Structural Machine Learning Homework #1

Ting-Xuan, Hsu
Department of Applied Mathematics
National Chung Hsing University
Taichung, Taiwan

I. Derive for the Regression Problem

將3-layer multilayer perceptron 應用於迴歸問題。

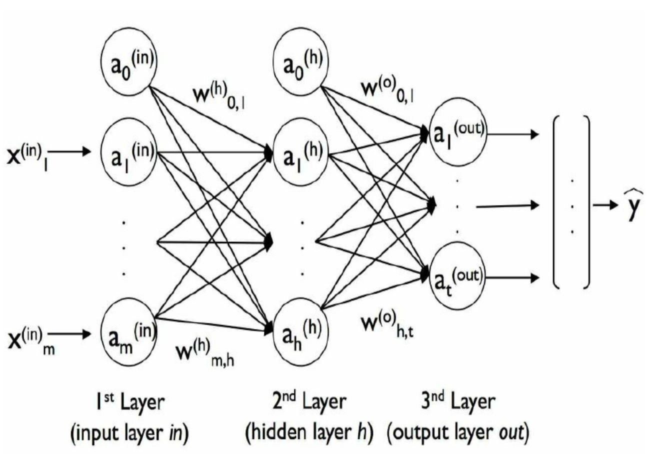

Fig. 1.

A. Forward Propagation

由於隱藏層中的每一個unit都與輸入層所有的unit相連接，可以由(1)、(2)兩式計算出隱藏層中的第一個activation unit $a_1^{(h)}$

$$z_1^{(h)} = \begin{bmatrix} a_1^{(in)} & a_2^{(in)} & \dots & a_m^{(in)} \end{bmatrix}^T \begin{bmatrix} w_{0,1}^{(h)} & w_{1,1}^{(h)} & \dots & w_{m,1}^{(h)} \end{bmatrix} \tag{1}$$

$$a_1^{(h)} = \phi\left(z_1^{(h)}\right) \tag{2}$$

其中，$a_i^{(in)}$為輸入層第i個unit，$w_{i,j}^{(h)}$為輸入層第i個unit連接到隱藏層第j個unit的權重，$\phi(\cdot)$為activation function(此處為sigmoid)，以向量表示為

$$\begin{aligned} \vec{z}^{\,(h)} &= \begin{bmatrix} z_1^{(h)} & z_2^{(h)} & \dots & z_n^{(h)} \end{bmatrix}^T \\ &= \begin{bmatrix} \vec{w_1}^{(h)} & \vec{w_2}^{(h)} & \dots & \vec{w_n}^{(h)} \end{bmatrix}^T \vec{a}^{(in)} \\ &= \left(\mathbf{W}^{(h)}\right)^T \vec{a}^{(in)} \end{aligned} \tag{3}$$

$$\vec{a}^{(h)} = \phi\left(\vec{x}^{(h)}\right) \tag{4}$$

以矩陣形式分別寫出隱藏層(h)及輸出層(out)的結果

$$\mathbf{Z}^{(h)} = \left(\mathbf{w}^{(h)}\right)^T \mathbf{A}^{(in)} \tag{5}$$

$$\mathbf{A}^{(h)} = \phi\left(\mathbf{Z}^{(h)}\right) \tag{6}$$

$$\mathbf{Z}^{(out)} = \left(\mathbf{W}^{(out)}\right)^T \mathbf{A}^{(h)} \tag{7}$$

$$\mathbf{A}^{(out)} = \phi\left(\mathbf{Z}^{(out)}\right) \tag{8}$$

B. Cost Function

$$\begin{aligned} J(W) &= \sum_i \left\| \hat{y}^{(i)} - y^{(i)} \right\|^2 \\ &= \sum_i \left\| \vec{a}^{(i)} - \vec{y}^{(i)} \right\| \\ &= Tr\left(\left(\mathbf{A}^{(out)} - \mathbf{Y}\right)^T \left(\mathbf{A}^{(out)} - \mathbf{Y}\right)\right) \end{aligned} \tag{9}$$

C. Backward Propagation

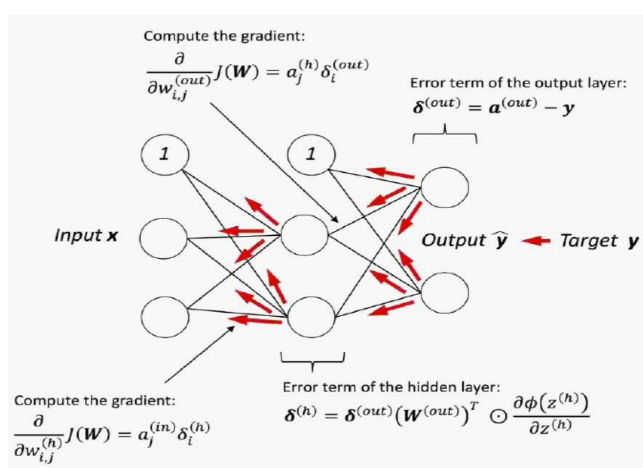

Fig. 2.

