Machine Learning Homework 1

唐永承
4105053128
國立中興大學應用數學系

I. Derive The Forward And Backward Schemes

三層的 Multilayer Perceptron 包含一層輸入層(Input Layer)，一層隱藏(Hidden Layer)以及一層輸出層(Output Layer)。

A. *Forward Schemes*

1. Net input of the hidden layer

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

$\mathrm{A}^{(in)} \in \mathbb{R}^{\mathrm{k}\times\mathrm{n}}$，$\mathrm{W}^{(\mathrm{h})} \in \mathbb{R}^{\mathrm{k}\times\mathrm{h}}$，$Z^{(h)} \in \mathbb{R}^{h\times n}$，n 為樣本數量，k 為特徵數量，h 為 hidden layer 的 output 維度。

2. Activation function of the hidden layer

$$\mathrm{A}^{(\mathrm{h})} = \emptyset\left(\mathrm{Z}^{(\mathrm{h})}\right)$$

$\mathrm{A}^{(h)} \in \mathbb{R}^{\mathrm{h}\times\mathrm{n}}$，這裡的激勵函數(Activation function)使用 sigmoid function。

3. Net input of the output layer:

$$\mathrm{Z}^{(\mathrm{out})} = \left(\mathrm{W}^{(\mathrm{out})}\right)^{\mathrm{T}}$$

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

4. Activation of the output layer:

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

$\mathrm{A}^{(\mathrm{out})} \in \mathbb{R}^{1\times\mathrm{n}}$，輸出層則沒有激勵函數。

B. *Cost Function*

在此使用平方誤差總合 (Sum of Squared Errors)作為 Cost Function。

$$\mathrm{J}_{(\mathrm{W})} = \sum_{\mathrm{i}} \|\hat{y}_i - y_i\|^2$$
$$= \mathrm{Tr}\left(\left(\mathrm{A}^{(\mathrm{out})} - \mathrm{Y}\right)^{\mathrm{T}}\left(\mathrm{A}^{(\mathrm{out})} - \mathrm{Y}\right)\right) i$$

$\hat{y}_i$ 為第 i 筆的資料預測，y_i 是第 i 筆的真實標籤，Y 為所有資料的真實標籤。

C. *Backward Schemes*

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

$$\frac{\partial \mathrm{J(w)}}{\partial \mathrm{W}_{\mathrm{j}}^{(\mathrm{out})}} = \frac{\partial \mathrm{J(w)}}{\partial(\mathrm{A}^{(\mathrm{out})} - Y)}\frac{\partial(\mathrm{A}^{(\mathrm{out})} - Y)}{\partial \mathrm{W}_{\mathrm{j}}^{(\mathrm{out})}}$$
$$= 2(\mathrm{A}^{(\mathrm{out})} - Y)\frac{\partial(\mathrm{A}^{(\mathrm{out})} - Y)}{\partial \mathrm{A}^{(\mathrm{out})}}\frac{\partial \mathrm{A}^{(\mathrm{out})}}{\partial \mathrm{W}_{\mathrm{j}}^{(\mathrm{out})}}$$
$$= 2(\mathrm{A}^{(\mathrm{out})} - Y)\frac{\partial}{\partial \mathrm{W}_{\mathrm{j}}^{(\mathrm{out})}}\left(\mathrm{W}^{(\mathrm{out})}\right)^{\mathrm{T}}\mathrm{A}^{(\mathrm{h})}$$

可得到：

$$\delta^{(\mathrm{out})} = \mathrm{A}^{(\mathrm{out})} - \mathrm{Y}$$
$$\frac{\partial}{\partial \mathrm{W}_{\mathrm{j}}^{(\mathrm{out})}} J(w) = 2a_j^{(h)}\delta^{(out)}$$

