CSE – 426 Homework 13

Griffin Kent

1) With the following log-likelihood of CRF (Conditional Random Field)

$$\ell(X_1, \dots, X_n, Z_1, \dots, Z_n; \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{D} \theta_j f_j(Z_i, Z_{i-1}, X_i) - \log Z,$$

Where

$$Z = \sum_{Z_1,\dots,Z_n} \prod_{i=1}^{n} \psi_i(Z_i, Z_{i-1}, X_i),$$

Prove that the partial derivative of ℓ with respect to the parameter θ_j is:

$$\frac{\partial \ell}{\partial \theta_j} = \sum_{i=1}^{n} f_j(Z_i, Z_{i-1}, X_i) - \frac{1}{Z} \sum_{Z_1,\dots,Z_n} \sum_{i=1}^{n} \prod_{i=1}^{n} \psi_i(Z_i, Z_{i-1}, X_i) f_j(Z_i, Z_{i-1}, X_i).$$

(Proof): We can calculate the partial of ℓ with respect to θ_j with the following steps:

$$\frac{\partial \ell}{\partial \theta_j} = \frac{\partial}{\partial \theta_j}\left[\sum_{i=1}^{n} \sum_{j=1}^{D} \theta_j f_j(Z_i, Z_{i-1}, X_i) - \log Z\right]$$

$$= \frac{\partial}{\partial \theta_j}\left(\sum_{i=1}^{n} \sum_{j=1}^{D} \theta_j f_j(Z_i, Z_{i-1}, X_i)\right) - \frac{\partial}{\partial \theta_j}(\log Z) = \sum_{i=1}^{n} f_j(Z_i, Z_{i-1}, X_i) - \frac{1}{Z}\frac{\partial}{\partial \theta_j}(Z)$$

$$= \sum_{i=1}^{n} f_j(Z_i, Z_{i-1}, X_i) - \frac{1}{Z}\frac{\partial}{\partial \theta_j}\left(\sum_{Z_1,\dots,Z_n} \prod_{i=1}^{n} \psi_i(Z_i, Z_{i-1}, X_i)\right)$$

$$= \sum_{i=1}^{n} f_j(Z_i, Z_{i-1}, X_i) - \frac{1}{Z} \sum_{Z_1,\dots,Z_n} \frac{\partial}{\partial \theta_j} \prod_{i=1}^{n} \psi_i(Z_i, Z_{i-1}, X_i).$$

Now, we can use the property of the potential function that

$$\psi_i(Z_i, Z_{i-1}, X_i; \boldsymbol{\theta}) = \exp\left\{\sum_{j=1}^{D} \theta_j f_j(Z_i, Z_{i-1}, X_i)\right\}.$$

For simplicity and readability, we will denote $\phi_i \triangleq \sum_{j=1}^{D} \theta_j f_j(Z_i, Z_{i-1}, X_i)$. Then we can take the derivative of the following expression:

$$\frac{\partial}{\partial \theta_j} \prod_{i=1}^{n} \psi_i(Z_i, Z_{i-1}, X_i) = \frac{\partial}{\partial \theta_j} \prod_{i=1}^{n} \exp\left\{\sum_{j=1}^{D} \theta_j f_j(Z_i, Z_{i-1}, X_i)\right\} = \frac{\partial}{\partial \theta_j} \prod_{i=1}^{n} e^{\phi_i}.$$

Using the product rule, we have

$$\frac{\partial}{\partial \theta_j} \prod_{i=1}^{n} e^{\phi_i} = \frac{\partial}{\partial \theta_j}\left(e^{\phi_1} e^{\phi_2} \dots e^{\phi_n}\right)$$

$$= \frac{\partial}{\partial \theta_j}\left(e^{\phi_1}\right)\left(e^{\phi_2} e^{\phi_3} \dots e^{\phi_n}\right) + \left(e^{\phi_1}\right)\frac{\partial}{\partial \theta_j}\left(e^{\phi_2} e^{\phi_3} \dots e^{\phi_n}\right)$$
$$= \left(e^{\phi_1} e^{\phi_2} \dots e^{\phi_n}\right)\frac{\partial \phi_1}{\partial \theta_j} + \left(e^{\phi_1}\right)\frac{\partial}{\partial \theta_j}\left(e^{\phi_2} e^{\phi_3} \dots e^{\phi_n}\right).$$

