

TP2

Thibault Duhamel 18026048, Heng Shi 18171434

February 21, 2019

Question 1

The loss function is defined as follows:

$$E(\vec{w}) = \sum_n^N (t_n - \vec{w}^T \cdot \vec{\phi}_n)^2 + \lambda \vec{w}^T \cdot \vec{w}$$

As we want to minimize the above expression, it is convenient to calculate its gradient:

$$\frac{dE(\vec{w})}{d\vec{w}} = \sum_n^N [-2(t_n - \vec{w}^T \cdot \vec{\phi}_n) \vec{\phi}_n^T] + 2\lambda \vec{w}$$

Let us now set this gradient to zero in order to deduce \vec{w} :

$$\frac{dE(\vec{w})}{d\vec{w}} = 0$$

$$\sum_n^N [-2(t_n - \vec{w}^T \cdot \vec{\phi}_n) \vec{\phi}_n^T] + 2\lambda \vec{w} = 0$$

$$\sum_n^N [(\vec{w}^T \cdot \vec{\phi}_n) \vec{\phi}_n^T] - \sum_n^N [t_n \vec{\phi}_n^T] + \lambda \vec{w} = 0$$

We can write the exact same equality using matrices, with Φ being the same matrix as defined in the book from Bishop:

$$\Phi^T \Phi \vec{w} - \Phi^T \vec{t} + \lambda \vec{w} = 0$$

$$\Phi^T \Phi \vec{w} + \lambda \vec{w} = \Phi^T \vec{t}$$

$$(\Phi^T \Phi + \lambda I) \vec{w} = \Phi^T \vec{t}$$

And finally, assuming the matrix $(\Phi^T \Phi + \lambda I)$ is invertible:

$$\boxed{\vec{w} = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T \vec{t}}$$

Question 2

In this question, we do not write the transpose symbol to simplify expressions.

The cross-entropy loss function is defined as follows:

$$E(\vec{w}) = - \sum_n^N [t_n \ln(\sigma(\vec{w} \cdot \vec{\phi}_n)) + (1 - t_n) \ln(1 - \sigma(\vec{w} \cdot \vec{\phi}_n))]$$

Let us first compute the gradient of the sigmoid function:

$$\frac{d\sigma}{d\vec{w}} = \frac{\vec{\phi}_n e^{-\vec{w} \cdot \vec{\phi}_n}}{(1 + e^{-\vec{w} \cdot \vec{\phi}_n})^2}$$

Then, we can use this intermediate result to calculate the gradient of the 2 logarithms in the sum:

$$\begin{aligned} \frac{d \ln(\sigma)}{d\vec{w}} &= \frac{\frac{d\sigma}{d\vec{w}}}{\sigma} \\ &= \frac{\vec{\phi}_n e^{-\vec{w} \cdot \vec{\phi}_n}}{(1 + e^{-\vec{w} \cdot \vec{\phi}_n})^2} \times (1 + e^{-\vec{w} \cdot \vec{\phi}_n}) \\ &= \frac{\vec{\phi}_n e^{-\vec{w} \cdot \vec{\phi}_n}}{1 + e^{-\vec{w} \cdot \vec{\phi}_n}} \end{aligned}$$

and:

$$\begin{aligned} \frac{d \ln(1 - \sigma)}{d\vec{w}} &= \frac{-\frac{d\sigma}{d\vec{w}}}{1 - \sigma} \\ &= \frac{-\vec{\phi}_n e^{-\vec{w} \cdot \vec{\phi}_n}}{(1 + e^{-\vec{w} \cdot \vec{\phi}_n})^2} \times \frac{1 + e^{-\vec{w} \cdot \vec{\phi}_n}}{e^{-\vec{w} \cdot \vec{\phi}_n}} \\ &= \frac{-\vec{\phi}_n}{1 + e^{-\vec{w} \cdot \vec{\phi}_n}} \end{aligned}$$

