Appendix for "Lifelong Learning for Sentiment Classification"

Zhiyuan Chen, Nianzu Ma, Bing Liu
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607, USA
{czyuanacm,jingyima005}@gmail.com,liub@cs.uic.edu

This appendix includes the detailed derivation steps for the proposed Lifelong Sentiment Classification (LSC) model.

Recall that our objective function under stochastic gradient descent for each target domain training document d_i is defined as below:

$$F_{+,i} = P\left(c_j|d_i\right) - P(c_f|d_i) \tag{9}$$

where c_j is the correct label of document d_i and c_f is the wrong label of document d_i. Eq. 9 is written as below after plugging probabilities from Naïve Bayesian text classification:

$$\frac{P(c_j)\prod_{w\in d_i} P(w|c_j)^{n_{w,d_i}}}{\sum_{r=1}^{|C|} P(c_r)\prod_{w\in d_i} P(w|c_r)^{n_{w,d_i}}} - \frac{P(c_f)\prod_{w\in d_i} P(w|c_f)^{n_{w,d_i}}}{\sum_{r=1}^{|C|} P(c_r)\prod_{w\in d_i} P(w|c_r)^{n_{w,d_i}}} \tag{10}$$

Below, we first work on the derivation for a positive document d_i, i.e., $c_j = +$ and $c_f = -$ for document d_i, which gives us:

$$P(+)\prod_{w\in d_i} P(w|+)^{n_{w,d_i}} - P(-)\prod_{w\in d_i} P(w|-)^{n_{w,d_i}} \tag{11}$$

Here we leave out the denominator of Eq. 10 for the time being and work only on the numerators (we will bring the denominator back in Eq. 13).

Now we plug Eq. 1 (in the submitted paper) into Eq. 11:

$$\frac{P\left(+\right)\prod_{w\in d_i}\left(\lambda + X_{+,w}\right)^{n_{w,d_i}}}{\left(\lambda\left|V\right| + \sum_{v=1}^{|V|} X_{+,v}\right)^{|d_i|}} - \frac{P\left(-\right)\beta^{|d_i|}\prod_{w\in d_i}\left(\lambda + X_{-,w}\right)^{n_{w,d_i}}}{\left(\lambda\left|V\right| + \sum_{v=1}^{|V|} X_{+,v}\right)^{|d_i|}} \tag{12}$$

where $|d_i|$ is the number of words in d_i and $\beta = (\lambda\left|V\right| + \sum_{v=1}^{|V|} X_{+,v})/(\lambda\left|V\right| + \sum_{v=1}^{|V|} X_{-,v})$.

Now let us bring back the denominator in Eq. 10, which is nothing but Eq. 12 except that instead of subtraction, it uses summation. After canceling the common denominator, we obtain:

$$\frac{P(+)\prod_{w\in d_i}\left(\lambda + X_{+,w}\right)^{n_{w,d_i}} - P(-)\beta^{|d_i|}\prod_{w\in d_i}\left(\lambda + X_{-,w}\right)^{n_{w,d_i}}}{P(+)\prod_{w\in d_i}\left(\lambda + X_{+,w}\right)^{n_{w,d_i}} + P(-)\beta^{|d_i|}\prod_{w\in d_i}\left(\lambda + X_{-,w}\right)^{n_{w,d_i}}} \tag{13}$$

To make sure Eq. 13 gives a positive value for taking log, we first add 1 to it. Then we take the log. These do not change the maximization solution. Last, we negate the equation to make it a minimization problem for gradient descent:

$$log\left(P(+)\prod_{w\in d_i}\left(\lambda + X_{+,w}\right)^{n_{w,d_i}} + P(-)\beta^{|d_i|}\prod_{w\in d_i}\left(\lambda + X_{-,w}\right)^{n_{w,d_i}}\right) - log\left(2\times P(+)\prod_{w\in d_i}\left(\lambda + X_{+,w}\right)^{n_{w,d_i}}\right) \tag{14}$$

Eq. 14 is the objective function that we want to minimize for a positive training document d_i. Note that this objective function is not convex.