Taking the next derivative, we have

$$= \left(e^{\phi_1} e^{\phi_2} \dots e^{\phi_n}\right)\frac{\partial \phi_1}{\partial \theta_j} + \left(e^{\phi_1} e^{\phi_2} \dots e^{\phi_n}\right)\frac{\partial \phi_2}{\partial \theta_j} + \left(e^{\phi_2}\right)\frac{\partial}{\partial \theta_j}\left(e^{\phi_3} e^{\phi_4} \dots e^{\phi_n}\right).$$

Furthermore, we can see that the derivative of the product $\frac{\partial}{\partial \theta_j}\prod_{i=1}^{n} e^{\phi_i}$ is equivalent to the sum of the products of e^{ϕ_i} over $\forall i$ multiplied by the partials $\frac{\partial \phi_i}{\partial \theta_j}$:

$$\frac{\partial}{\partial \theta_j}\prod_{i=1}^{n} e^{\phi_i} = \left(e^{\phi_1} e^{\phi_2} \dots e^{\phi_n}\right)\frac{\partial \phi_1}{\partial \theta_j} + \left(e^{\phi_1} e^{\phi_2} \dots e^{\phi_n}\right)\frac{\partial \phi_2}{\partial \theta_j} + \cdots + \left(e^{\phi_1} e^{\phi_2} \dots e^{\phi_n}\right)\frac{\partial \phi_n}{\partial \theta_j}$$
$$= \sum_{i=1}^{n}\prod_{i=1}^{n} e^{\phi_i}\frac{\partial \phi_i}{\partial \theta_j} = \sum_{i=1}^{n}\prod_{i=1}^{n} \exp\left\{\sum_{j=1}^{D} \theta_j f_j(Z_i, Z_{i-1}, X_i)\right\}\frac{\partial}{\partial \theta_j}\left(\sum_{j=1}^{D} \theta_j f_j(Z_i, Z_{i-1}, X_i)\right)$$
$$= \sum_{i=1}^{n}\prod_{i=1}^{n} \psi_i(Z_i, Z_{i-1}, X_i)\, f_j(Z_i, Z_{i-1}, X_i).$$

Finally, plugging this back into our original expression for $\frac{\partial \ell}{\partial \theta_j}$:

$$\frac{\partial \ell}{\partial \theta_j} = \sum_{i=1}^{n} f_j(Z_i, Z_{i-1}, X_i) - \frac{1}{Z}\sum_{Z_1,\dots,Z_n}\sum_{i=1}^{n}\prod_{i=1}^{n} \psi_i(Z_i, Z_{i-1}, X_i) f_j(Z_i, Z_{i-1}, X_i).$$

Therefore, completing the proof.

■

2) Using the forward and backward messages on a linear chain of random variables $X_1, \dots, X_n$, prove that

$$\sum_{X_1,\dots,X_n}\prod_{j=1}^{n-1} \psi_{j,j+1}(X_j, X_{j+1}) = \sum_{X_i} \mu_\alpha(X_i)\mu_\beta(X_i).$$

That is, prove Eq. (35) in the lecture notes of graphical models.

(Proof): From Eq. (34) of the lecture notes on graphical models, we know that

$$\mu_\alpha(X_i)\mu_\beta(X_i) = \sum_{\backslash X_i}\left(\prod_{j=1}^{i} \psi_{j,j+1}(X_j, X_{j+1}) \prod_{j=i+1}^{n-1} \psi_{j,j+1}(X_j, X_{j+1})\right).$$

From the right-hand side of the desired equation, we then have

$$\sum_{X_i} \mu_\alpha(X_i)\mu_\beta(X_i) = \sum_{X_i}\left[\sum_{\backslash X_i}\left(\prod_{j=1}^{i} \psi_{j,j+1}(X_j, X_{j+1}) \prod_{j=i+1}^{n-1} \psi_{j,j+1}(X_j, X_{j+1})\right)\right]$$

$$= \sum_{X_1,\dots,X_n} \left(\prod_{j=1}^{i} \psi_{j,j+1}(X_j, X_{j+1}) \prod_{j=i+1}^{n-1} \psi_{j,j+1}(X_j, X_{j+1}) \right)$$

$$= \sum_{X_1,\dots,X_n} \left(\prod_{j=1}^{n-1} \psi_{j,j+1}(X_j, X_{j+1}) \right).$$

Since this is equivalent to the left-hand side expression, this completes the proof.

