Machine Learning Homework1

張格恩
應用數學系
國立中興大學

I. Derivation for the regression problem

這裡我是用 3 層的 Multilayer Perceptron 來推導 forward 和 backward schemes for the Regression problem。3 層的 Multilayer Perceptron 包含 1 層 input layer，1 層 hidden layer 以及 1 層 output layer。

A. Forward schemes

我是以訓練數據 (training data) 中 n 筆樣本 (sample) 來推導 forward schemes。

1.Net input of the hidden layer:

$$Z^{(h)} = (W^{(h)})^T A^{(in)}$$

$A^{(in)} \in \mathbb{R}^{k\times n}$，$W^{(h)} \in \mathbb{R}^{k\times h}$，$Z^{(h)} \in \mathbb{R}^{h\times n}$，$n$ 是樣本的數量，k 是 feature 的數量，h 是 output of the hidden layer 的維度。

2.Activation of the hidden layer:

$$A^{(h)} = \phi(Z^{(h)})$$

$A^{(h)} \in \mathbb{R}^{h\times n}$，這裡我的 activation function(ϕ) 是用 sigmod function。

3.Net input of the output layer:

$$Z^{(out)} = (W^{(out)})^T A^{(h)}$$

$W^{(out)} \in \mathbb{R}^{h\times 1}$，$Z^{(out)} \in \mathbb{R}^{1\times n}$

4.Activation of the output layer:

$$A^{(out)} = Z^{(out)}$$

$A^{(out)} \in \mathbb{R}^{1\times n}$，這裡我的 activation function 是用 identity function。

B. Cost function

這裡 cost function 是用 sum of square error，其表示如下:

$$\begin{aligned} J(w) &= \sum_i \|\hat{y}_i - y_i\|^2 \\ &= Tr((A^{(out)} - Y)^T (A^{(out)} - Y)) \end{aligned}$$

$\hat{y_i}$ 是模型對第 i 筆資料的預測，y_i 是第 i 筆資料的真實標籤，Y 是 n 筆樣本的真實標籤。

C. Backward schemes

1.The gradient used to update $W^{(out)}$

$$\begin{aligned} \frac{\partial J(w)}{\partial W_j^{(out)}} &= \frac{\partial J(w)}{\partial (A^{(out)} - Y)} \frac{\partial (A^{(out)} - Y)}{\partial W_j^{(out)}} \\ &= 2(A^{(out)} - Y) \frac{\partial (A^{(out)} - Y)}{\partial A^{(out)}} \frac{\partial A^{(out)}}{\partial W_j^{(out)}} \\ &= 2(A^{(out)} - Y) \frac{\partial}{\partial W_j^{(out)}} (W^{(out)})^T A^{(h)} \end{aligned}$$

最後，我們可以獲得:

$$\begin{aligned} \delta^{(out)} &= A^{(out)} - Y \\ \frac{\partial}{\partial W_j^{(out)}} J(w) &= 2a_j^{(h)} \delta^{(out)} \end{aligned}$$

這裡 $a_j^{(h)}$ 代表 $A^{(h)}$ 的第 j 列 (row)。

2.The gradient used to update $W^{(h)}$

$$\begin{aligned} \frac{\partial J(w)}{\partial W_{j,k}^{(h)}} &= \frac{\partial J(w)}{\partial (A^{(out)} - Y)} \frac{\partial (A^{(out)} - Y)}{\partial W_{j,k}^{(h)}} \\ &= 2(A^{(out)} - Y) \frac{\partial (A^{(out)} - Y)}{\partial A^{(out)}} \frac{\partial A^{(out)}}{\partial W_{j,k}^{(h)}} \\ &= 2(A^{(out)} - Y) \frac{\partial (W^{(out)})^T A^{(h)}}{\partial A^{(h)}} \frac{\partial A^{(h)}}{\partial W_{j,k}^{(h)}} \end{aligned}$$