此處$a_j^{(h)}$表示$\mathrm{A}^{(\mathrm{h})}$的第 j 列。

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

$$\frac{\partial \mathrm{J(w)}}{\partial \mathrm{W}_{\mathrm{j,k}}^{(\mathrm{h})}} = \frac{\partial \mathrm{J(w)}}{\partial(\mathrm{A}^{(\mathrm{out})} - Y)}\frac{\partial(\mathrm{A}^{(\mathrm{out})} - Y)}{\partial \mathrm{W}_{\mathrm{j,k}}^{(\mathrm{h})}}$$
$$= 2(\mathrm{A}^{(\mathrm{out})} - Y)\frac{\partial(\mathrm{A}^{(\mathrm{out})} - Y)}{\partial \mathrm{A}^{(\mathrm{out})}}\frac{\partial \mathrm{A}^{(\mathrm{out})}}{\partial \mathrm{W}_{\mathrm{j,k}}^{(\mathrm{h})}}$$
$$= 2(\mathrm{A}^{(\mathrm{out})} - Y)\frac{\partial\left(\mathrm{W}^{(\mathrm{out})}\right)^{\mathrm{T}}\mathrm{A}^{(\mathrm{h})}}{\partial \mathrm{A}^{(\mathrm{h})}}\frac{\partial \mathrm{A}^{(\mathrm{h})}}{\partial \mathrm{W}_{\mathrm{j,k}}^{(\mathrm{h})}}$$

再將上述$\frac{\partial\left(\mathrm{W}^{(\mathrm{out})}\right)^{\mathrm{T}}\mathrm{A}^{(\mathrm{h})}}{\partial \mathrm{A}^{(\mathrm{h})}}\frac{\partial \mathrm{A}^{(\mathrm{h})}}{\partial \mathrm{W}_{\mathrm{j,k}}^{(\mathrm{h})}}$改成以$\mathrm{A}_{\mathrm{j,i}}^{(\mathrm{h})}$繼續偏微分。

$$\frac{\partial \mathrm{J(w)}}{\partial \mathrm{W}_{\mathrm{j,k}}^{(\mathrm{h})}} = 2(\mathrm{A}^{(\mathrm{out})} - Y)\frac{\partial\left(\mathrm{W}^{(\mathrm{out})}\right)^{\mathrm{T}}\mathrm{A}^{(\mathrm{h})}}{\partial \mathrm{A}^{(\mathrm{h})}}\frac{\partial \mathrm{A}^{(\mathrm{h})}}{\partial \mathrm{W}_{\mathrm{j,k}}^{(\mathrm{h})}}$$
$$= 2(\mathrm{A}^{(\mathrm{out})} - Y)\frac{\partial\left(\mathrm{W}^{(\mathrm{out})}\right)^{\mathrm{T}}\mathrm{A}^{(\mathrm{h})}}{\partial \mathrm{A}_{\mathrm{j,i}}^{(\mathrm{h})}}\frac{\partial \mathrm{A}^{(\mathrm{h})}}{\partial \mathrm{W}_{\mathrm{j,k}}^{(\mathrm{h})}}$$
$$= 2(\mathrm{A}^{(\mathrm{out})} - Y)\frac{\partial(\mathrm{W}_{\mathrm{j}}^{(\mathrm{out})})\emptyset\left(\mathrm{Z}_{\mathrm{j,i}}^{(\mathrm{h})}\right)}{\partial \mathrm{Z}_{\mathrm{j,i}}^{(\mathrm{h})}}\frac{\partial \mathrm{Z}_{\mathrm{j,i}}^{(\mathrm{h})}}{\partial \mathrm{W}_{\mathrm{j,k}}^{(\mathrm{h})}}$$
$$= 2(\mathrm{A}^{(\mathrm{out})} - Y)\mathrm{W}_{\mathrm{j}}^{(\mathrm{out})}\emptyset\left(\mathrm{Z}_{\mathrm{j,i}}^{(\mathrm{h})}\right)[1 - \emptyset\left(\mathrm{Z}_{\mathrm{j,i}}^{(\mathrm{h})}\right)]\frac{\partial\left[W_j^{(h)}A_i^{(in)}\right]}{\partial W_{j,k}^{(h)}}$$
$$= 2(\mathrm{A}^{(\mathrm{out})} - Y)\mathrm{W}_{\mathrm{j}}^{(\mathrm{out})}\emptyset\left(\mathrm{Z}_{\mathrm{j,i}}^{(\mathrm{h})}\right)[1 - \emptyset\left(\mathrm{Z}_{\mathrm{j,i}}^{(\mathrm{h})}\right)A_{k,i}^{(in)}]$$

再對$\mathrm{A}_{\mathrm{j,i}}^{(\mathrm{h})}$的所有數值做偏微分後，可得到：

$$\delta^{(\mathrm{h})} = \mathrm{W}^{(0\cup t)\delta\mathrm{out}} \odot \frac{\emptyset \mathrm{Z}^{(\mathrm{h})}}{\partial \mathrm{Z}^{(\mathrm{h})}}$$
$$\frac{\partial}{\partial \mathrm{W}_{\mathrm{j,k}}^{(\mathrm{h})}} J(w) = 2a_k^{(in)}\delta_j^{(h)}$$

