CS688: Graphical Models - Spring 2016

Assignment 2: Part A

Shamya Karumbaiah

Collaborated with Rafael Lizarralde

1) *Code file - Q1Inference.m*

1.1 The node potential for each position in the first test word is given in the table below -

Position 1	Position 2	Position 3	Position 4
-7.6444	-4.0745	-10.2081	6.4649
18.4684	5.7448	0.8973	24.5313
-6.3286	1.1764	17.1910	-13.3429
10.4225	-1.7931	-12.0177	5.8712
-4.9672	-1.2122	5.5794	-10.9548
-1.9340	-1.7849	-0.5940	-11.4965
-0.9452	-8.2999	-21.4264	-5.4946
-5.6571	3.0952	9.1489	-7.1956
5.3953	6.8066	9.4824	8.0457
-6.8098	0.3416	1.9472	3.5715

1.2 The value of the negative energy of the true label sequence after conditioning on the corresponding observed image sequence -

Word	Negative Energy
that	63.9793
hire	89.6109
rises	96.9406

1.3 The log partition function for the CRF model after conditioning on the corresponding observed image sequence -

Word	Log partition
that	67.6019
hire	89.6144
rises	103.5276

1.4 The most likely joint labeling and its probability under the model -

Test Word	Most likely label	Probability
that	trat	.7958188
hire	hire	.9965205
rises	riser	.9370071

1.5 The marginal probability distribution over character labels for each position in the word -

Label	Position1	Position2	Position3	Position4
e	7.2227e-012	12.6584e-006	1.1321e-012	8.8683e-009
t	999.5246e-003	172.4732e-003	22.9451e-009	999.9999e-003
a	26.2617e-012	2.7314e-003	999.4589e-003	21.3566e-018
i	472.7213e-006	175.2834e-006	161.1855e-015	7.4054e-009
n	71.5555e-012	200.7356e-006	3.6976e-006	329.0041e-018
o	2.1138e-009	140.0475e-006	17.6109e-009	144.1003e-018
s	3.2960e-009	106.4607e-009	5.1721e-018	53.7109e-015
h	43.4927e-012	26.7353e-003	283.5253e-006	13.1781e-015
r	2.6281e-006	796.5951e-003	253.7649e-006	63.9398e-009
d	10.6937e-012	936.2852e-006	94.6377e-009	637.3628e-012

2) Code file - Q2.1_4SumProdMsgPass

2.1 Clique potential for the labels t,a,h for each of the three clique potentials

For clique C1:

/	T	A	H
T	17.8146	18.7491	18.8339
A	-6.0479	-6.5593	-6.2812
H	-5.2916	-5.6098	-5.7933

For clique C2:

/	T	A	H
T	5.0911	6.0255	6.1103
A	1.4571	.9456	1.2237
H	3.4607	3.1425	2.9590

For clique C3:

/	T	A	H
T	24.7749	-12.1649	-5.9328
A	42.0030	3.6174	10.0427
H	34.0456	-4.1467	1.8171

2.2 Log-space sum-product messages -

Character	$\delta'_{3\to2}(Y_3)$	$\delta'_{2\to1}(Y_2)$	$\delta'_{1\to2}(Y_2)$	$\delta'_{2\to3}(Y_3)$
e	14.4439	37.7353	18.5893	25.6511
t	24.7749	48.0291	17.8153	25.2369
a	42.0030	42.9495	18.7494	25.5984
i	12.5677	40.4300	18.5227	25.5779
n	29.8224	40.9076	18.1808	25.2716
o	24.1459	40.0510	18.6773	25.6012
s	2.7272	33.4551	18.0913	25.0715
h	34.0456	45.1460	18.8341	25.3880
r	33.9083	49.0110	18.3634	25.4145
d	26.2260	42.4119	18.2164	25.2026