■

3) In HMM (Hidden Markov Model) learning, given the inferred probabilities

$$\xi_i(j,k) = \mathbb{P}(Z_{i-1} = j, Z_i = k | \boldsymbol{\theta}^{old}),$$

With the old parameter $\boldsymbol{\theta}^{old}$, prove that the transition matrix can be updated as

$$A_{jk} = \frac{\sum_{i=1}^{n} \xi_i(j,k)}{\sum_{k=1}^{K} \sum_{i=1}^{n} \xi_i(j,k)}.$$

[*Hints: Use Eq.* (52) *in the lecture notes of graphical models and fix* k *to extract the summation* $\sum_{k=1}^{K} \sum_{i=1}^{n} \xi_i(j,k)\, ln\, A_{jk}$. *Then perform MLE of* A_{jk} *by treating* $\xi_i(j,k)$ *as observed data. You can use the Lagrangian method to solve this problem: introduce the Lagrangian multiplier* λ *and construct the Lagrangian function* $\mathcal{L}(A_{j1}, \dots, A_{jK}, \lambda) = \sum_{i=1}^{n} \sum_{k=1}^{K} \xi_i(j,k) \ln A_{jk} + \lambda\left(1 - \sum_{k=1}^{K} A_{jk}\right)$. *You may find question* 1 *of HW3 useful. After you solve* A_{jk}, *claim that the entire matrix* A *can be obtained by having* j *go from* 1 *to* K.]

(Proof): From Eq. (52) of the lecture notes, we are given the following lower-bound for the incomplete-data likelihood $\mathbb{P}(X_1, \dots, X_n | \boldsymbol{\theta})$ as

$$\sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{K} \xi_i(j,k) \ln A_{jk}.$$

If j is fixed, then this is equivalent to

$$\sum_{k=1}^{K} \sum_{i=1}^{n} \xi_i(j,k) \ln A_{jk}.$$

If we let $\xi_i(j,k)$ be observed data, then we can define the likelihood function of A_{jk} that we will maximize as

$$L\left(A_{jk}; \xi_i(j',k)\right) \triangleq \sum_{k=1}^{K} \sum_{i=1}^{n} \xi_i(j,k) \ln A_{jk}.$$

Since we know that $\sum_{k=1}^{K} A_{jk} = 1$, we can construct the Lagrangian function

$$\mathcal{L}(A_{jk}, \lambda) = \sum_{i=1}^{n} \sum_{k=1}^{K} \xi_i(j,k) \ln A_{jk} + \lambda \left(1 - \sum_{k=1}^{K} A_{jk} \right).$$

In order to maximize this function, we will take the derivative with respect to A_{jk} and set it to zero:

$$\frac{\partial \mathcal{L}(A_{jk},\lambda)}{\partial A_{jk}} = \sum_{i=1}^{n} \frac{\partial}{\partial A_{jk}} \sum_{k=1}^{K} \xi_i(j,k) \ln A_{jk} + \lambda \frac{\partial}{\partial A_{jk}} \left(1 - \sum_{k=1}^{K} A_{jk}\right)$$

$$= \sum_{i=1}^{n} \frac{\xi_i(j,k)}{A_{jk}} - \lambda = 0 \rightarrow \frac{1}{A_{jk}} \sum_{i=1}^{n} \xi_i(j,k) = \lambda$$

$$\rightarrow A_{jk} = \frac{\sum_{i=1}^{n} \xi_i(j,k)}{\lambda}.$$

Substituting this into the constraint, we can solve for λ at optimality:

$$\sum_{k=1}^{K} A_{jk} = 1 \rightarrow \sum_{k=1}^{K} \frac{\sum_{i=1}^{n} \xi_i(j,k)}{\lambda} = 1$$

$$\rightarrow \lambda = \sum_{k=1}^{K} \sum_{i=1}^{n} \xi_i(j,k).$$

Now, substituting this back into the above equation, we have the MLE estimate of A_{jk} as:

$$A_{jk} = \frac{\sum_{i=1}^{n} \xi_i(j,k)}{\sum_{k=1}^{K} \sum_{i=1}^{n} \xi_i(j,k)}.$$

Finally, the entire matrix A can be obtained by having j go from 1 to K; therefore, completing the proof.

■

4) Explain why the time complexity of running the Viterbi algorithm on a linear chain of n random variables, which each have K choices, is $O(nK^2)$.