$a_k^{(in)}$代表每筆資料第 k 項。

II. Data And Data Preprocessing

A. *Data*

這次作業使用 Housing data set ，此訓練資料集，共有 13 個資料特徵 (features)，1 個標籤 (label)，總共 506 筆資料。

特徵：

1. CRIM: 城鎮人均犯罪率
2. ZN: 佔地面積超過 25,000 平方呎的住宅用地比例。
3. INDUS: 每個城鎮的非零售業務面積比例
4. CHAS: 查爾斯河虛擬變量，如果是大片土地則為 1，否則為 0
5. NOX： 一氧化氮濃度（10ppm）
6. RM: 每棟住宅的平均房間數量
7. AGE: 1940 年以前的業主單位比例
8. DIS: 到波士頓五個就業中心的加權距離
9. RAD: 高速公路的可達性指數
10. TAX: 每 10,000 美元的全價物業稅率
11. PTRATIO: 城鎮的師生比例
12. B: $1000(Bk - 0.63)^2$，其中 Bk 是非洲裔美國人後裔的人數比例
13. LSTAT: 低端人口比例

標籤：

1. MEDV: 自住房屋的中位數價值(1000 美金為單位)

B. *Data Preprocessing*

我先將資料做前處理，先算出每欄資料的平均數以及標準差，再替資料做標準化。

$$Z = \frac{X - \mu}{\sigma}$$

Z 為標準化後數值，X 為資料，μ為資料平均數，σ為資料標準差。

完成標準化後，會將資料進行切割，其中 80%作為訓練集，剩餘的 20%做為測試集，因此訓練集會有 404 筆資料，測試集則有 102 筆資料

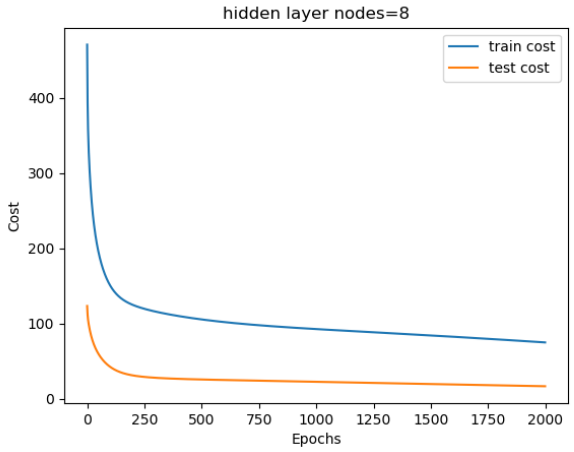
III. Implement The 3-Layer MLP

這次實作部分我會做四種 3-Layer MLP 和三種 4-Layer MLP，3-Layer MLP 分別為隱藏層有 8、15、30 和 50 個節點(hidden nodes)， 4-Layer MLP 中兩層隱藏層節點分別為 30-10 個、50-20 個以及 100-50 個。

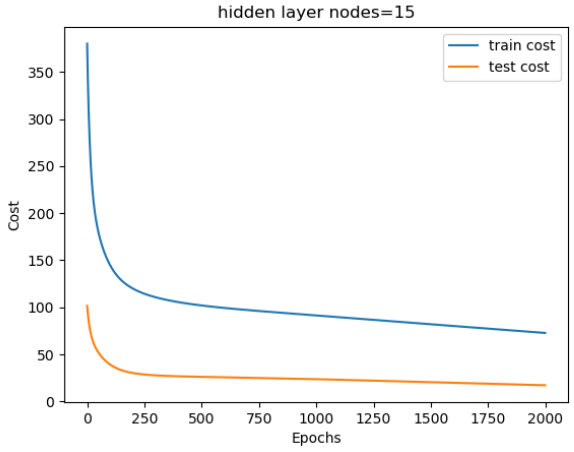
這幾個 MLP 的其餘參數皆相同， epoch 次數設定為 2000，學習率(learning rate) 設定為 0.0001，激勵函數(activation function) 選用 sigmoid function 。