2.3 Log belief 1 -

/	T	A
T	65.8437	61.6986
A	41.9812	36.3903

Log belief 2 -

/	T	A
T	47.6812	65.8438
A	44.9813	61.6980

Log belief 3 -

/	T	A
T	50.0117	13.0720
A	67.6013	29.2158

2.4 Marginal Distribution -

Label	Position1	Position2	Position3	Position4
e	7.2227e-012	12.6584e-006	1.1321e-012	8.8683e-009
t	999.5246e-003	172.4732e-003	22.9451e-009	999.9999e-003
a	26.2617e-012	2.7314e-003	999.4589e-003	21.3566e-018
i	472.7213e-006	175.2834e-006	161.1855e-015	7.4054e-009
n	71.5555e-012	200.7356e-006	3.6976e-006	329.0041e-018
o	2.1138e-009	140.0475e-006	17.6109e-009	144.1003e-018
s	3.2960e-009	106.4607e-009	5.1721e-018	53.7109e-015
h	43.4927e-012	26.7353e-003	283.5253e-006	13.1781e-015
r	2.6281e-006	796.5951e-003	253.7649e-006	63.9398e-009
d	10.6937e-012	936.2852e-006	94.6377e-009	637.3628e-012

Pairwise Marginals -
Belief 1:

/	T	A	H
T	172.3604e-003	2.7305e-003	26.7298e-003
A	7.4658e-012	27.8595e-015	330.8640e-015
H	15.9043e-012	72.0009e-015	538.9681e-015

Belief 2:

/	T	A	H
T	2.2314e-009	172.3712e-003	65.6861e-006
A	149.9698e-012	2.7288e-003	1.2616e-006
H	1.2104e-009	26.7203e-003	7.7862e-006

Belief 3:

/	T	A	H
T	22.9451e-009	2.0796e-024	1.0581e-021
A	999.4588e-003	21.3368e-018	13.1708e-015
H	283.5253e-006	7.3432e-021	2.8571e-018

2.5 Code file - Q2.5TestAccuracy
Predictions for the first 5 test words -

Test Word	Predicted Word
that	trat
hire	hire
rises	riser
edison	edison
shore	shore

The average character-level accuracy over the complete test set is 972/1081 which is 89.917% accuracy.

3) 3.1 The average log likelihood function can be written as -

$$l(\theta) = \frac{1}{N}\sum_{i=1}^{N} logP(\mathbf{y}^{(i)}|\mathbf{x}^{(i)};\theta)$$

Substituting our conditional model $P_w(y_i|x_i)$, we have -

$$l(w) = \frac{1}{N}\sum_{i=1}^{N} logP_w(\mathbf{y}^{(i)}|\mathbf{x}^{(i)})$$

This could be written as -

$$l(w) = \frac{1}{N}\sum_{i=1}^{N} log(\frac{exp(-E_w(\mathbf{x}_i,\mathbf{y}_i))}{Z})$$

where,

$$-E_w(\mathbf{x}_i,\mathbf{y}_i) = \sum_{j=1}^{L_i}\sum_{c=1}^{C}\sum_{f=1}^{F} W_{cf}^{F}[y_{ij}=c]x_{ijf} + \sum_{j=1}^{L_i-1}\sum_{c=1}^{C}\sum_{c'=1}^{C} W_{cc'}^{T}[y_{ij}=c][y_{ij+1}=c']$$

and

$$Z = \sum_{\mathbf{y}'_i} exp(-E_w(\mathbf{x'}_i,\mathbf{y'}_i))$$

Simplifying this, we get -

$$l(w) = \frac{1}{N}\sum_{i=1}^{N}(-E_w(\mathbf{x}_i, \mathbf{y}_i) - logZ)$$

3.2 Differentiating the average log likelihood function with respect to the feature parameter W_{cf}^F -

$$\frac{\partial l}{\partial W_{cf}^F} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial(-E_w(\mathbf{x}_i, \mathbf{y}_i))}{\partial W_{cf}^F} - \frac{\partial logZ}{\partial W_{cf}^F}$$

