1.1.

training data: $\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$

where: $x_i \in R^d$, $y_i \in R^k$ for $i=1,2,\cdots,N$

consider: $\min_{W\in R^{k\times d}, b\in R^k} \sum_{i=1}^{N} \| y_i - Wx_i - b\|_2^2$ (least-square problem)

Q1: set $\tilde{x}_i = \begin{bmatrix} x_i \\ 1 \end{bmatrix} \in R^{d+1}$

$\tilde{W} = [W|b] \in R^{k\times(d+1)}$

then origin question $\Leftrightarrow \min_{\tilde{W}} \sum_{i=1}^{N} \| y_i - \tilde{W}\tilde{x}_i\|_2^2$

set $X = \begin{bmatrix} \tilde{x}_1^T \\ \tilde{x}_2^T \\ \vdots \\ \tilde{x}_N^T \end{bmatrix} = \begin{bmatrix} x_1^T & 1 \\ x_2^T & 1 \\ \vdots & \\ x_N^T & 1 \end{bmatrix} \in R^{N\times(d+1)}$; $Y = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_N^T \end{bmatrix} \in R^{N\times k}$

then $L(\tilde{W}) = \| Y - X\tilde{W}^T\|_F^2 = tr((Y-X\tilde{W}^T)^T(Y-X\tilde{W}^T))$

$= tr(Y^TY) - 2tr(Y^TX\tilde{W}^T) + tr(\tilde{W}X^TX\tilde{W}^T)$

$\therefore \frac{\partial L}{\partial \tilde{W}} = -2(Y^TX)^T + 2(X^TX\tilde{W}^T)^T = -2X^TY + 2X^TX\tilde{W}^T$

let $\frac{\partial L}{\partial \tilde{W}} = 0 \Rightarrow -2X^TY + 2X^TX\tilde{W}^T = 0$

$\Rightarrow X^TX\tilde{W}^T = X^TY$

assume X^TX is invertible then

$\tilde{W}^T = (X^TX)^{-1}X^TY$

$\because \tilde{W} = [W|b]$

$\therefore W^* = \tilde{W}[:,:d]$, $b^* = \tilde{W}[:,d]$

where $\tilde{W}^T = (X^TX)^{-1}X^TY \in R^{(d+1)\times k}$

$\therefore$ final closed form

$\begin{bmatrix} W^* \\ b^* \end{bmatrix} = (X^TX)^{-1}X^TY \in R^{(d+1)\times k}$ where $X = \begin{bmatrix} \tilde{x}_1^T \\ \tilde{x}_2^T \\ \vdots \\ \tilde{x}_N^T \end{bmatrix} = \begin{bmatrix} x_1^T & 1 \\ x_2^T & 1 \\ \vdots & \\ x_N^T & 1 \end{bmatrix} \in R^{N\times(d+1)}$; $Y = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_N^T \end{bmatrix} \in R^{N\times k}$

Q2:

for a single sample (x_i, y_i), the loss is

$l_i(W,b) = \| y_i - Wx_i - b\|_2^2 = (y_i - Wx_i - b)^T(y_i - Wx_i - b)$

we have

$\frac{\partial l_i}{\partial W} = -2(y_i - Wx_i - b)x_i^T$; $\frac{\partial l_i}{\partial b} = -2(y_i - Wx_i - b)$

$\therefore \frac{\partial L}{\partial W} = -2\sum_{i=1}^{N}(y_i - Wx_i - b)x_i^T$ and $\frac{\partial L}{\partial b} = -2\sum_{j=1}^{N}(y_i - Wx_i - b)$

$\Rightarrow$ At iteration t, we have

$W^{(t+1)} = W^{(t)} + 2\eta\sum_{i=1}^{N}(y_i - W^{(t)}x_i - b^{(t)})x_i^T$

$b^{(t+1)} = b^{(t)} + 2\eta\sum_{i=1}^{N}(y_i - W^{(t)}x_i - b^{(t)})$

where $\eta > 0$ is the learning rate.

in matrix form.

