HW2

1. derivation of optimal W for a binary SVM
2. derivation of optimal W's for a 3-layer MLP
3. define a loss-function for SVM

derivation of optimal W for a binary SVM

$$\boldsymbol{margin = \rho = \frac{2}{\|W\|}}$$

$$\boldsymbol{max\rho \Leftrightarrow max\rho^2 \Leftrightarrow min\frac{1}{2}\ \|W\|^2}$$

$$X_i^T W + b \geq +1, y_i = +1$$

$$X_i^T W + b \geq -1, y_i = -1$$

$$\boldsymbol{minJ(W) = min\frac{1}{2}\ \|W\|^2}$$

$$s.t. \quad yi(X_i^T W + b) \geq 1, y_i = 1,2, \dots, n$$

$$\boldsymbol{L(W, b, \alpha) = \frac{1}{2}\ \|W\|^2 - \sum_{i=1}^{n} \alpha_i\ [y_i(X_i^T W + b) - 1]}$$

$$\boldsymbol{maxL(W, b, \alpha) = +\infty}$$

$$\boldsymbol{maxL(W, b, \alpha) = J(W) = \frac{1}{2}\ \|W\|^2}$$

$$\boldsymbol{minmaxL(W, b, \alpha)}$$

$$\boldsymbol{\nabla_W L(W, b, \alpha) = W - \sum_{i=1}^{n} \alpha_i\, y_i X_i = 0 \Rightarrow W = \sum_{i=1}^{n} \alpha_i\, y_i X_i}$$

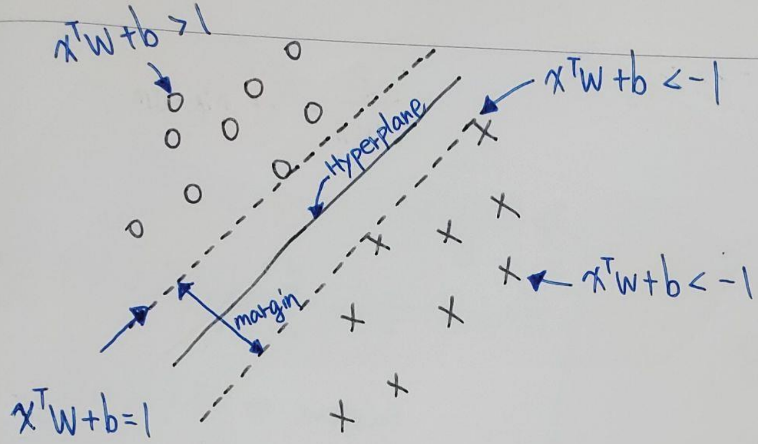
$$\boldsymbol{\nabla_b L(W, b, \alpha) = -\sum_{i=1}^{n} \alpha_i\, y_i = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i\, y_i = 0}$$

$$\boldsymbol{minL(W, b, \alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i\, a_j y_i y_j X_i^T X_i + \sum_{i=1}^{n} a_i}$$

$$\boldsymbol{min = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i\, a_j y_i y_j X_i^T X_i - \sum_{i=1}^{n} a_i}$$

$$\boldsymbol{\sum_{i=1}^{n} \alpha_i\, y_i = 0}$$

$$\alpha_i \geq 0, i = 1,2, \dots, n$$



$$\text{margin} = \rho = \frac{2}{\|w\|}$$

$$\max \rho \iff \max \rho^2 \iff \min \frac{1}{2}\|w\|^2$$

$$x_i^T W + b \geq +1, \; y_i = +1$$
$$x_i^T W + b \leq -1, \; y_i = -1$$

$$\min J(w) = \min \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad y_i(x_i^T W + b) \geq 1, \; i = 1, 2, \cdots, n$$

$$y_i(x_i^T W + b) = 1$$

$$L(W, b, \alpha) = \frac{1}{2}\|W\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(x_i^T W + b) - 1]$$

$$\max L(W, b, \alpha) = +\infty$$
$$\max L(W, b, \alpha) = J(W) = \frac{1}{2}\|W\|^2$$
$$\min\max L(W, b, \alpha)$$