The gradient used to update $\mathbf{W}^{(out)}$ can be calculate as follows:

$$
\begin{aligned}
\frac{\partial}{\partial \mathbf{W}^{(out)}_{i,j}} J(\mathbf{W}) &= \frac{\partial [Tr((\mathbf{A}^{(out)} - \mathbf{Y})^T(\mathbf{A}^{(out)} - \mathbf{Y}))]}{\partial \mathbf{W}^{(out)}_{i,j}} \\
&= \frac{\partial [Tr((\mathbf{A}^{(out)} - \mathbf{Y})^T(\mathbf{A}^{(out)} - \mathbf{Y}))]}{\partial \mathbf{A}^{(out)} - \mathbf{Y}} \\
&\qquad \frac{\partial(\mathbf{A}^{(out)} - Y)}{\partial \mathbf{W}^{(out)}_{i,j}} \\
&= 2(\mathbf{A}^{(out)} - \mathbf{Y}) \frac{\partial(\mathbf{A}^{(out)} - \mathbf{Y})}{\partial \mathbf{A}^{(out)}} \frac{\partial \mathbf{A}^{(out)}}{\partial \mathbf{W}^{(out)}_{i,j}} \\
&= 2(\mathbf{A}^{(out)} - \mathbf{Y}) I \frac{\partial((\mathbf{W}^{(out)})^T)\mathbf{A}^{(h)})}{\partial \mathbf{W}^{(out)}_{i,j}} \\
&= 2\mathbf{A}^{(h)}_j \delta^{(out)} \\
\delta^{(out)}_i &= \mathbf{A}^{(out)} - \mathbf{Y}
\end{aligned}
\tag{10}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \mathbf{W}^{(h)}_{i,j}} J(\mathbf{W}) &= \frac{\partial [Tr((\mathbf{A}^{(out)} - \mathbf{Y})^T(\mathbf{A}^{(out)} - \mathbf{Y}))]}{\partial(\mathbf{A}^{(out)} - \mathbf{Y})} \frac{\partial(\mathbf{A}^{(out)} - \mathbf{Y})}{\partial \mathbf{W}^{(h)}_{i,j}} \\
&= 2(\mathbf{A}^{(out)} - \mathbf{Y}) \frac{\partial(\mathbf{A}^{(out)} - \mathbf{Y})}{\partial \mathbf{A}^{(out)}} \frac{\partial \mathbf{A}^{(out)}}{\partial \mathbf{W}^{(h)}_{i,j}} \\
&= 2(\mathbf{A}^{(out)} - \mathbf{Y}) I \frac{\partial((\mathbf{W}^{(out)})^T)\mathbf{A}^{(h)}}{\partial \mathbf{A}^{(h)}} \frac{\partial \mathbf{A}^{(h)}}{\partial \mathbf{W}^{(h)}_{i,j}} \\
&= 2(\mathbf{A}^{(out)} - \mathbf{Y}) I \frac{\partial((\mathbf{W}^{(out)})^T)\mathbf{A}^{(h)}}{\partial \mathbf{A}^{(h)}} \\
&\qquad \frac{\partial \phi(\mathbf{Z}^{(h)})}{\partial \mathbf{Z}^{(h)}} \frac{\mathbf{Z}^{(h)}}{\partial \mathbf{W}^{(h)}_{i,j}} \\
&= 2(\mathbf{A}^{(out)} - \mathbf{Y}) I \frac{\partial((\mathbf{W}^{(out)})^T)\mathbf{A}^{(h)}}{\partial \mathbf{A}^{(h)}} \\
&\qquad [\phi(\mathbf{Z}^{(h)}) \odot (\mathbf{C} - \phi(\mathbf{Z}^{(h)})] \frac{\partial[(\mathbf{W}^{(h)})^T \mathbf{A}^{(in)}]}{\partial \mathbf{W}^{(h)}_{i,j}} \\
&= 2\mathbf{A}^{(in)}_j \delta^{(h)}_i \\
\delta^{(h)}_i &= \mathbf{W}^{(out)} \delta^{(out)} \odot \frac{\partial \phi(\mathbf{Z}^{(h)})}{\partial(\mathbf{Z}^{(h)})}
\end{aligned}
\tag{11}
$$

II. Prepare Data and Preprocessing

A. Prepare Data

將Boston Housing Data 用於迴歸問題以預測房價。

資料筆數: 506
屬性個數: 14

1. CRIM ：per capita crime rate by town
2. ZN ：proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS ：proportion of non-retail business acres per town
4. CHAS ：Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX ：nitric oxides concentration (parts per 10 million)
6. RM ：average number of rooms per dwelling
7. AGE ：proportion of owner-occupied units built prior to 1940
8. DIS ：weighted distances to five Boston employment centres
9. RAD ：index of accessibility to radial highways
10. TAX ：full-value property-tax rate per 10,000
11. PTRATIO ：pupil-teacher ratio by town
12. B ：$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
13. LSTAT ：
14. MEDV ：Median value of owner-occupied homes in 1000's

B. Data preprocessing

將特徵及房價分別做Normalization.

Random 70% data are used in the training phase.

III. Implemention and Result

A. Result

n_hidden=50, epochs=1500, eta=0.002

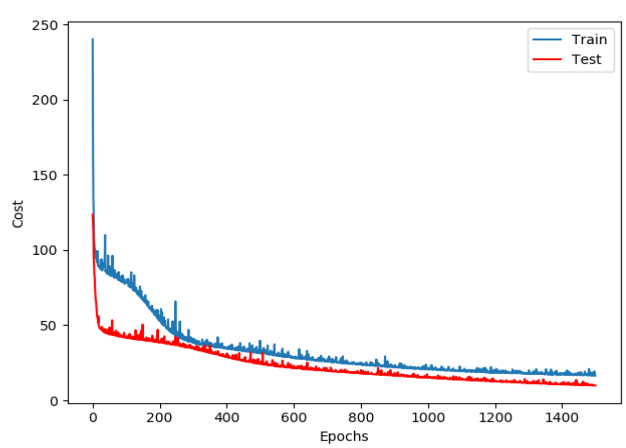

Fig. 3.