1. [50 points] Consider a linear model of the form

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i$$

together with a sum-of-squares error function of the form

$$error_D(\mathbf{w}) = \frac{1}{2} \sum_{t=1}^{N} \{y(\mathbf{x}_t, \mathbf{w}) - r_t\}^2$$

Now suppose that Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. Show that minimizing $error_D$ averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of some weight-decay regularization term (you should clearly identify the regularization term)

2. [50 points] Consider a 3-layer perceptron (1 hidden layer) for the classification task with K classes. Suppose the activation function for the output unit is the softmax function:

$$y_i^t = \frac{\exp[net_i]}{\sum_{j=1}^{K} \exp[net_j]}$$

where $net_i = \sum_{h=1}^{H} w_{hi} z_h^t + w_{0i}$ is the input to the i-th output unit ($1 \le i \le K$). The activation function for the hidden layer unit is the sigmoid function:

$$z_h^t = \frac{1}{1 + \exp(-\sum_{i=1}^{D} w_{ih} x_i + w_{0h})}$$

Derive the learning rule if the error function is the sum squared error, i.e.

$$Err(\mathcal{X}|\mathbf{w}) = \frac{1}{2} \sum_t \sum_i (r_i^t - y_i^t)^2$$