Solving the first part,

$$\frac{\partial(-E_w(\mathbf{x}_i, \mathbf{y}_i))}{\partial W_{cf}^F} = \frac{\partial}{\partial W_{cf}^F}(\sum_{j=1}^{L_i}\sum_{c=1}^{C}\sum_{f=1}^{F} W_{cf}^F[y_{ij} = c]x_{ijf} + \sum_{j=1}^{L_i-1}\sum_{c=1}^{C}\sum_{c'=1}^{C} W_{cc'}^T[y_{ij} = c][y_{ij+1} = c'])$$

$$= \sum_{j=1}^{L_i}\sum_{c=1}^{C}\sum_{f=1}^{F}[y_{ij} = c]x_{ijf}$$

This is the nothing but the expectation of the feature count in the actual data.

Solving the second part,

$$\frac{\partial logZ}{\partial W_{cf}^F} = \frac{1}{Z}\frac{\partial Z}{\partial W_{cf}^F} = \frac{1}{Z}\frac{\partial}{\partial W_{cf}^F}\sum_{\mathbf{y}'} exp(-E_w(\mathbf{x}_i, \mathbf{y}'))$$

$$= \frac{1}{Z}\sum_{\mathbf{y}'} exp(-E_w(\mathbf{x}_i, \mathbf{y}'))\frac{\partial(-E_w(\mathbf{x}_i, \mathbf{y}'))}{\partial W_{cf}^F}$$

$$= \sum_{\mathbf{y}'}\frac{exp(-E_w(\mathbf{x}_i, \mathbf{y}'))}{Z}\frac{\partial(-E_w(\mathbf{x}_i, \mathbf{y}'))}{\partial W_{cf}^F}$$

$$= \sum_{\mathbf{y}'} P(\mathbf{y}'|\mathbf{x}^{(i)})\sum_{j=1}^{L_i}\sum_{c=1}^{C}\sum_{f=1}^{F}[y'_j = c]x_{ijf}$$

This is nothing but the expectation of the feature count from the CRF model.

In the final form the derivation of the average log likelihood function with respect to the feature parameter W^F_{cf} is -

$$\frac{\partial l}{\partial W^F_{cf}} = \frac{1}{N}\sum_{i=1}^{N}(\sum_{j=1}^{L_i}\sum_{c=1}^{C}\sum_{f=1}^{F}[y_{ij}=c]x_{ijf} - \sum_{\mathbf{y}'}P(\mathbf{y}'|\mathbf{x}^{(i)})\sum_{j=1}^{L_i}\sum_{c=1}^{C}\sum_{f=1}^{F}[y'_j=c]x_{ijf})$$

Since we are differentiating wrto W^F_{cf}, we could drop the sum over C and F -

$$\frac{\partial l}{\partial W^F_{cf}} = \frac{1}{N}\sum_{i=1}^{N}(\sum_{j=1}^{L_i}[y_{ij}=c]x_{ijf} - \sum_{\mathbf{y}'}P(\mathbf{y}'|\mathbf{x}^{(i)})\sum_{j=1}^{L_i}[y'_j=c]x_{ijf})$$

3.3 Differentiating the average log likelihood function with respect to the transition parameter $W^T_{cc'}$ -

$$\frac{\partial l}{\partial W^T_{cc'}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial(-E_w(\mathbf{x}_i,\mathbf{y}_i))}{\partial W^T_{cc'}} - \frac{\partial logZ}{\partial W^T_{cc'}}$$

Solving the first part,

$$\frac{\partial(-E_w(\mathbf{x}_i,\mathbf{y}_i))}{\partial W^T_{cc'}} = \frac{\partial}{\partial W^T_{cc'}}(\sum_{j=1}^{L_i}\sum_{c=1}^{C}\sum_{f=1}^{F}W^F_{cf}[y_{ij}=c]x_{ijf} + \sum_{j=1}^{L_i-1}\sum_{c=1}^{C}\sum_{c'=1}^{C}W^T_{cc'}[y_{ij}=c][y_{ij+1}=c'])$$

$$= \sum_{j=1}^{L_i-1}\sum_{c=1}^{C}\sum_{c'=1}^{C}[y_{ij}=c][y_{ij+1}=c']$$

This is the nothing but the expectation of the transition in the actual data.