Gradient: $\frac{\partial L}{\partial W} = -2X^T(Y - X\tilde{W}^T)$

Update rule: $\tilde{W}^{(t+1)T} = \tilde{W}^{(t)T} + 2\eta X^T(Y - X\tilde{W}^{(t)T})$

Stopping Criteria

1. Stop when the gradient norm is small enough:

$\| \frac{\partial L}{\partial \tilde{W}} \|_F < \epsilon_g$

2. Stop when the change in loss between consecutive iteration is small enough

$|L(W^{(t+1)}, b^{(t+1)}) - L(W^{(t)}, b^{(t)})| < \epsilon_L$ or $\frac{|L^{(t)} - L^{(t+1)}|}{|L^{(t)} + \epsilon|} < \epsilon_L$

Q3.

consider loss for single sample

$$\ell_i(W) = -\sum_{k=1}^{K} y_{i,k} \log(\hat{y}_{i,k})$$

then $\frac{\partial \ell_i}{\partial W} = \frac{\partial \ell_i}{\partial z_i} \cdot \frac{\partial z_i}{\partial W}$

where $\frac{\partial \ell_i}{\partial z_{i,k}} = -\sum_{j=1}^{K} y_{i,j} \frac{\partial \log(\hat{y}_{i,j})}{\partial z_{i,k}}$

for $\frac{\partial \hat{y}_{i,j}}{\partial z_{i,k}}$ when $j=k$, $\frac{\partial \hat{y}_{i,j}}{\partial z_{i,k}} = \hat{y}_{i,k}(1-\hat{y}_{i,k})$

for softmax

when $j \neq k$, $\frac{\partial \hat{y}_{i,j}}{\partial z_{i,k}} = -\hat{y}_{i,j}\hat{y}_{i,k}$

combined, we have

$$\frac{\partial \hat{y}_{i,j}}{\partial z_{i,k}} = \hat{y}_{i,j}(\delta_{jk} - \hat{y}_{i,k})$$ where δ_{jk} is the Kronecker delta.

$$\therefore \frac{\partial \ell_i}{\partial z_{i,k}} = -\sum_{j=1}^{K} y_{i,j} \frac{1}{\hat{y}_j} \frac{\partial \hat{y}_{i,j}}{\partial z_{i,k}}$$

$$= -\sum_{j=1}^{K} y_{i,j} \frac{1}{\hat{y}_{i,j}} (\delta_{jk} - \hat{y}_{i,k}) \cdot \hat{y}_{i,j}$$

$$= -\sum_{j=1}^{K} y_{i,j}(\delta_{jk} - \hat{y}_{i,k})$$

$$= -\sum_{j=1}^{K} y_{i,j}\delta_{jk} + \sum_{j=1}^{K} y_{i,j}\hat{y}_{i,k}$$

$$= -y_{ik} + \hat{y}_{i,k}\sum_{j=1}^{K} y_{i,j}$$

$\because y_i$ is one-hot encoded ($\sum_{j=1}^{K} y_{i,j} = 1$)

$$\therefore \frac{\partial \ell_i}{\partial z_{i,k}} = \hat{y}_{i,k} - y_{i,k}$$

in vector form

$$\frac{\partial \ell_i}{\partial z_i} = \hat{y}_i - y_i$$

$\because z_i = W^T x_i$

$\therefore z_{i,k} = \sum_{m=1}^{d} (W_{m,k} x_{i,m})$

$\therefore \frac{\partial z_{i,k}}{\partial W_{m,n}} = x_{i,m}\delta_{kn}$

in matrix form

$\frac{\partial z_i}{\partial W} = x_i \otimes I_K$ where $\otimes$ is the outer product.

$$\therefore \frac{\partial \ell_i}{\partial W} = x_i(\hat{y}_i - y_i)^T$$

$$\therefore \frac{\partial J}{\partial W} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \ell_i}{\partial W}$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i(\hat{y}_i - y_i)^T$$

set $X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix} = R^{N\times d}$; $Y = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_N^T \end{bmatrix} \in R^{N\times K}$, $\hat{Y} = \begin{bmatrix} \hat{y}_1^T \\ \hat{y}_2^T \\ \vdots \\ \hat{y}_N^T \end{bmatrix}$

then $\sum_{i=1}^{N} x_i(\hat{y}_i - y_i)^T = X^T(\hat{Y} - Y)$

$$\therefore \nabla_W J = \frac{1}{N} X^T(\hat{Y} - Y)$$

1.2.

Q1.

Slack Variables

$\xi_i \geq 0$: measures the violation when $y_i - f(x_i) > \varepsilon$

$\xi_i^* \geq 0$: measures the violation when $f(x_i) - y_i > \varepsilon$

$\therefore$ origin problem becomes: $\min_{w,\xi,\xi^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$

Subject to

$$y_i - w^T x_i \le \epsilon + \xi_i^*$$
$$w^T x_i - y_i \le \epsilon + \xi_i$$

Q2.

introduce

$\alpha_i \geq 0$ for constrains $w^T x_i - y_i \le \epsilon + \xi_i$

$\alpha_i^* \geq 0$ for constrains $y_i - w^T x_i \le \epsilon + \xi_i^*$

$\lambda_i \geq 0$ for constrains $\xi_i \geq 0$

$\lambda_i^* \geq 0$ for constrains $\xi_i^* \geq 0$

$$\therefore L(w, \xi, \xi^*, \alpha, \alpha^*, \lambda, \lambda^*) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) + \sum_{i=1}^{n}\alpha_i(w^T x_i - y_i - \epsilon - \xi_i) + \sum_{i=1}^{n}\alpha_i^*(w^T x_i - y_i - \epsilon - \xi_i^*) - \sum_{i=1}^{n}\lambda_i\xi_i - \sum_{i=1}^{n}\lambda_i^*\xi_i^*$$

Q3.

