Machine Learning and Applications

Assignment 1

1. (2) Consider a two-layer network function of the form

$$y_k(x, \mathbf{w}) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} g\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right)$$

in which the hidden unit nonlinear activation functions $g(\cdot)$ are given by logistic sigmoid functions of the form

$$\sigma(a) = (1 + \exp(a))^{-1}$$

Show that there exists an equivalent network, which computes exactly the same function, but with hidden unit activation functions given by $\tanh(a)$ where the tanh function is defined by

$$\tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}}.$$

2. (2) Consider a binary classification problem in which the target values are $t \in \{0, 1\}$, with a network output $y(x, w)$ that represents the probability $p(t = 1|x)$, and suppose that there is a probability ϵ that the class label on a training data point has been incorrectly set.

Let us introduce a *likelihood function*. Let $\mathcal{D}$ be your data set and $\mathbf{w}$ be a set of parameters of your model. Then the likelihood function is defined as conditional probability

$$\mathcal{L}(\mathbf{w}) = p(\mathcal{D}|\mathbf{w}).$$

For example, let us consider linear regression problem in the following form:

$$y_i = \mathbf{w}^T\mathbf{x} + \xi_i, \quad \xi_i \sim \mathcal{N}(0, \sigma^2).$$

It means that $y_i \sim \mathcal{N}(\mathbf{w}^T\mathbf{x}_i, \sigma^2)$. Then the logarithm of likelihood function (log likelihood) is the following:

$$\log \mathcal{L}(\mathbf{w}) = \log \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}_N),$$

where $\mathbf{y} = (y_1, \ldots, y_N)$, $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T)^T$ and $\mathbf{I}_N$ is an identity matrix of size N.

Assuming independent and identically distributed data, write down the error function corresponding to the negative log likelihood. Verify that the error function

$$E(\mathbf{w}) = -\sum_{i=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\},$$

where y_n is a prediction of neural network, t_n is actual target value and N is the sample size, is obtained when $\epsilon = 0$. Note that this error function makes the model robust to incorrectly labeled data, in contrast to the usual error function.

3. (2) The outer product approximation to the Hessian matrix for a neural network using a sum-of-squares error function $E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y_n - t_n)^2$ is given by

$$\mathbf{H} \simeq \sum_{n=1}^{N} \nabla y_n (\nabla y_n)^T$$

Derive similar result in case of multiple outputs, when the error function is defined by

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} \|\mathbf{y}_n(\mathbf{x}_n, \mathbf{w}) - \mathbf{t}_n\|^2,$$

where $\mathbf{t}_n = (t_n^{(1)}, \ldots, t_n^{(K)})$, $\mathbf{y}(\mathbf{x}_n, \mathbf{w}) = (y^{(1)}(\mathbf{x}_n, \mathbf{w}), \ldots, y^{(K)}(\mathbf{x}_n, \mathbf{w}))$ and K is the number of output components.

4. (3) Consider a two-layer network of the form shown in Figure 1 with the addition of extra parameters corresponding to skip-layer connections that go directly from the inputs to the outputs. Consider the the sum-of-squares errors loss function and suppose that the output units have linear activation functions and hidden units have tanh activation function.

   Write down the equations for the derivatives of the error function with respect to these additional parameters.

5. (3) One way to encourage invariance of a model to a set of transformations is to expand the training set using transformed versions of the original input patterns.

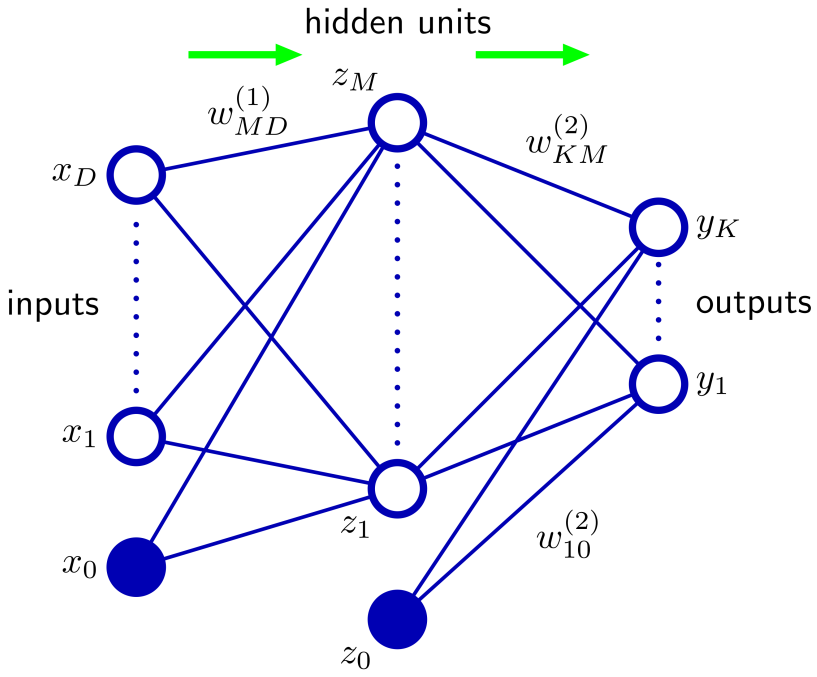

Figure 1: Neural Network

Consider the framework for training with transformed data in the special case in which the transformation consists simply of the addition of random noise $x \to x + \xi$ where ξ has a Gaussian distribution with zero mean and unit covariance.

The error function for untransformed inputs can be written (in the infinite data set limit) in the form

$$E = \frac{1}{2} \int \int (y(\mathbf{x}) - t)^2 p(t|\mathbf{x}) p(\mathbf{x}) \mathrm{d}\mathbf{x} \mathrm{d}t$$

If we now consider an infinite number of copies of each data point, each of which is perturbed by the transformation, then the error function can be written as

$$\widetilde{E} = \frac{1}{2} \int \int (y(x + \xi) - t)^2 p(t|\mathbf{x}) p(\mathbf{x}) p(\xi) \mathrm{d}\mathbf{x} \mathrm{d}t \mathrm{d}\xi$$

Using Taylor series expansion of $y(\mathbf{x} + \xi)$ show that

$$\widetilde{E} \simeq E + \lambda \Omega$$

where Ω is a regularizer which takes the form of Tikhonov regularizer

$$\Omega = \frac{1}{2} \int \|\nabla y(\mathbf{x})\|^2 p(\mathbf{x}) \mathrm{d}\mathbf{x}.$$