We now compute the gradients by taking partial derivatives on Eq. 14. We define $g(\boldsymbol{X})$, a function of $\boldsymbol{X}$ where $\boldsymbol{X}$ is a vector consisting of $X_{+,w}$ and $X_{-,w}$ of each word w:

$$g\left(\boldsymbol{X}\right) = \beta^{|d_i|} = \left(\frac{\lambda\left|V\right| + \sum_{v=1}^{|V|} X_{+,v}}{\lambda\left|V\right| + \sum_{v=1}^{|V|} X_{-,v}}\right)^{|d_i|} \tag{15}$$

The partial derivatives for a word u, i.e., $\frac{\partial g}{\partial X_{+,u}}$ and $\frac{\partial g}{\partial X_{-,u}}$, are quite straightforward and thus not shown here. The final partial derivatives for a word u on Eq. 14 is shown below:

$$\frac{\partial F_{+,i}}{\partial X_{+,u}} = \frac{\frac{n_{u,d_i}}{\lambda+X_{+,u}} + \frac{P(-)}{P(+)} \prod_{w \in d_i} \left(\frac{\lambda+X_{-,w}}{\lambda+X_{+,w}}\right)^{n_{w,d_i}} \times \frac{\partial g}{\partial X_{+,u}}}{1 + \frac{P(-)}{P(+)} \prod_{w \in d_i} \left(\frac{\lambda+X_{-,w}}{\lambda+X_{+,w}}\right)^{n_{w,d_i}} \times g(\boldsymbol{X})} - \frac{n_{u,d_i}}{\lambda + X_{+,u}} \tag{16}$$

$$\frac{\partial F_{+,i}}{\partial X_{-,u}} = \frac{\frac{n_{u,d_i}}{\lambda+X_{-,u}} \times g(\boldsymbol{X}) + \frac{\partial g}{\partial X_{-,u}}}{\frac{P(+)}{P(-)} \prod_{w \in d_i} \left(\frac{\lambda+X_{+,w}}{\lambda+X_{-,w}}\right)^{n_{w,d_i}} + g(\boldsymbol{X})} \tag{17}$$

Negative document. We can follow the same process and get the corresponding objective function $F_{-,i}$ for a negative document. Then the final partial derivatives can be obtained following the same process. We gave the final results directly:

$$\frac{\partial F_{-,i}}{\partial X_{+,u}} = \frac{\frac{n_{u,d_i}}{\lambda+X_{+,u}} + \frac{P(-)}{P(+)} \prod_{w \in d_i} \left(\frac{\lambda+X_{-,w}}{\lambda+X_{+,w}}\right)^{n_{w,d_i}} \times \frac{\partial g}{\partial X_{+,u}}}{1 + \frac{P(-)}{P(+)} \prod_{w \in d_i} \left(\frac{\lambda+X_{-,w}}{\lambda+X_{+,w}}\right)^{n_{w,d_i}} \times g(\boldsymbol{X})} - \frac{\partial g}{\partial X_{+,u}} \times \frac{1}{g(\boldsymbol{X})} \tag{18}$$

$$\frac{\partial F_{-,i}}{\partial X_{-,u}} = \frac{\frac{n_{u,d_i}}{\lambda+X_{-,u}} \times g(\boldsymbol{X}) + \frac{\partial g}{\partial X_{-,u}}}{\frac{P(+)}{P(-)} \prod_{w \in d_i} \left(\frac{\lambda+X_{+,w}}{\lambda+X_{-,w}}\right)^{n_{w,d_i}} + g(\boldsymbol{X})} - \frac{\frac{n_{u,d_i}}{\lambda+X_{-,u}} \times g(\boldsymbol{X}) + \frac{\partial g}{\partial X_{-,u}}}{g(\boldsymbol{X})} \tag{19}$$