We have

$$L(w, \xi, \xi^*, \alpha, \alpha^*, \lambda, \lambda^*) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) + \sum_{i=1}^{n}\alpha_i(w^T x_i - y_i - \epsilon - \xi_i) + \sum_{i=1}^{n}\alpha_i^*(w^T x_i - y_i - \epsilon - \xi_i^*) - \sum_{i=1}^{n}\lambda_i\xi_i - \sum_{i=1}^{n}\lambda_i^*\xi_i^*$$

$\because$ KKT

$$\therefore \frac{\partial L}{\partial w} = w + \sum_{i=1}^{n}\alpha_i x_i - \sum_{i=1}^{n}\alpha_i^* x_i = 0$$
$$\Rightarrow w = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)x_i$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \lambda_i = 0$$
$$\Rightarrow C = \alpha_i + \lambda_i$$

$$\frac{\partial L}{\partial \xi_i^*} = C - \alpha_i^* - \lambda_i^* = 0$$
$$\Rightarrow C = \alpha_i^* + \lambda_i^*$$

$$\because \alpha_i(w^T x_i - y_i - \epsilon - \xi_i) = 0$$
$$\alpha_i^*(y_i - w^T x_i - \epsilon - \xi_i^*) = 0$$
$$\lambda_i\xi_i = 0$$
$$\lambda_i^*\xi_i^* = 0$$

$$\therefore \frac{1}{2}\|w\|^2 = \frac{1}{2}\left\|\sum_{i=1}^{n}(\alpha_i^* - \alpha_i)x_i\right\|^2 = \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)x_i^T x_j$$

$$\because C = \alpha_i + \lambda_i \quad C = \alpha_i^* + \lambda_i^*$$

$$\therefore C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}\lambda_i\xi_i - \sum_{i=1}^{n}\lambda_i^*\xi_i^* = \sum_{i=1}^{n}\alpha_i\xi_i + \sum_{i=1}^{n}\alpha_i^*\xi_i^*$$

$$\therefore L = -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)x_i^T x_j + \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)y_i - \epsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*)$$

$\therefore$ we have Dual From of SVR

Maximize

$$L_D(\alpha, \alpha^*) = -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)x_i^T x_j + \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)y_i - \epsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*)$$

Subject to:

$$0 \le \alpha_i \le C, \; i = 1, \dots, n$$
$$0 \le \alpha_i^* \le C, \; i = 1, \dots, n$$

Using $w = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)x_i$ we have $f(x) = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)x_i^T x$

Q4. yes:

Dual Problem rewritten:

Minimize

$$\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)x_i^Tx_j-\sum_{i=1}^{n}(\alpha_i^*-\alpha_i)y_i+\epsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)$$

Subject to:

$0\le\alpha_i\le C,\ i=1,\dots,n$

$0\le\alpha_i^*\le C,\ i=1,\dots,n$

We define

$u=[\alpha_1\cdots\alpha_n,\alpha_1^*\cdots,\alpha_n^*]^T\in R^{2n}$

$Q=\begin{bmatrix}K & -K\\ -K & K\end{bmatrix}$ where $K_{ij}=x_i^Tx_j$

$p=[-y_1+\epsilon,\dots,-y_n+\epsilon,y_1+\epsilon,\dots y_n+\epsilon]^T$

then we have

$\min \frac{1}{2}u^TQu+p^Tu$

subject to

$0\le u\le C1$

Q5.

define margin Support Vectors:

Points where $0<\alpha_i<C$ or $0<\alpha_i^*<C$

For $\alpha_i>0$, $w^Tx_i-y_i=\epsilon$

For $\alpha_i^*>0$, $y_i-w^Tx_i=\epsilon$

and $|f(x_i)-y_i|=\epsilon$

$\because$ KKT

$\therefore$ $\alpha_i>0$, $w^Tx_i-y_i=\epsilon+\xi_i$

$\alpha_i^*>0$, $y_i-w^Tx_i=\epsilon+\xi_i^*$

for $0<\alpha_i<C$, $\xi_i=0$

for $\alpha_i=C$, $\xi_i>0$