$$\nabla_W L(W, b, \alpha) = W - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \Rightarrow W = \sum_{i=1}^{n} \alpha_i y_i x_i$$
$$\nabla_b L(W, b, \alpha) = -\sum_{i=1}^{n} \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$L(W, b, \alpha) = \frac{1}{2}\|W\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(x_i^T W + b) - 1]$$
$$= \frac{1}{2}\sum_{i=1}^{n} \alpha_i y_i x_i^T \cdot \sum_{j=1}^{n} \alpha_j y_j x_j - \sum_{i=1}^{n} \alpha_i y_i x_i^T \cdot \sum_{j=1}^{n} \alpha_j y_j x_j - b\sum_{i=1}^{n} \alpha_i y_i + \sum_{i=1}^{n} \alpha_i$$
$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n} \alpha_i y_i x_i^T \sum_{j=1}^{n} \alpha_j y_j x_j = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\min L(W, b, \alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{n}\alpha_i$$

$$\min \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^{n}\alpha_i$$

$$\sum_{i=1}^{n}\alpha_i y_i = 0 \quad . \quad \alpha_i \geq 0, 1, 2, \ldots, n$$

derivation of optimal W's for a 3-layer MLP

運算法則

$$y = XW + b$$

$$y = \sum x_i * w_i + b$$

輸入 X 乘以權重 W 得到 y，再通過啟動函數得到輸出（O）。 在這裡，啟動函數是 sigmoid 函數

$$\sigma(y) = \frac{1}{1+e^{-y}}$$

E 是 loss 函數值，這裡是輸出值（output）與真實值（target）的歐式距離

$$E = \frac{1}{2}(O_0^1 - t)^2$$

E 的大小是評價感知器模型好壞的指標之一，w 權重是描述這個感知器模型的參數，通過計算 E 來優化感知器模型，即優化 w 的值

$\boldsymbol{w_{jk}^{I}}$表示第 I 層，第 j 個輸入連結第 k 個輸出的權值 w。 以下先對一個權重（值）w 求得感知器模型的梯度

$$\frac{\partial E}{\partial w_{j0}^1} = (O_0^1 - t)\frac{\partial O_0^1}{\partial w_{j0}^1}$$

$$\frac{\partial E}{\partial w_{j0}^1} = (O_0^1 - t)\frac{\partial \sigma(x_0^1)}{\partial w_{j0}^1}$$

$$\frac{\partial E}{\partial w_{j0}^1} = (O_0^1 - t)\frac{\partial \sigma(x_0^1)}{\partial x_0^1}\frac{\partial x_0^1}{\partial w_{j0}^1}$$

$$\frac{\partial E}{\partial w_{j0}^1} = (O_0^1 - t)\sigma(x_0^1)(1-\sigma(x_0^1))\frac{\partial x_0^1}{\partial w_{j0}^1}$$

$$\frac{\partial E}{\partial w_{j0}^1} = (O_0^1 - t)O_0^1(1-O_0^1)x_j^0\frac{\partial x_0^1}{\partial w_{j0}^1}$$

$$\frac{\partial E}{\partial w_{j0}^1} = (O_0^1 - t)O_0^1(1-O_0^1)x_j^0$$

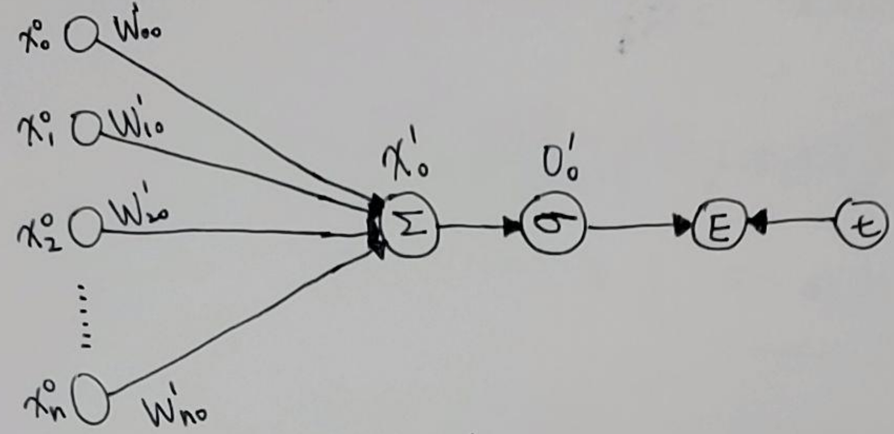
現在把單個輸出的感知器模型推廣成多輸出感知器模型

$$\frac{\partial E}{\partial w_{jk}^1} = (O_k^1 - t_k)\frac{\partial O_k^1}{\partial w_{jk}^1}$$