這裡我們將上述的 $\frac{\partial (W^{(out)})^T A^{(h)}}{\partial A^{(h)}} \frac{\partial A^{(h)}}{\partial W_{j,k}^{(h)}}$ 變成對這個數值為例 ($A_{j,i}^{(h)}$)，而繼續偏微分。

$$\begin{aligned} \frac{\partial J(w)}{\partial W_{j,k}^{(h)}} &= 2(A^{(out)} - Y) \frac{\partial (W^{(out)})^T A^{(h)}}{\partial A^{(h)}} \frac{\partial A^{(h)}}{\partial W_{j,k}^{(h)}} \\ &= 2(A^{(out)} - Y) \frac{\partial (W^{(out)})^T A^{(h)}}{\partial A_{j,i}^{(h)}} \frac{\partial A_{j,i}^{(h)}}{\partial W_{j,k}^{(h)}} \\ &= 2(A^{(out)} - Y) W_j^{(out)} \frac{\partial \phi(Z_{j,i}^{(h)})}{\partial Z_{j,i}^{(h)}} \frac{\partial Z_{j,i}^{(h)}}{\partial W_{j,k}^{(h)}} \\ &= 2(A^{(out)} - Y) W_j^{(out)} \phi(Z_{j,i}^{(h)})[1 - \phi(Z_{j,i}^{(h)})] \frac{\partial [W_j^{(h)} A_i^{(in)}]}{\partial W_{j,k}^{(h)}} \\ &= 2(A^{(out)} - Y) W_j^{(out)} \phi(Z_{j,i}^{(h)})[1 - \phi(Z_{j,i}^{(h)})] A_{k,i}^{(in)} \end{aligned}$$

最後，我們對 $A^{(h)}$ 的所有數值進行偏微分後，可以獲得:

$$\delta^{(h)} = W^{(out)}\delta^{out} \odot \frac{\phi(Z^{(h)})}{\partial Z^{(h)}}$$
$$\frac{\partial}{\partial W^{(h)}_{j,k}} J(w) = 2a^{(in)}_k \delta^{(h)}_j$$

這裡 $a^{(in)}_k$ 代表每筆樣本的第 k 項。

II. Preparation and Preprocessing for the Data

A. Preparation for the data

這裡的數據我們使用 Housing dataset。
1.Feature(x) 我們用:
(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 oxide 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 centers
(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 [people of African American descent] by town
(13)LSTAT: Percentage of lower status of the population

2.Label(y) 則是:
MEDV: Median value of owner-occupied homes in $1000s

B. preprocessing for the data

數據的樣本數: 506 筆。我們的 Training data/Test data 分別是: 354 筆/152 筆 (70% / 30%)。
我們會把 Feature(x) 根據各項特徵做 Normalization，同時也會把 Label(y) 做 Normalization。

III. Implement the 3-layer MLP for the Regression Problem

A. Model

這裡我以 1 筆樣本為例，units of hidden layer 我設計為 50 個。下圖為我們的 3-layer MLP for regression 的模型圖。

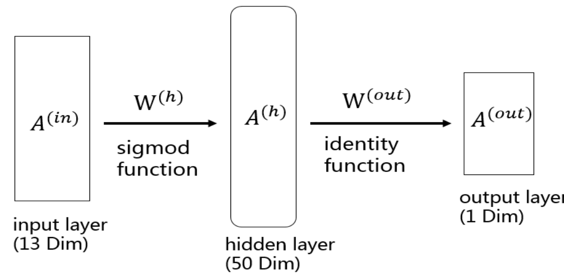

Fig. 1: 3-layer MLP for regression

B. Experiment1

我們這裡是用 sum of square error(SSE) 來當成更新權重的 cost function。
參數設定是用 hidden units: 50 ， epochs: 2000 ， learning rate: 0.001 ， batch size: 3 。
1.sum of square error(SSE) for training:

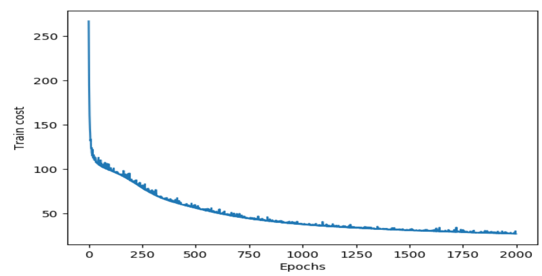

Fig. 2: SSE for training

2.sum of square error(SSE) for testing:

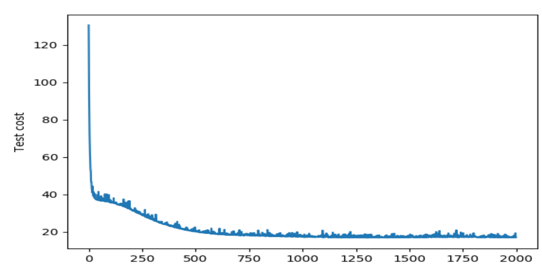

Fig. 3: SSE for testing

3.mean square error for training/testing:
這裡是將上述的 SSE of training 和 SSE of testing 各別取平均，並且我們發現訓練過程在 1000 ∼ 1250 epochs 之間，MSE of training 和 MSE of testing 會交集。

