CSCI 567 Homework 2

Shamim Samadi

October 8, 2017

Problem 1. Neural networks

- **Question 1.1**

1. $\frac{\partial l}{\partial u}$

$$\frac{\partial l}{\partial u} = \frac{\partial l}{\partial a}\frac{\partial a}{\partial u} = \frac{\partial l}{\partial a}\frac{\partial a}{\partial h}\frac{\partial h}{\partial u} = (W^{(2)^T}.\frac{\partial l}{\partial a}). * H(u) \tag{1}$$

2. $\frac{\partial l}{\partial a}$

$$\frac{\partial l}{\partial a} = \frac{\partial l}{\partial z}\frac{\partial z}{\partial a} \tag{2}$$

Using the definition $l = -\sum_k y_k \log z_k$, we have: $\frac{\partial l}{\partial a_m} = \frac{\partial l}{\partial z_k}\frac{\partial z_k}{\partial a_m} = (\frac{-y_k}{z_k}).\frac{\partial z_k}{\partial a_m}$. Now let's compute $\frac{\partial z_k}{\partial a_m}$:

$$\frac{\partial z_k}{\partial a_m} = \begin{cases} \frac{e^{a_k}}{\sum_k e^{a_k}} - (\frac{e^{a_k}}{\sum_k e^{a_k}})^2, & k = m \\ -\frac{e^{a_m}e^{a_k}}{(\sum_k e^{a_k})^2}, & k \neq m \end{cases} \Rightarrow \frac{\partial z_k}{\partial a_m} = \begin{cases} z_k - z_k^2 = z_k(1-z_k), & k = m \\ -z_k z_m, & k \neq m \end{cases}$$

Thus, $\frac{\partial l}{\partial a_m} = (\frac{-y_K}{z_K}).z_K(1-z_K) + \sum_{K\neq m}(-z_K z_m).(\frac{-y_K}{z_K}) = -y_m(1-z_m) + \sum_{K\neq m} z_m y_K = z_m \sum_K y_K - y_m = z_m - y_m$

3. $\frac{\partial l}{\partial W^{(1)}}$

$$\frac{\partial l}{\partial W^{(1)}} = \frac{\partial l}{\partial u}.\frac{\partial u}{\partial W^{(1)}} = \frac{\partial l}{\partial u}.x^T \tag{3}$$

4. $\frac{\partial l}{\partial b^{(1)}}$

$$\frac{\partial l}{\partial b^{(1)}} = \frac{\partial l}{\partial u}.\frac{\partial u}{\partial b^{(1)}} = \frac{\partial l}{\partial u}.(1) = \frac{\partial l}{\partial u} \tag{4}$$

5. $\frac{\partial l}{\partial a}$

$$\frac{\partial l}{\partial W^{(2)}} = \frac{\partial l}{\partial a}.\frac{\partial a}{\partial W^{(2)}} = \frac{\partial l}{\partial a}.h^T \tag{5}$$

- **Question 1.2**
In gradient descent, to update a parameter, we take steps proportional to the negative of the gradient, and thus no learning happens if the gradient is zero (parameter value always stays at the initial value).

- **Question 1.3**

$$a = W^{(2)}u + b^{(2)}, u = W^{(1)}x + b^{(1)} \Rightarrow \tag{6}$$

$$a = W^{(2)}(W^{(1)}x + b^{(1)}) + b^{(2)} = W^{(2)}W^{(1)}x + (W^{(2)}b^{(1)} + b^{(2)}) \tag{7}$$

i.e. $U = W^{(2)}W^{(1)}$ and $v = W^{(2)}b^{(1)} + b^{(2)}$.

Problem 2. Kernel Methods

- **Question 2.1**

Assume $\frac{\partial l(s,y)}{\partial s}$ is a know quantity: $\frac{\partial l(s,y)}{\partial s} = s'$

$$\frac{\partial l(s,y)}{\partial w} = \sum_n \frac{\partial l}{\partial s}.\phi(x_n) + \lambda w^T = s' \sum_n \phi(x_n) + \lambda w^T = 0 \tag{8}$$

thus

$$w^* = \sum_n \frac{-s'}{\lambda}\phi(x_n) \tag{9}$$

or

$$w^* = \sum_{n=1}^{N} \alpha_n \phi(x_n), \tag{10}$$

where: $\alpha_n = \frac{-s'}{\lambda}$.

- **Question 2.2**

$$w^* = \sum_{n=1}^{N} \alpha_n \phi(x_n) = \Phi^T \alpha \tag{11}$$

$$J(\alpha) = \min_\alpha \sum_n l((\phi^T\alpha)^T \phi(x_n), y_n) + \frac{\lambda}{2}(\phi^T\alpha)^T(\phi^T\alpha) \tag{12}$$

$$J(\alpha) = \min_\alpha \sum_n l(\alpha^T \Phi\phi(x_n), y_n) + \frac{\lambda}{2}\alpha^T\Phi\Phi^T\alpha \tag{13}$$

which using $K = \Phi\Phi^T$, leads to

$$J(\alpha) = \min_\alpha \sum_n l(\alpha^T K_n, y_n) + \frac{\lambda}{2}\alpha^T K\alpha \tag{14}$$