$$\frac{\partial E}{\partial w_{jk}^1} = (O_k^1 - t_k)\sigma(x_k^1)(1-\sigma(x_k^1))\frac{\partial x_k^1}{\partial w_{jk}^1}$$

$$\frac{\partial E}{\partial w_{jk}^1} = (O_k^1 - t_k)O_k^1(1-O_k^1)x_j^0$$

單輸出



$$y = xw + b$$

$$y = \sum x_i \cdot w_i + b$$

$$\sigma(y) = \frac{1}{1+e^{-y}}$$

$$E = \frac{1}{2}(O^1_0 - t)^2$$

$$\frac{\partial E}{\partial w^1_{j0}} = (O^1_0 - t)\frac{\partial O^1_0}{\partial w^1_{j0}}$$

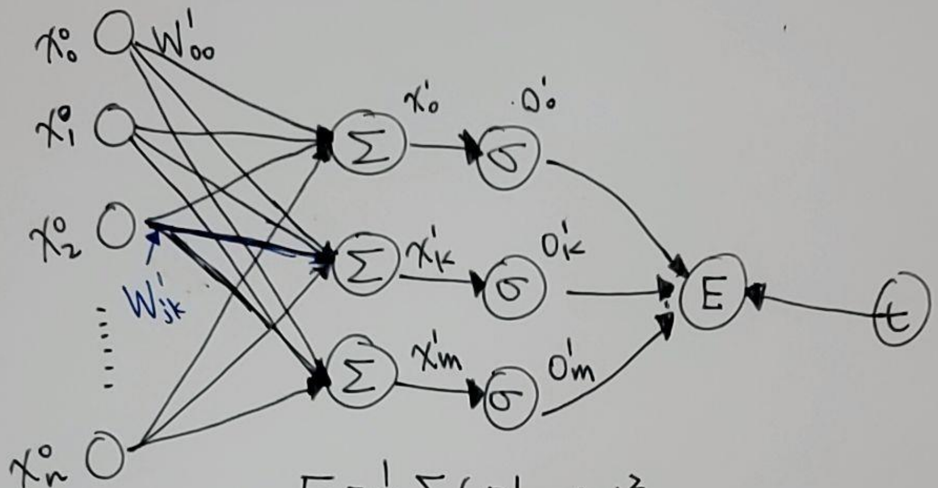
$$\frac{\partial E}{\partial w^1_{j0}} = (O^1_0 - t)\frac{\partial \sigma(x^1_0)}{\partial w^i_j}$$

$$\frac{\partial E}{\partial w^1_{j0}} = (O^1_0 - t)\frac{\partial \sigma(x^1_0)}{\partial x^1_0}\frac{\partial x^1_0}{\partial w^1_{j0}}$$

$$\frac{\partial E}{\partial w^1_{j0}} = (O^1_0 - t)\sigma(x^1_0)(1 - \sigma(x^1_0))\frac{\partial x^1_0}{\partial w^1_{j0}}$$

$$\frac{\partial E}{\partial w^1_{j0}} = (O^1_0 - t)O^1_0(1 - O^1_0)\frac{\partial x^1_0}{\partial w^1_{j0}}$$

$$\frac{\partial E}{\partial w^1_{j0}} = (O^1_0 - t)O^1_0(1 - O^1_0)x^0_j$$



$$E = \frac{1}{2} \sum (o_i^1 - t_i)^2$$

$$\frac{\partial E}{\partial w_{jk}^1} = (o_k^1 - t_k) \frac{\partial o_k^1}{\partial w_{jk}^1}$$

$$\frac{\partial E}{\partial w_{jk}^1} = (o_k^1 - t_k) \frac{\partial \sigma(x_k^1)}{\partial w_{jk}^1}$$

$$\frac{\partial E}{\partial w_{jk}^1} = (o_k^1 - t_k) \sigma(x_k^1)(1-\sigma(x_k^1)) \frac{\partial x_k^1}{\partial w_{jk}^1}$$

$$\frac{\partial E}{\partial w_{jk}^1} = (o_k^1 - t_k) o_k^1 (1 - o_k^1) x_j^0$$

<u>define a loss-function for SVM</u>

用來評估模型的預測值與真實值不一致的程度，也是神經網絡中優化的目標函數，神經網絡訓練或者優化的過程就是最小化損失函數的過程，損失函數越小，說明模型的預測值就越接近真實值，模型的健壯性也就越好。

Hinge Loss function：

$$L(y, f(x)) = \max(0, 1 - yf(x))$$