Solving the second part,

$$\frac{\partial logZ}{\partial W^T_{cc'}} = \frac{1}{Z}\frac{\partial Z}{\partial W^T_{cc'}} = \frac{1}{Z}\frac{\partial}{\partial W^T_{cc'}}\sum_{\mathbf{y}'_i}exp(-E_w(\mathbf{x}_i,\mathbf{y}'))$$

$$= \frac{1}{Z}\sum_{\mathbf{y}'}exp(-E_w(\mathbf{x}_i,\mathbf{y}'))\frac{\partial(-E_w(\mathbf{x}_i,\mathbf{y}'))}{\partial W^T_{cc'}}$$

$$= \sum_{\mathbf{y}'}\frac{exp(-E_w(\mathbf{x}_i,\mathbf{y}'))}{Z}\frac{\partial(-E_w(\mathbf{x}_i,\mathbf{y}'))}{\partial W^T_{cc'}}$$

$$= \sum_{\mathbf{y}'} P(\mathbf{y}'|\mathbf{x}^{(i)}) \sum_{j=1}^{L_i-1} \sum_{c=1}^{C} \sum_{c'=1}^{C} [y'_j = c][y'_{j+1} = c']$$

This is nothing but the expectation of the transition from the CRF model.

In the final form the derivation of the average log likelihood function with respect to the feature parameter $W^T_{cc'}$ is -

$$\frac{\partial l}{\partial W^T_{cc'}} = \frac{1}{N} \sum_{i=1}^{N} (\sum_{j=1}^{L_i-1} \sum_{c=1}^{C} \sum_{c'=1}^{C} [y_{ij} = c][y_{ij+1} = c'] - \sum_{\mathbf{y}'} P(\mathbf{y}'|\mathbf{x}^{(i)}) \sum_{j=1}^{L_i-1} \sum_{c=1}^{C} \sum_{c'=1}^{C} [y'_j = c][y'_{j+1} = c'])$$

Since we are differentiating wrto $W^T_{cc'}$, we could drop the sum over C -

$$\frac{\partial l}{\partial W^T_{cc'}} = \frac{1}{N} \sum_{i=1}^{N} (\sum_{j=1}^{L_i-1} [y_{ij} = c][y_{ij+1} = c'] - \sum_{\mathbf{y}'} P(\mathbf{y}'|\mathbf{x}^{(i)}) \sum_{j=1}^{L_i-1} [y'_j = c][y'_{j+1} = c'])$$

3.4 As a byproduct of the sum-product algorithm's computation we get the single-variable and pairwise marginal probabilities.
$\frac{\partial l}{\partial W^F_{cf}}$ needs single variable marginal probability and $\frac{\partial l}{\partial W^T_{cc'}}$ needs the pairwise marginal probabilities. Hence, we can efficiently compute both the value of the log-likelihood function and the values of the derivatives by directly substituting these values obtained from the sum-product algorithm's computation.

3.5 Code file - Q3AvgLogLikelihood
Using a data set consisting of the first 50 training data cases only, the average log likelihood of the true label sequences given the image sequences using the supplied model parameters is -4.5840.

4) Code file - Q4maximize
4.1 Consider the objective function, $f_w(x, y) = -(1 - x)^2 - 100(y - x^2)^2$
Differentiating the equation with respect to x, we have

$$\begin{aligned} \frac{\partial f_w(x, y)}{\partial x} &= (-2(1 - x)(-1)) - (100 * 2(y - x^2)(-2x)) \\ &= 2(1 - x) - 400x(y - x^2) \\ &= 2 - 2x + 400xy - 400x^3 \end{aligned}$$

Differentiating the equation with respect to y, we have

$$\begin{aligned} \frac{\partial f_w(x, y)}{\partial y} &= -100 * 2(y - x^2) \\ &= 200x^2 - 200y \end{aligned}$$

4.2 The location of the maximum is x =1 and y = 1 and the value of the objective function at the maximum is 0.