The core of Neural Network

This is to give a very brief summary of essence of Neural Network.

Model

Simple Neural Network

Forward Neural Network

$$[z]_{1\times r} = [h]_{1\times n}[w]_{n\times r} + [b]_{1\times r}$$
$$[\hat{y}]_{1\times r} = soft\max([z]_{1\times r})$$
$$L = [y]_{1\times r} \circ \log[\hat{y}]_{1\times r}$$

Back Propagation

$$\frac{\partial L}{\partial z} = [\hat{y}]_{1\times r} - [y]_{1\times r}$$
$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial w} = \left([\hat{y}]_{1\times r} - [y]_{1\times r}\right)_{n\times 1} \circ transpose([h]_{r\times 1})$$
$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial b} = [\hat{y}]_{1\times r} - [y]_{1\times r}$$
$$\frac{\partial L}{\partial h} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial h} = \left([\hat{y}]_{1\times r} - [y]_{1\times r}\right)[w]^T_{n\times r}$$

Variant with dropout

Forward

$$[r]_{1\times n} \sim Ber(p)$$
$$[\widetilde{h}]_{1\times n} = [r]_{1\times n} \circ [h]_{1\times n}$$
$$[z]_{1\times r} = [\widetilde{h}]_{1\times n}[w]_{n\times r} + [b]_{1\times r}$$
$$[\hat{y}]_{1\times r} = soft\max([z]_{1\times r})$$
$$L = [y]_{1\times r} \circ \log[\hat{y}]_{1\times r}$$

Backward

$$\frac{\partial L}{\partial z} = [\hat{y}]_{1\times r} - [y]_{1\times r}$$
$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial w} = \left([\hat{y}]_{1\times r} - [y]_{1\times r}\right)_{n\times 1} \circ transpose([h]_{r\times 1})$$
$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial b} = [\hat{y}]_{1\times r} - [y]_{1\times r}$$
$$\frac{\partial L}{\partial \widetilde{h}} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial \widetilde{h}} = \left([\hat{y}]_{1\times r} - [y]_{1\times r}\right)[w]^T_{n\times r}$$
$$\frac{\partial L}{\partial h} = \frac{\partial L}{\partial \widetilde{h}}\frac{\partial \widetilde{h}}{\partial h} = \left([\hat{y}]_{1\times r} - [y]_{1\times r}\right)[w]^T_{n\times r} \circ [r]_{1\times n}$$

Solver

Stochastic Gradient Descent

Momentum

$$v_t = \gamma v_{t-1} + \eta \nabla_\theta J(\theta)$$
$$\theta = \theta - v_t$$

NAG

$$v_t = \gamma v_{t-1} + \eta \nabla_\theta J(\theta - \gamma v_{t-1})$$
$$\theta = \theta - v_t$$

Conjugate Gradient

$$\min \frac{1}{2} x^T Q x + b^T x + c$$

$$\textit{choose } x^1,\ p_1 = -\nabla f\left(x^1\right)$$

$$\textit{if } \nabla f\left(x^k\right) = 0,\ \textit{stop};$$

$$\textit{otherwise } x^{k+1} = x^k + a_k p_k$$

$$a_k = -\frac{p_k^T \nabla f\left(x^k\right)}{p_k^T Q p_k}$$

$$p_{k+1} = -\nabla f\left(x^{k+1}\right) + \lambda_k p_k$$

$$\lambda_k = \frac{p_k^T Q \nabla f\left(x^k\right)}{p_k^T Q p_k}$$

BFGS

$$d_k = H_k \nabla f\left(x^k\right)$$
$$x^{k+1} = x^k - a_k d_k$$
$$f(x) = f\left(x^{k+1}\right) + \nabla f\left(x^{k+1}\right)\left(x - x^{k+1}\right) + \frac{1}{2}\left(x - x^{k+1}\right)^T \nabla^2 f\left(x^{k+1}\right)\left(x - x^{k+1}\right)$$
$$\nabla f(x) - \nabla f\left(x^{k+1}\right) = \nabla^2 f\left(x^{k+1}\right)\left(x - x^{k+1}\right)$$
$$\nabla f\left(x^{k+1}\right) - \nabla f\left(x^k\right) = \nabla^2 f\left(x^{k+1}\right)\left(x^{k+1} - x^k\right)$$
$$g_{k+1} - g_k = \nabla^2 f\left(x^{k+1}\right)\left(x^{k+1} - x^k\right)$$
$$s_k = x^{k+1} - x^k$$
$$y_k = g_{k+1} - g_k$$
$$H_{k+1} = \left(\nabla^2 f\left(x^{k+1}\right)\right)^{-1} = B_{k+1}^{-1}$$
$$B_{k+1} = B_k + E_k$$
$$E_k = a_k U_k U_k^T + b_k V_k V_k^T$$
$$y_k = \left(B_k + a_k U_k U_k^T + b_k V_k V_k^T\right)s_k$$
$$y_k - B_k s_k = a_k U_k U_k^T s_k + b_k V_k V_k^T s_k$$
$$U_k = y_k \Rightarrow a_k = \frac{1}{y_k^T s_k}$$
$$V_k = -B_k s_k \Rightarrow b_k = -\frac{1}{s_k^T B_k s_k}$$
$$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k}$$
$$H_{k+1} = \left(I - \frac{s_k y_k^T}{y_k^T s_k}\right) H_k \left(I - \frac{y_k s_k^T}{y_k^T s_k}\right) + \frac{s_k s_k^T}{y_k^T s_k}$$

L-BFGS

$$H_{k+1} = \left(I - \frac{s_k y_k^T}{y_k^T s_k}\right) H_k \left(I - \frac{y_k s_k^T}{y_k^T s_k}\right) + \frac{s_k s_k^T}{y_k^T s_k}$$
$$\rho_k = \frac{1}{y_k^T s_k},\ V_k = 1 - \rho_k y_k s_k^T$$
$$H_{k+1} = V_k^T H_k V_k + \rho_k s_k s_k^T$$
$$H_k = V_{k-1}^T H_{k-1} V_{k-1} + \rho_{k-1} s_{k-1} s_{k-1}^T$$
$$\vdots$$
$$H_1 = V_0^T H_0 V_0 + \rho_0 s_0 s_0^T$$
$$\begin{aligned} H_k = &\left(V_{k-1}^T \cdots V_0^T\right) H_0 \left(V_0 \cdots V_{k-1}\right) + \\ &\left(V_{k-1}^T \cdots V_1^T\right) \rho_0 s_0 s_0^T \left(V_1 \cdots V_{k-1}\right) + \\ &\left(V_{k-1}^T \cdots V_2^T\right) \rho_1 s_1 s_1^T \left(V_2 \cdots V_{k-1}\right) + \\ &\cdots \\ &V_{k-1}^T \rho_{k-2} s_{k-2} s_{k-2}^T V_{k-1} + \\ &\rho_{k-1} s_{k-1} s_{k-1}^T \end{aligned}$$