Machine Learnig Homework 11

Problem 1

Consider a neural network with an input layer, a hidden layer and an output layer. Assume a linear activation function in the output layer and let the output layer take form of a single neuron. The output of the network can be expressed as

$$y(x) = V^T a(x) + c = V^T f(z) = V^T f(W^T x + b) \tag{1}$$

where V weight matrix from the hidden layer to the output, a the activation of the hidden layer, f the hidden layer's nonlinearity, which acts on its input element wise, W the weight matrix from the input to the hidden layer, b and c are biases and x the input vector.

Sigmoid is a scaled and translated version of tanh, since

$$2\sigma(2x-1) = \frac{2}{1+e^{-2x}} - 1 = \frac{2e^{2x}}{e^{2x}+1} - 1 = \frac{e^{2x}-1}{e^{2x}+1} \quad = tanh(x) \tag{2}$$

therefore, if we take $W = 2\widetilde{W}$ and $b = 2\widetilde{b}$ it is obvious that

$$2\sigma(W^T x + b) - 1 = tanh(\widetilde{W}^T x + \widetilde{b}) \tag{3}$$

Weights V and the bias c of the final layer can easily account for scaling and translation of the resulting hidden's layer activation.

Problem 2

$$\frac{d}{dx}\sigma(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1+e^{-x}}{(1+e^{-x})^2} - \frac{1}{(1+e^{-x})^2} = \sigma(x) - \sigma^2(x) = \sigma(x)(1-\sigma(x)) \tag{4}$$

$$\begin{aligned} \frac{d}{dx}tanh(x) &= \frac{2e^{2x}\left((e^{2x}+1) - (e^{2x}-1)\right)}{(e^{2x}+1)^2} \\ &= \frac{\left((e^{2x}+1) + (e^{2x}-1)\right)\left((e^{2x}+1) - (e^{2x}-1)\right)}{(e^{2x}+1)^2} \\ &= \frac{(e^{2x}+1)^2 - (e^{2x}-1)^2}{(e^{2x}+1)^2} \\ &= 1 - \frac{(e^{2x}-1)^2}{(e^{2x}+1)^2} = 1 - tanh^2(x) \end{aligned} \tag{5}$$

Problem 3

The joint probability distribution over target variables z_i is given by

$$p(\{z_i\}_{i=1}^N | \mathcal{D}, w) = \prod_{i=1}^N p(z_i | x_i, w) \tag{6}$$

The log likelihood with respect to z_i

$$\begin{aligned} l(\{z_i\}_{i=1}^N | \mathcal{D}, w) &= \sum_{i=1}^N log\left(p(z_i | x_i, w)\right) \\ &= \sum_{i=1}^N log\left(\sqrt{\frac{\beta}{(2\pi)^D}} e^{\frac{-\beta}{2}(z_i - y_i)^T (z_i - y_i)}\right) \\ &\propto \frac{-\beta}{2} \sum_{i=1}^N (z_i - y_i)^T (z_i - y_i) \end{aligned} \tag{7}$$

Therefore minimizing the negative log likelihood $nl(z) = -l(z)$ is equivalent to minimizing the sum of squared errors.

Problem 4

Done.