以下實驗結果的圖片中，藍色線為 train cost ，橘色線為 test cost，在這裡會造成 train cost 的值 大於 test cost 的原因是因為使用平方誤差總合 (Sum of Squared Errors, SSE) 作為 cost function，而這使用的訓練資料(404 筆)比測試資料多(102 筆)，因此總和上 train cost 較大。

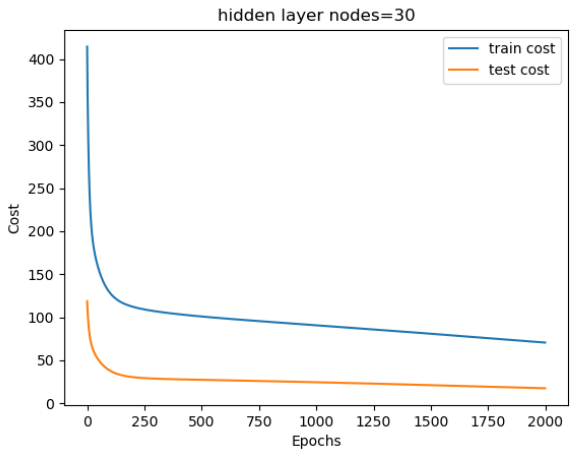
A. *8 Hidden Layer Nodes*



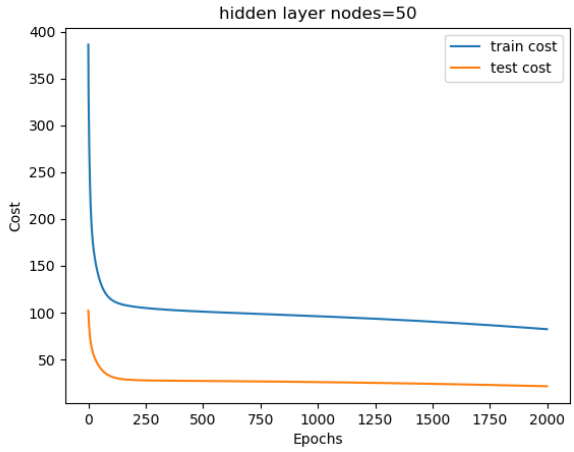
B. *15 Hidden Layer Nodes*



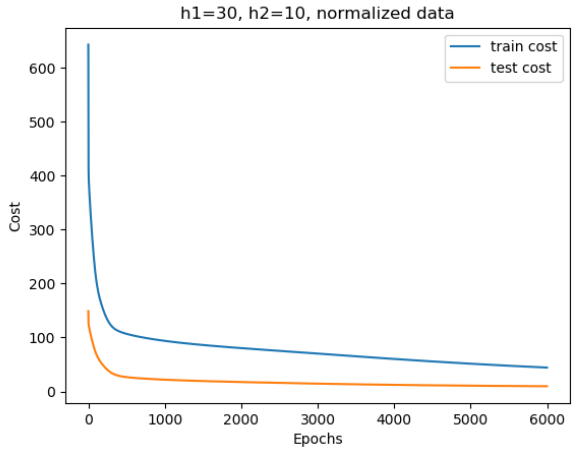
C. *30 Hidden Layer Nodes*



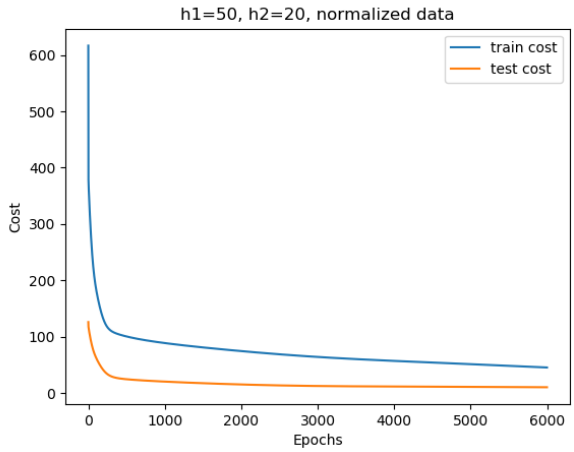
D. 50 Hidden Layer Nodes



E. Hidden Layer 1 30 Nodes, Hidden Layer 2 10 Nodes



F. Hidden Layer 1 50 Nodes, Hidden Layer 2 20 Nodes



G. Hidden Layer 1 100 Nodes, Hidden Layer 2 50 Nodes,