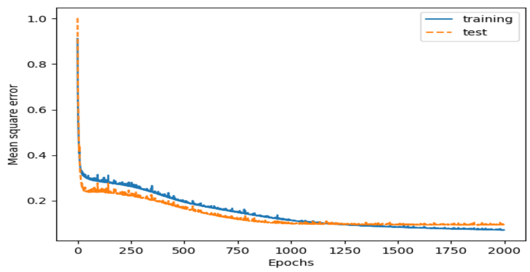

Fig. 4: MSE for training and testing

C. *Experiment2*

我們這裡改用 mean square error(MSE) 來當成更新權重的 cost function。
參數設定是用 hidden units: 50 ， epochs: 3000 ， learning rate: 0.001 ， batch size: 3 。
我們發現訓練過程大概在 3000 epochs 時，MSE of training 和 MSE of testing 會交集。

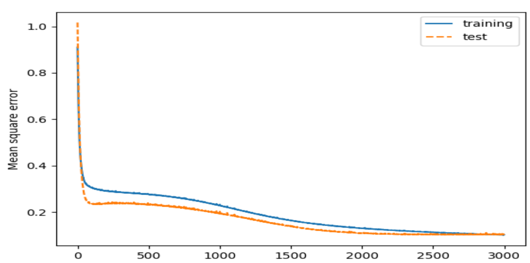

Fig. 5: MSE for training and testing

D. *Experiment3*

這裡我們想嘗試用上述的模型，但對資料做不同的前處理 (preprocessing)，以此確認前處理 (preprocessing) 對實驗的影響。
我們這裡用 mean square error(MSE) 來當成更新權重的 cost function。
參數設定是用 hidden units: 50 ， epochs: 2000 ， learning rate: 0.001 ， batch size: 3 。

1.normalization for feature and label:
我們將 feature 和 label 都做 normalization。其 MSE for training and testing 如下。

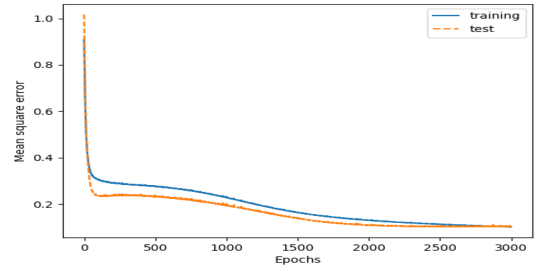

Fig. 6: normalization for feature and label

2.normalization for label:
我們只將 label 做 normalization。其 MSE for training and testing 如下。

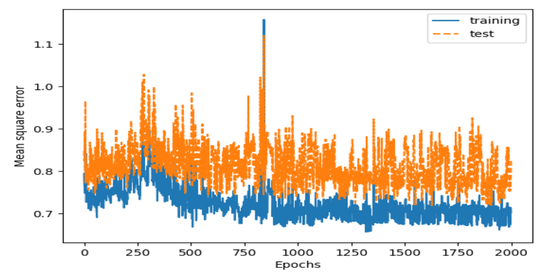

Fig. 7: normalization for label

3.normalization for feature:
我們只將 feature 做 normalization。其 MSE for training and testing 如下。

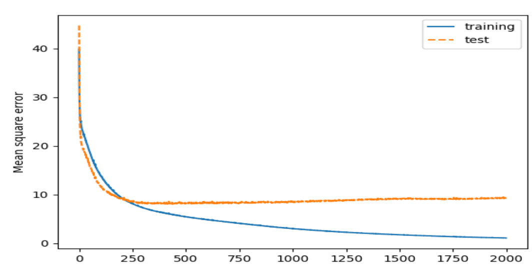

Fig. 8: normalization for feature

4.not normalization:
我們不做任何的前處理 (preprocessing)。其 MSE for training and testing 如下。

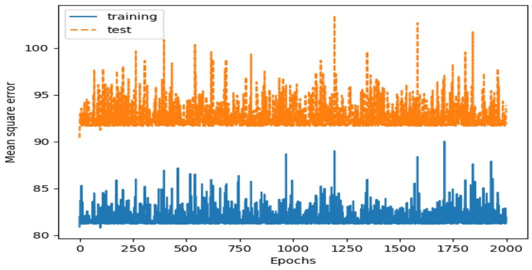

Fig. 9: not normalization