From those 2 expressions, we are now able to express the gradient of the whole sum:

$$\begin{aligned} \frac{dE(\vec{w})}{d\vec{w}} &= - \sum_n^N \left[t_n \frac{\vec{\phi}_n e^{-\vec{w} \cdot \vec{\phi}_n}}{1 + e^{-\vec{w} \cdot \vec{\phi}_n}} + (1 - t_n) \frac{-\vec{\phi}_n}{1 + e^{-\vec{w} \cdot \vec{\phi}_n}} \right] \\ &= - \sum_n^N \left[t_n \frac{e^{-\vec{w} \cdot \vec{\phi}_n}}{1 + e^{-\vec{w} \cdot \vec{\phi}_n}} + (1 - t_n) \frac{-1}{1 + e^{-\vec{w} \cdot \vec{\phi}_n}} \right] \vec{\phi}_n \end{aligned}$$

$$\begin{aligned}
&= - \sum_n^N \left[t_n y_n e^{-\vec{w} \cdot \vec{\phi}_n} + (t_n - 1) y_n \right] \vec{\phi}_n \\
&= - \sum_n^N \left[t_n y_n (1 + e^{-\vec{w} \cdot \vec{\phi}_n}) - y_n \right] \vec{\phi}_n \\
&= - \sum_n^N \left[t_n y_n \frac{1}{y_n} - y_n \right] \vec{\phi}_n \\
&= - \sum_n^N [t_n - y_n] \vec{\phi}_n \\
&\boxed{\frac{dE(\vec{w})}{d\vec{w}} = \sum_n^N [y_n - t_n] \vec{\phi}_n}
\end{aligned}$$

Question 3

Let us first write down the definition of the entropy in this context:

$$H(X) = -p_1 \log_2(p_1) - p_2 \log_2(p_2) - p_3 \log_2(p_3)$$

As we want to maximize this function (with $p_1 + p_2 + p_3 = 1$) we can express the problem with a Lagrange multiplier:

$$L = p_1 \log_2(p_1) + p_2 \log_2(p_2) + p_3 \log_2(p_3) + \lambda(p_1 + p_2 + p_3 - 1)$$

However, we do have another condition on those probabilities (that is $p_1 = 2p_2$), so we can modify the expression consequently:

$$L = 2p_2 \log_2(2p_2) + p_2 \log_2(p_2) + p_3 \log_2(p_3) + \lambda(3p_2 + p_3 - 1)$$

$$L = 2p_2 + 2p_2 \log_2(p_2) + p_2 \log_2(p_2) + p_3 \log_2(p_3) + \lambda(3p_2 + p_3 - 1)$$

$$L = 2p_2 + 3p_2 \log_2(p_2) + p_3 \log_2(p_3) + \lambda(3p_2 + p_3 - 1)$$

Let us now compute the gradients of such a function:

$$\begin{cases} \frac{dL}{dp_2} = 2 + 3 \log_2(p_2) + \frac{3}{\ln(2)} + 3\lambda \\ \frac{dL}{dp_3} = \log_2(p_3) + \frac{1}{\ln(2)} + \lambda \end{cases}$$

We can then set those gradients to zero:

$$\begin{cases} 2 + 3\log_2(p_2) + \frac{3}{\ln(2)} + 3\lambda = 0 \\ \log_2(p_3) + \frac{1}{\ln(2)} + \lambda = 0 \end{cases}$$

Let us first divide the first equation by 3:

$$\begin{cases} \frac{2}{3} + \log_2(p_2) + \frac{1}{\ln(2)} + \lambda = 0 \\ \log_2(p_3) + \frac{1}{\ln(2)} + \lambda = 0 \end{cases}$$

As both equations contain the same terms, let us subtract L_2 from L_1 :

$$\log_2\left(\frac{p_3}{p_2}\right) = \frac{2}{3}$$

It is now possible to extract a relation between p_3 and p_2 :

$$p_3 = 2^{2/3}p_2$$

As all probabilities now depend on p_2 , we can insert those terms inside the condition $p_1 + p_2 + p_3 = 1$:

$$2p_2 + p_2 + 2^{2/3}p_2 = 1$$

Which gives:

$$p_2 = \frac{1}{2^{2/3} + 3}$$

Now that we know p_2 , it is eventually trivial to compute p_1 and p_3 :

$$p_1 = \frac{2}{2^{2/3} + 3} \text{ and } p_3 = \frac{2^{2/3}}{2^{2/3} + 3}$$