(Solution): In the Viterbi algorithm, the following equation is used to calculate all messages $v_i(k)$ for $\forall k = 1, \dots, K$ and $\forall i = 1, \dots, n$ on a linear chin with n states:

$$v_i(k) = \max_j \left[v_{j-1}(j) A_{jk} \mathbb{P}(X_i | Z_i = k)\right].$$

From this expression, for a single $v_i(k)$ (with i fixed), j will iterate from $1, \dots, K$, which will yield a total of K operations. Since there are K choices for each i, j will be iterated from 1 to K a total of K times, which will yield a total of K^2 operations. Finally, since this process must be completed for $\forall i = 1, \dots, n$, there will be a total of nK^2 operations. Therefore, we can see that the Viterbi algorithm will have a time complexity of $O(nK^2)$.

■

5) (Graduate Only)

Derive the logistic regression model from a maximum entropy formulation.

Let training data be $\left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{m}$, where $x^{(i)} \in \mathbb{R}$ and $y^{(i)} \in \{0,1\}$. We are looking for a probabilistic model $\mathbb{P}(y|x)$ so that the expectation of the feature x under the distribution of the model $\mathbb{P}(y|x)$ equals the empirical average of the feature based on the training data:

$$\sum_{i=1}^{m} \mathbb{P}\left(y \middle| x^{(i)}\right) x^{(i)} = \sum_{i=1}^{m} 1\left[y^{(i)} = y\right] x^{(i)}, \quad \forall y = 0,1.$$

If $\mathbb{P}\left(y \middle| x^{(i)}\right)$ is a good predictor, then it should behave like the indicator function $1\left[y^{(i)} = y\right]$, which is 1 if $y^{(i)} = y$ and 0 otherwise. This is called the "balancing equation".

We have further constraints over $\mathbb{P}(y|x)$: $\sum_{y=0}^{1} \mathbb{P}(y|x) = 1$ and $\mathbb{P}(y|x) \geq 0$.

Subject to these constraints, we don't want to over-commit to a specific label y, but would like to make $\mathbb{P}(y|x)$ distributed as close to a uniform distribution as possible. This goal can be achieved by maximizing the entropy of the distribution $\mathbb{P}\left(y \middle| x^{(i)}\right)$ over the random variable $y \in \{0,1\}$ for each training example $i = 1, \ldots, m$:

$$-\sum_{i=1}^{m} \sum_{y=0}^{1} \mathbb{P}\left(y \middle| x^{(i)}\right) \log \mathbb{P}\left(y \middle| x^{(i)}\right).$$

If we let the two probabilities $\mathbb{P}(y|x)$ be two primal variables, then we have the following optimization problem:

$$\max_{\mathbb{P}(y|x):y=0,1} -\sum_{i=1}^{m} \sum_{y=0}^{1} \mathbb{P}\left(y \middle| x^{(i)}\right) \log \mathbb{P}\left(y \middle| x^{(i)}\right),$$

$$s.t. \quad \sum_{y=0}^{1} \mathbb{P}\left(y \middle| x^{(i)}\right) = 1,$$

$$\mathbb{P}\left(y \middle| x^{(i)}\right) \geq 0, \quad y \in \{0,1\},$$

$$\sum_{i=1}^{m} \mathbb{P}\left(y \middle| x^{(i)}\right) x^{(i)} = \sum_{i=1}^{m} 1\left[y^{(i)} = y\right] x^{(i)}, \quad \forall y = 0,1; \ \forall i = 1, \ldots, m.$$

Work on the following steps to find $\mathbb{P}(y|x) = \frac{e^{\lambda_y x}}{\sum_{y=0}^{1} e^{\lambda_y x}}$, where λ_y is the coefficient of the logistic regression model on the single input feature x for predicting class y.

- Introduce the Lagrangian multipliers λ_y for the last equality constraint for $y = 0,1$, and β_i for the first equality constraint for $\forall i = 1, \ldots, m$. Construct the Lagrangian function $\mathcal{L}(\mathbb{P}(y = 0|x), \mathbb{P}(y = 1|x), \lambda_0, \lambda_1, \beta_1, \ldots, \beta_m)$ of the above optimization problem, ignoring the constraint $\mathbb{P}\left(y \middle| x^{(i)}\right) \geq 0$.
- Take the partial derivative of $\mathcal{L}$ with respect to the primal variable $\mathbb{P}\left(y \middle| x^{(i)}\right)$ for a fixed y and $i = 1, \ldots, m$.
- Equate the partial derivative to 0 and use the constraint $\sum_{y=0}^{1} \mathbb{P}\left(y \middle| x^{(i)}\right) = 1$ to eliminate β_i and find $\mathbb{P}\left(y \middle| x^{(i)}\right) = \frac{e^{\lambda_y x^{(i)}}}{\sum_{y=0}^{1} e^{\lambda_y x^{(i)}}}$.

