Homework #1

7107018016 統計碩一翁婉庭

Data: 我們取 housing data，X 共有 13 個變數，y 共有 506 筆資料，取 MEDV(房價) 作為預測目標，透過訓練去找到最適合的預測結果。

I. 題目

1.Derive the forward and backward schemes for the Regression Problem, where the cost is $\sum_{i=1}^{n}\left\|\hat{y}^{(i)}-y^{(i)}\right\|^2$
2.Prepare the house data set, and then do the preprocessing to the data
3.Implement the 3-layer MLP (including the cost, forward and backward schemes) for the regression problem.

II. 變數設定

我們將 MLP 架構中，先將 X 標準化，*n_hidden*(隱藏層數量)=30 層, epochs(迭代次數)=100 次，X: 輸入的變數，W: 權重，b:bias,Z: 輸出值，a: 經過 (sigmoid) 激活函數後的輸出值，$\hat{y}^{(i)}$: 預測值,$y^{(i)}$: 真實值

III. 理論架構

我們根據資料去做 3 層 MLP 先將資料分割為 test 及 training，*test_size*=0.3,*training_size*=0.7, 先將 X 標準化，接著進行 forward，在隱藏層，透過 sigmoid 函數去活化在激活我們的值，將值轉成範圍在 0 和 1 之間，同樣在輸出層，也是做一次的 sigmoid 函數去轉換我們的值，最後得到的 output 值 $\hat{y}^{(i)}$ 經過 cost function, 去降低我們預測的損失，經過 100 次的重複 forward 和 backward 來回訓練，得到最後的預測值。

IV. MLP 流程

A. 前向傳遞 *(Forward propagation)*

- 輸入層到隱藏層

$$Z^{(h)}=(X\cdot W^{(h)})+b^{(h)}$$

- 激活隱藏層

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

  a_h 將 Z_h 的值用 sigmoid function 轉成範圍從 0 到 1 之間。
- 隱藏層到輸出層

$$Z^{(out)}=(a^{(h)}\cdot W^{(out)})+b^{(out)}$$

- 激活輸出層

$$a^{(out)}=\phi(Z^{(out)})$$

  a_{out} 將 Z_{out} 的值用 sigmoid function 轉成範圍從 0 到 1 之間。
- cost function

$$\sum_{i=1}^{n}\left\|a^{(out)}-y^{(i)}\right\|^2$$

B. 反向傳遞 *(Backward propagation)*

反向傳遞的目的就是利用最後的目標函數 (cost function) 來進行參數的更新，在這裡利用 SSE 當作目標函數。如果誤差值越大，代表參數學得不好，所以需要繼續學習，直到參數或是誤差值收斂。

* delta gradient
1.)$\delta^{(out)}=\frac{\partial J(W)}{\partial a^{(out)}}\cdot\frac{\partial a^{(out)}}{\partial Z^{(out)}}=\sigma'(Z^{(out)})$
2.)$\delta^{(h)}=\frac{\partial J(W)}{\partial a^{(h)}}\cdot\frac{\partial a^{(h)}}{\partial Z^{(h)}}=\sigma'(Z^{(h)})$

* W gradient
1.)$\frac{\partial}{\partial W^{(out)}}J(W)=\frac{\partial J(W)}{\partial a^{(out)}}\cdot\frac{\partial a^{(out)}}{\partial Z^{(out)}}\cdot\frac{\partial Z^{(out)}}{\partial W^{(out)}}$

$$=\delta^{(out)}\cdot a^{(h)}$$

$$\frac{\partial J(W)}{\partial a^{(out)}}=\frac{\partial}{\partial a^{(out)}}\sum_{i=1}^{n}\left\|a^{(out)}-y^{(i)}\right\|^2=2\sum_{i=1}^{n}(a^{(out)}-y^{(i)})$$

$$\frac{\partial a^{(out)}}{\partial Z^{(out)}}=\partial(\frac{1}{1+e^{-a^{(out)}}})=\frac{\partial}{\partial W_{out}}\sum_{i=1}^{n}(a^{(out)}-y^{(i)})^2$$

$$=\sigma'(Z^{(out)})$$

$$\frac{\partial Z^{(out)}}{\partial W^{(out)}}=\frac{\partial(a^{(h)}\cdot W^{(out)}+b^{(out)})}{\partial W^{(out)}}=a^{(h)}$$

2.) $\frac{\partial}{\partial W^{(h)}}J(W)=\frac{\partial J(W)}{\partial a^{(h)}}\cdot\frac{\partial a^{(h)}}{\partial Z^{(h)}}\cdot\frac{\partial Z^{(h)}}{\partial W^{(h)}}=\delta^{(h)}\cdot X$

$$\frac{\partial J(W)}{\partial a^{(h)}}=\frac{\partial J(w)}{\partial a^{(out)}}\cdot\frac{\partial a^{(out)}}{\partial z^{(out)}}\cdot\frac{\partial z^{(out)}}{\partial a^{(h)}}$$

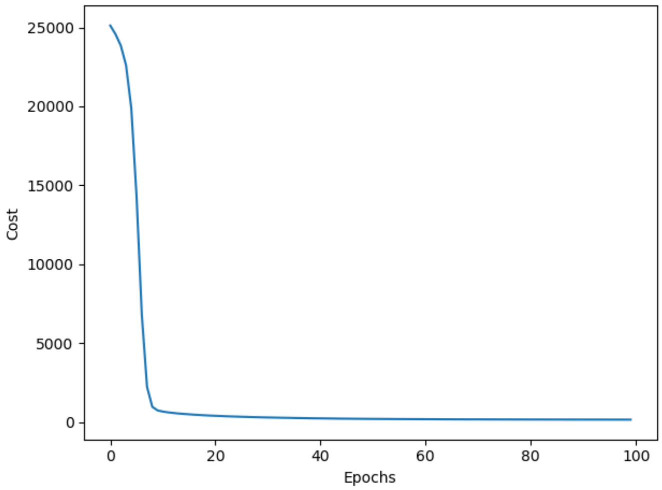
$$=2\sum(a^{(out)}-y^{(i)})\cdot W^{(out)}$$

$$\frac{\partial a^{(h)}}{\partial Z^{(h)}}=a^{(h)}(1-a^{(h)})=\sigma'(Z^{(h)})$$

$$\frac{\partial Z^{(h)}}{\partial W^{(h)}}=\frac{\partial(X\cdot W^{h}+b^{(h)})}{\partial W^{(h)}}=X$$

* b gradient
1.)$\frac{\partial}{\partial b^{(h)}}J(W)=\frac{\partial J(W)}{\partial a^{(h)}}\cdot\frac{\partial a^{(h)}}{\partial Z^{(h)}}\cdot\frac{\partial Z^{(h)}}{\partial b^{(h)}}=\delta^{(h)}$
2.)$\frac{\partial}{\partial b^{(out)}}J(W)=\frac{\partial J(W)}{\partial a^{(out)}}\cdot\frac{\partial a^{(out)}}{\partial Z^{(out)}}\cdot\frac{\partial Z^{(out)}}{\partial b^{(out)}}=\delta^{(out)}$

V. 圖



VI. 結論

從圖形可看出，我們一開始第一次的 cost 值很大，是因為還沒做訓練，經過 100 次的迭代訓練後, 經過梯度下降法，降低誤差, 我們的預測值會很接近真實值，到最後讓我們的預測值很接近真實值。