Assignment 2 & 3

Paul Thillen
IFT3395/6390 - Machine learning

November 20, 2017

1 Theoretical part A

1.1

To show:

$$sigmoid(x) = \frac{1}{2}(tanh(\frac{x}{2}) + 1)$$

Which is equivalent to showing:

$$tanh(x) = 2 \cdot sigmoid(2x) - 1$$

We have:

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

and:

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

consequently:

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

1.2

To show:

$$ln(sigmoid(x)) = -softplus(-x)$$

We have:

$$ln(sigmoid(x)) = ln(\frac{1}{1+e^{-x}}) = -ln(1+e^{-x}) = -softplus(-x)$$

1.3

To show:

$$sigmoid'(x) = sigmoid(x) \cdot (1 - sigmoid(x))$$

We have:

$$\begin{aligned} sigmoid'(x) &= (\frac{e^x}{1+e^x})' \\ &= \frac{e^x(1+e^x) - e^x \cdot e^x}{(1+e^x)^2} \\ &= \frac{e^x}{1+e^x}(1 - \frac{e^x}{1+e^x}) \\ &= sigmoid(x) \cdot (1 - sigmoid(x)) \end{aligned}$$

1.4

To show:

$$tanh'(x) = 1 - tanh^2(x)$$

We have:

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

1.5 Write sign using only indicator functions

$$sgn(x) = \mathbb{1}_{\mathbb{R}_+}(x) - \mathbb{1}_{\mathbb{R}_-}(x)$$

1.6 Derivative of abs

$$abs(x) = \begin{cases} x, \text{ if } x > 0 \\ 0, \text{ if } x = 0 \\ -x, \text{ if } x < 0 \end{cases}$$

$$abs'(x) = \begin{cases} 1, \text{ if } x > 0 \\ 0, \text{ if } x = 0 \\ -1, \text{ if } x < 0 \end{cases}$$

1.7 Derivative of rect

$$rect(x) = \begin{cases} x, \text{ if } x > 0 \\ 0, \text{ else} \end{cases} = \mathbb{1}_{\{x>0\}}(x) \cdot x$$

$$rect'(x) = \mathbb{1}_{\{x>0\}}(x)$$

1.8 L2 gradient

$$\frac{\partial ||x||_2^2}{\partial x} = \begin{pmatrix} \frac{\partial}{\partial x_1} ||x||_2^2 \\ ... \\ \frac{\partial}{\partial x_d} ||x||_2^2 \end{pmatrix} = \begin{pmatrix} 2x_1 \\ ... \\ 2x_d \end{pmatrix}$$

1.9 L1 gradient

$$\frac{\partial ||x||_1}{\partial x} = \begin{pmatrix} \frac{\partial}{\partial x_1} ||x||_1 \\ ... \\ \frac{\partial}{\partial x_d} ||x||_1 \end{pmatrix} = \begin{pmatrix} abs'(x_1) \\ ... \\ abs'(x_d) \end{pmatrix}$$

2 Theoretical part B

2.1

Dimensions of $W^{(1)}$ and $b^{(1)}$:

$$dim(W^{(1)}) = d_h \times d$$
$$dim(b^{(1)}) = d_h$$

Preactivation vector of neurons of the hidden layer h^a where $w_j^{(1)}$ is the j-th row of $W^{(1)}$.

$$h^a = W^{(1)} \cdot x + b^{(1)}$$
$$h_j^a = w_j^{(1)} \cdot x + b_j^{(1)}$$

Ouput vector of the hidden layer h^s:

$$h^s = rect(h^a)$$
$$h_k^s = max(0, h_k^a)$$

2.2

Dimensions of $W^{(2)}$ and $b^{(2)}$:

$$dim(W^{(2)}) = m \times d_h$$
$$dim(b^{(2)}) = m$$

Preactivation vector of neurons of the output layer o^a where $w_j^{(2)}$ is the j-th row of $W^{(2)}$.

$$o^a = W^{(1)} \cdot h^s + b^{(1)}$$
$$o_j^a = w_j^{(1)} \cdot h^s + b_j^{(1)}$$

2.3

Ouput vector of the output layer o^s:

$$o^s = softmax(o^a)$$
$$o_k^s = \frac{e^{o_k^a}}{\sum\limits_{i=1}^{m} e^{o_i^a}}$$

Since exponentials are always positive and both denominator and numerator are exponentials or sum of exponentials, o_k^s has to be positive too.

If we sum over all k for o_k^s, we receive:

$$\sum_{j=1}^{m} \frac{e^{o_j^a}}{\sum\limits_{i=1}^{m} e^{o_i^a}} = \frac{\sum\limits_{j=1}^{m} e^{o_j^a}}{\sum\limits_{i=1}^{m}} = \frac{\sum\limits_{i=1}^{m} e^{o_i^a}}{\sum\limits_{i=1}^{m} e^{o_i^a}} = 1$$

These two properties are important because o^s is a probability distribution for each possible class.

2.4

2.5

2.6

2.7

2.8

2.9

2.10

2.11

2.12

2.13

2.14

2.15

2.16

2.17