(Proof): To begin, we can construct the following Lagrangian function:

$$\mathcal{L}(\mathbb{P}(y = 0|x), \mathbb{P}(y = 1|x), \lambda_0, \lambda_1, \beta_1, \ldots, \beta_m)$$

$$= -\sum_{i=1}^{m}\sum_{y=0}^{1}\mathbb{P}\left(y\middle|x^{(i)}\right)\log\mathbb{P}\left(y\middle|x^{(i)}\right) + \sum_{i=1}^{m}\beta_i\left(\sum_{y=0}^{1}\mathbb{P}\left(y\middle|x^{(i)}\right) - 1\right)$$
$$+ \sum_{y=0}^{1}\lambda_y\left(\sum_{i=1}^{m}\mathbb{P}\left(y\middle|x^{(i)}\right)x^{(i)} - \sum_{i=1}^{m}1\left[y^{(i)} = y\right]x^{(i)}\right).$$

Now, taking the partial derivative of $\mathcal{L}$ with respect to the primal variable $\mathbb{P}\left(y\middle|x^{(i)}\right)$ for a fixed y and $i = 1, \dots, m$, and setting it equal to 0, we have

$$\frac{\partial\mathcal{L}}{\partial\mathbb{P}\left(y\middle|x^{(i)}\right)}$$
$$= \frac{\partial}{\partial\mathbb{P}\left(y\middle|x^{(i)}\right)}\left[-\sum_{i=1}^{m}\sum_{y=0}^{1}\mathbb{P}\left(y\middle|x^{(i)}\right)\log\mathbb{P}\left(y\middle|x^{(i)}\right)\right] + \beta_i\frac{\partial}{\partial\mathbb{P}\left(y\middle|x^{(i)}\right)}\left(\sum_{y=0}^{1}\mathbb{P}\left(y\middle|x^{(i)}\right) - 1\right)$$
$$+ \lambda_y\frac{\partial}{\partial\mathbb{P}\left(y\middle|x^{(i)}\right)}\left(\sum_{i=1}^{m}\mathbb{P}\left(y\middle|x^{(i)}\right)x^{(i)} - \sum_{i=1}^{m}1\left[y^{(i)} = y\right]x^{(i)}\right)$$
$$= -(1)\log\mathbb{P}\left(y\middle|x^{(i)}\right) - \mathbb{P}\left(y\middle|x^{(i)}\right)\frac{1}{\mathbb{P}\left(y\middle|x^{(i)}\right)} + \beta_i(1) + \lambda_y x^{(i)} = 0$$
$$\rightarrow -\log\mathbb{P}\left(y\middle|x^{(i)}\right) - 1 + \beta_i + \lambda_y x^{(i)} = 0$$
$$\rightarrow \log\mathbb{P}\left(y\middle|x^{(i)}\right) = \beta_i - 1 + \lambda_y x^{(i)}$$
$$\rightarrow \mathbb{P}\left(y\middle|x^{(i)}\right) = e^{\beta_i - 1 + \lambda_y x^{(i)}}.$$

Using the constraint $\sum_{y=0}^{1}\mathbb{P}\left(y\middle|x^{(i)}\right) = 1$, we have

$$\sum_{y=0}^{1}e^{\beta_i + \lambda_y x^{(i)} - 1} = 1 \rightarrow e^{\beta_i - 1}\sum_{y=0}^{1}e^{\lambda_y x^{(i)}} = 1$$
$$\rightarrow e^{\beta_i - 1} = \frac{1}{\sum_{y=0}^{1}e^{\lambda_y x^{(i)}}}.$$

Plugging this into the above expression, we obtain

$$\mathbb{P}\left(y\middle|x^{(i)}\right) = e^{\beta_i - 1 + \lambda_y x^{(i)}} = e^{\beta_i - 1}e^{\lambda_y x^{(i)}}$$
$$= \frac{e^{\lambda_y x^{(i)}}}{\sum_{y=0}^{1}e^{\lambda_y x^{(i)}}}.$$

Therefore, we have completed the proof.

■