陈羿侨 Yiqiao Chen 2024/10/21

① Complexity penalty

I. High model complexity can lead to overfitting

weight w

II. One approch is to add complexity penalty term to the objective function, there by reducing the model complexity during the training process.

regularization parameter

$$R(w) = \underbrace{\mathcal{E}_{av}(w)}_{\text{learning to reduce error}} + \underbrace{\lambda \mathcal{E}_c(w)}_{\text{learning to reduce model complexity}}$$

III. How we measure $\mathcal{E}_c(w)$?

Empirically, we believe that model complexity is related to the magnitude of weight.
This leads to the concept of weight decay:

$$\text{define } \mathcal{E}_c(w) := \|w\|^2 = \sum_{i \in \mathcal{N}} w_i^2$$

$$\text{weight decay: } R(w) = \mathcal{E}_{av}(w) + \lambda \|w\|^2$$

② Network Pruning: Optimal brain surgeon

1. Defination to excess weights

We believe that the weights that cause a slight reduction in train error that causing the network overfit are relatively redundant.

II. Derivation

We aim to find the weight w_i such that when the weight is modified $w' = w + \Delta w$ and $\mathbb{1}_i^T \Delta w + w_i = 0$, the change in error remains within an acceptable range.

$$\mathcal{E}_{av}(w + \Delta w) - \mathcal{E}_{av}(w) < k$$

Using Taylor expansion:

hessian matrix

$$\mathcal{E}_{av}(w+\Delta w) = \mathcal{E}_{av}(w) + \cancel{\Delta w \nabla \mathcal{E}_{av}(w)} + \frac{1}{2}\Delta w^T \underset{\smile}{H} \Delta w + \cancel{O(\|\Delta w\|^3)}$$

We do OBS after the training process has converged, so the gradient $\nabla \mathcal{E}_{av}$ may set to zero.

Assume that the error surface around minimum is nearly quadratic. High order terms may also be neglected.

Now we get the error changing term:

$$\Delta \mathcal{E}_{av}(w) = \mathcal{E}_{av}(w+\Delta w) - \mathcal{E}_{av}(w)$$

$$= \frac{1}{2}\Delta w^T H \Delta w$$

And then we want to establish the relationship between the error term and w_i.

for objective function: $\min\limits_{\Delta w} \frac{1}{2}\Delta w^T H \Delta w$ and constrain: $\mathbb{1}_i^T \Delta w + w_i = 0$

Using Lagrange function.

$$S = \frac{1}{2}\Delta w^T H \Delta w - \lambda(\mathbb{1}_i^T \Delta w + w_i)$$

$$\frac{\partial S}{\partial \Delta w} = H\Delta w - \lambda \mathbb{1}_i = 0$$ ⇐ extreme point has 0 gradient

and for $\Delta w = H^{-1}\lambda \mathbb{1}_i$, we have

$$\begin{bmatrix} \Delta w_1 \\ \Delta w_2 \\ \vdots \\ \Delta w_i \\ \vdots \\ \Delta w_m \end{bmatrix} = \begin{bmatrix} [H^{-1}]_1 & [H^{-1}]_2 & \cdots & [H^{-1}]_n \end{bmatrix} \begin{array}{l} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \lambda \\ \vdots \\ 0 \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \\ \leftarrow \text{index } i \\ \\ \\ \end{array}$$

recall that we have $\Delta w_i + w_i = 0$, we get

$$\Delta w_i = \lambda [H^{-1}]_{ii}$$

Now we hav

$$\text{i. } \lambda = \frac{\delta w_i}{[H^{-1}]_{ii}}$$

$$\text{ii. } \delta w = H^{-1} \frac{\delta w_i}{[H^{-1}]_{ii}} \mathbb{1}_i$$

Substitude these two terms into S, and we have:

$$S_i = \frac{w_i^2}{2[H^{-1}]_{ii}}$$

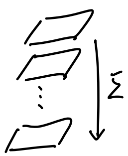
III. Compute the H^{-1}

for only one output neuron,

$$\mathcal{E}_{av}(w) = \frac{1}{2N} \sum_{n=1}^{N} (d(n) - F(w, x))$$

$$\frac{\partial \mathcal{E}_{av}(w)}{\partial w} = -\frac{1}{N} \sum_{n=1}^{N} \frac{\partial F(w,x)}{\partial w} (d(n) - F(w, x))$$

Apply multiplication rule,

$\boxed{H(N)} =$ 

$$H(N) = \frac{\partial^2 \mathcal{E}_{av}(w)}{\partial w^2} \approx \frac{1}{N} \sum_{n=1}^{N} \left(\frac{\partial F(w, x(n))}{\partial w} \right) \left(\frac{\partial F(w, x(n))}{\partial w} \right)^T$$

denote $\xi(n) = \frac{1}{\sqrt{N}} \frac{\partial F(w, x(n))}{\partial w}$

for m layer, we have

$$H(m) = \sum_{k=1}^{m} \xi(k) \xi^T(k) = H(m-1) + \xi(n) \xi^T(n)$$

Using Woodbury's Equality, we can calculate the inverse

$$H^{-1}(n) = H^{-1}(n-1) - \frac{H^{-1}(n-1)\xi(n)\xi^T(n)H^{-1}(n-1)}{1 + \xi^T(n)H^{-1}(n-1)\xi(n)}$$

and initial condition $H^{-1}(0) = \delta^{-1} I$