be added the f_1.

Amin Hosseiny Marani

(3)

$$\theta_{f_i} = \theta_i + \overset{\text{learning rate}}{\eta} \frac{\partial}{\partial \theta_c} \left[-\log Pr(q_t = c \mid o_t ; \theta) \right]$$

$$= \theta_i - \eta \left\{ - \left\{ [q_t = c] - Pr(q_t = c \mid o_t ; \theta) \right\} f_i(o_t, q_t = c) \right.$$

$$\underbrace{= \theta_i - \eta \left\{ - \left\{ \cancel{[q_t = N]}^{\,0} - Pr(q_t = N \mid o_t = race ; \theta) \right\} f_i(o_t = \text{'race'}, q_t = N) \right.}_{(1)}$$

(2)

$$Pr(q_t = N \mid o_t = race ; \theta) = \frac{1}{Z} \exp \left\{ \sum_{i=1}^{6} \theta_i^c f_i(o_t = race, q_t = N) \right\}$$

$$Z = \sum_{c' \in \{\text{all tags}\}} \exp \left\{ \sum_{i=1}^{6} \theta_i^{c'} f_i(o_t = race, q_t = c') \right\}$$

$$= \underbrace{\left[\exp\{0.8 - 1.3\} \right]}_{c' = NN} + \underbrace{\left[\exp\{0.8 + 0.01 + 0.1\} \right]}_{c' = VB} + \underbrace{\left[\exp\{0\} \right]}_{c' = VBG} = 0.607 + 2.484 + 1.0 = 4.091 \quad (3)$$

from (2) and (3): $Pr(q_t = N \mid o_t = race ; \theta) = \frac{1}{4.091} \times \exp(-0.5)$

$$= \frac{0.607}{4.091} = 0.148 \quad (4)$$

Amin Hosseiny Marani

③ Cont'd

from ④ and ①:

$$\theta_i = \theta_i - \eta(-[-0.148])f_i(o_t = race, y_t = N)$$

$$\theta_1 = 0.8 - \eta(0.148)1 = 0.8 - 0.148\eta$$

$$\theta_2 = 1 - \eta(0.148) \times 0 = 1$$

$$\theta_3 = -2 - \eta(\eta) \times 0 = -2$$

$$\theta_4 = 3 - \eta(\eta) \times 0 = 3$$

$$\theta_5 = 0.1 - \eta(\eta) \times 0 = 0.1$$

$$\theta_6 = -1.3 - \eta(0.148) \times 1 = -1.3 - \eta(0.148)$$

* the parameters/weights θ_i for those features (f_1 and f_6) that were wrong are getting smaller due to the wrong classification and tagging NN instead of VB.

on the other hand, f_2, f_4, and f_5 for $C = VB$ should increase for θ_{VB}.

(2)

$a^t = b + W h^{t-1} + U x^t$

$h^t = \tanh(a^t)$

$o^t = c + V h^t$

$\hat{y}^t = \mathrm{softmax}(o^t)$

o^t: len of 2

h^t: len of 3

x^t: len of 2

b, W, U, c, V size?

- h^t is of length 3, so a^t should be as of length 3 as well

So, $b \longrightarrow 3\times1$

$W \longrightarrow 3\times3$

$U \longrightarrow 3\times2$

- since h^t is of length 3 and o^t has length of 2, then

$c \longrightarrow 2\times1$ and $V \longrightarrow 2\times3$

(5)

$b = [1, 1, 1]$

$W = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ $U = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$

$c = [1, 1]$ $V = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

$h^0 = [1 \; 1 \; 1]^T$

$x^1 = [1, 1]^T$, $x^2 = [2, 2]^T$

$t = 1$

x^1 : $a^1 = [1, 1, 1] + [3 \; 3 \; 3] + [2 \; 2 \; 2]$
$= [6 \; 6 \; 6]$

$h^1 \simeq [1 \; 1 \; 1]^T = \tanh(a^1)$

$o^1 = [1 \; 1] + [3 \; 3] = [4 \; 4]$

$\hat{y}^1 = \text{Softmax}([4 \; 4]) = [0.5 \; 0.5]$

* Since $h^1 = h^0$, all the results would be the same for $t = 2$.

$a^1 = a^2$
$o^1 = o^2$
$h^1 = h^2$
$\hat{y}^1 = \hat{y}^2$

* Since $\tanh(6) \simeq 3.9999\ldots$
values are round.

5 con'td

$t=1$

x^2:

$a^1 = [8 \quad 8 \quad 8]$

$h^1 = \tanh([8 \quad 8 \quad 8]) \simeq [1 \quad 1 \quad 1]$

$o^1 = [4 \quad 4]$

$\hat{y}^1 = [0.5 \quad 0.5]$

again Since $h^1 = h^2 \longrightarrow a^1 = a^2 \,/\, o^1 = o^2 \,/\, \hat{y}^1 = \hat{y}^2$