Machine Learning - Deep Learning: Assignment 9

Harshita Agarwala

January 7, 2018

1 Problem 1

Basis functions are required in neural networks as they help in mapping the data set to a linearly separable space. It then becomes easier to find a hyperplane that would separate the data points.

2 Problem 2

We can show that tanh(x) can be written in terms of $\sigma(x)$

$$\begin{aligned} 2\sigma(2x) - 1 &= \frac{2}{e^{-2x}+1} - 1 \\ &= \frac{2 - e^{-2x} - 1}{e^{-2x}+1} \\ &= \frac{1 - e^{-2x}}{e^{-2x}+1} \\ &= \frac{e^{2x} - 1}{e^{2x}+1} \\ &= \frac{e^x(e^x - e^{-x})}{e^x(e^x + e^{-x})} \\ &= tanh(x) \end{aligned}$$

Therefore we can reproduce an equivalent network using tanh(x) activation function.

3 Problem 3

Now we know that $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$
Therefore derivative of tanh(x) is:

$$\begin{aligned} \frac{d}{dx}\tanh(x) &= \frac{d}{dx}\frac{e^x - e^{-x}}{e^x + e^{-x}} \\ &= \frac{(e^x - e^{-x})(e^x + e^{-x})}{(e^x + e^{-x})^2} \\ &= \frac{(e^x + e^{-x})(e^x + e^{-x}) - (e^x - e^{-x})(e^x - e^{-x})}{(e^x + e^{-x})^2} \\ &= 1 - \tanh(x)^2 \end{aligned}$$

This property is useful as it simplifies the derivative and reduces computation. Already computed values can be reused.

4 Problem 4

The appropriate choice of the activation function would be tanh(x). This is because tanh(x) is a scaled sigmoid function. The output of a sigmoid function is between 0 and 1. However, the tanh(x) function rescales the values between -1 and 1.
As shown in Problem 2, we can replace $f(x_i, w)$ with $(2f(2x_i, 2w)) - 1)$ in E(w)

5 Problem 5

$$E(w) = \frac{1}{m}\sum_{i=1}^{m} l(y_i - wx_i) - \frac{\lambda}{2}||w||^2$$

$$\nabla E(w) = \frac{1}{m}\sum_{i=1}^{m} \nabla l(y_i - wx_i) - \lambda w$$

$$\nabla l(y_i - wx_i) = \begin{cases} -x_i(y_i - wx_i) & \text{if } |(y_i - wx_i)| < 1 \\ -x_i sgn(y_i - wx_i) & \text{otherwise} \end{cases}$$

6 Problem 6

We know that over-fitting of the training data set leads to higher erros in the test and validation data sets. Therefore, relating to the graph, we should stop updating at approximately 50th iteration. At this stage the slope of the error graph for the training data set reduces which means that with every iteration only a small improvement occurs in the training set. Hence this is not a general improvemment and therefore leads to over-fitting.

7 Problem 7

$$\begin{aligned} a + \log\sum_{i=1}^{N} e^{x_i - a} &= a\log e + \log\sum_{i=1}^{N} e^{x_i - a} \\ &= \log e^a + \sum_{i=1}^{N} e^{x_i - a} \\ &= \log\left[e^a \sum_{i=1}^{N} e^{x_i - a}\right] \\ &= \log\sum_{i=1}^{N} e^{x_i - a}.e^a \\ &= \log\sum_{i=1}^{N} e^{x_i} \end{aligned}$$

8 Problem 8

For any arbitrary a, we can show that:

$$\frac{e^{x_i - a}}{\sum_{i=1}^{N} e^{x_i - a}} = \frac{e^{x_i - a}}{\sum_{i=1}^{N} e^{x_i - a}}.\frac{e^a}{e^a} = \frac{e^{x_i}}{\sum_{i=1}^{N} e^{x_i}}$$

9 Problem 9

$$\begin{aligned} & -(y\log(\sigma(x)) + (1-y)\log(1-\sigma(x))) \\ &= -\left[y\log\left(\frac{1}{1+e^{-x}}\right) + (1-y)\log\left(1 - \frac{1}{1+e^{-x}}\right)\right] \\ &= -\left[-y\log(1+e^{-x}) + (1-y)\log\left(\frac{e^{-x}}{1+e^{x}}\right)\right] \\ &= -\left[-y\log(1+e^{-x}) + \log(e^{-x}) - \log(1+-x) - y\log(e^{-x}) + y\log(1+e^{-x})\right] \\ &= -\left[-x - log(1+e^{-x}) + xy\right] \\ &= x + log(1+e^{-x}) - xy \end{aligned}$$

Now, as this is a cost function, we need to ensure that the value is positive. Therefore, we subsitute x with $abs(x)$ in the logarithmic function and with $\max(x, 0)$ in place of x

$$x + log(1+e^{-x}) - xy \;\sim\; \max(x,0) + log(1+e^{-abs(x)}) - xy$$