文杰 Homework #1 12309012

1.1 (i) Derivative with respect to w_0

Denote the error term as $e = y - Xw - w_0 1$

$\therefore$ The cost function is: $J(w, w_0) = e^T e + \lambda w^T w$

$$\frac{\partial J}{\partial w_0} = \frac{\partial}{\partial w_0}(e^T e) + \frac{\partial}{\partial w_0}(\lambda w^T w) = -2e^T 1 = -2(y - Xw - w_0 1)^T 1$$

Let $\frac{\partial J}{\partial w_0} = 0$: $(y - Xw - w_0 1)^T 1 = 0$

$\because$ The transpose of a scalar is the scalar itself: $1^T(y - Xw - w_0 1) = 0$

Expand: $1^T y - 1^T X w - w_0 1^T 1 = 0$

$1^T X = 0^T$ (X is centered, each column sums to zero) $\quad 1^T 1 = n$ (the number of samples)

$\therefore w_0 = \frac{1}{n} 1^T y = \bar{y} \quad \therefore \hat{w}_0 = \bar{y}$

(ii) Derivative with respect to w

$$\frac{\partial J}{\partial w} = \frac{\partial}{\partial w}(e^T e) + \frac{\partial}{\partial w}(\lambda w^T w) = 2e^T \frac{\partial e}{\partial w} + 2\lambda w$$

$$\frac{\partial}{\partial w}(e^T e) = 2e^T(-X) = -2(y - Xw - w_0 1)^T X$$

$\therefore \nabla_w [e^T e] = -2X^T(y - Xw - w_0 1)$

$$\frac{\partial}{\partial w}(\lambda w^T w) = 2\lambda w$$

$\nabla_w J = -2X^T(y - Xw - w_0 1) + 2\lambda w$

Set to 0: $X^T(y - Xw - w_0 1) = \lambda w$

Substitude $w_0 = \bar{y}$:

$X^T(y - Xw - \bar{y} 1) = \lambda w$

$\because X^T 1$ is a vector where the j-th element is $\sum_{i=1}^{n} x_{ij} = 0 \quad \therefore X^T 1 = 0$

$\therefore X^T(y - \bar{y} 1) = X^T y - \bar{y} X^T 1 = X^T y = X^T X w + \lambda w = (X^T X + \lambda I) w \quad \therefore w = (X^T X + \lambda I)^{-1} X^T y$

1.2 (i) Compute the feature vectors:

For $x_1 = 0$: $\phi(x_1) = [1, 0, 0]^T$

For $x_2 = 1$: $\phi(x_2) = [1, \sqrt{2}, 1]^T$

$\phi(x_2) - \phi(x_1) = [0, \sqrt{2}, 1]^T$

In SVM theory, $w = \sum \alpha_i y_i \phi(x_i)$

For two support vectors with equal $\alpha_i = \alpha$:

$$w = \alpha\left(y_1 \phi(x_1) + y_2 \phi(x_2)\right) = \alpha\left(-[1,0,0]^T + [1,\sqrt{2},1]^T\right) = \alpha[0,\sqrt{2},1]^T$$

$\therefore w$ is parallel to $[0, \sqrt{2}, 1]^T$

(iii)

For $y_1 = -1$: $y_1(w^T \phi(x_1) + w_0) = 1 \quad \therefore w^T \phi(x_1) + w_0 = -1$

For $y_2 = 1$: $y_2(w^T \phi(x_2) + w_0) = 1 \quad \therefore w^T \phi(x_2) + w_0 = 1$

Substitude the feature vectors ($w = [w_1, w_2, w_3]^T$):

$w^T \phi(x_1) = [w_1, w_2, w_3] \cdot [1, 0, 0] = w_1 \quad \therefore w_1 + w_0 = -1 \quad (1)$

$w^T \phi(x_2) = [w_1, w_2, w_3] \cdot [1, \sqrt{2}, 1] = w_1 + \sqrt{2} w_2 + w_3 \quad \therefore w_1 + \sqrt{2} w_2 + w_3 + w_0 = 1 \quad (2)$

Substitude (1) into (2): $\sqrt{2} w_2 + w_3 = 2 \quad (3)$

$\because w$ is parallel to $[0, \sqrt{2}, 1]^T$, Let $w = k[0, \sqrt{2}, 1]^T = [0, \sqrt{2}k, k]^T \quad \therefore \begin{cases} w_1 = 0 \\ w_2 = \sqrt{2}k \\ w_3 = k \end{cases}$

From (1): $w_0 = -1$
From (3): $k = \frac{2}{3}$
$\therefore w = [0, \frac{2\sqrt{2}}{3}, \frac{2}{3}]^T$

(ii) Margin $= \frac{1}{\|w\|} = \frac{1}{\frac{\sqrt{3}}{2}} = \frac{\sqrt{3}}{2}$

(iv) $w_0 = -1$

(v) $f(x) = w^T \phi(x) + w_0 = [0, \frac{2\sqrt{2}}{3}, \frac{2}{3}]^T [1, \sqrt{2}x, x^2] - 1 = \frac{2}{3}x^2 + \frac{\sqrt{2}}{3}x - 1$

2.1 Definitions:

- Input X, shape (N, D), where $N = 150$ (samples), $D = 3$ (features including bias)
- Weights W, shape (D, C), where $C = 3$ (classes)
- Logits $Z = XW$, shape (N, C)
- Predictions $\hat{Y} = \text{softmax}(Z)$, where $\hat{y}_{ic} = \frac{\exp(z_{ic})}{\sum_{j=1}^{C} \exp(z_{ij})}$
- True labels: Y_{onehot}, shape (N, C), one-hot encoded.
- Loss: $\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{ic} \log(\hat{y}_{ic})$

Now need $\frac{\partial \mathcal{L}}{\partial W_{dk}}$, where W_{dk} is the weight for feature d and class k.

$\mathcal{L}_i = -\sum_{c=1}^{C} y_{ic} \log(\hat{y}_{ic})$ Total loss: $\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_i$

$$\frac{\partial \mathcal{L}}{\partial W_{dk}} = \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial \hat{y}_{ic}} \frac{\partial \hat{y}_{ic}}{\partial z_{ik}} \frac{\partial z_{ik}}{\partial W_{dk}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_{ik} - y_{ik}) x_{id}$$

$\frac{\partial \mathcal{L}}{\partial \hat{y}_{ic}} = -\frac{1}{N} \frac{y_{ic}}{\hat{y}_{ic}}$

if $c = k$: $\frac{\partial \hat{y}_{ic}}{\partial z_{ik}} = \hat{y}_{ic}(1 - \hat{y}_{ic})$

if $c \neq k$: $\frac{\partial \hat{y}_{ic}}{\partial z_{ik}} = -\hat{y}_{ic} y_{ik}$

$z_{ik} = \sum_{d=1}^{D} x_{id} W_{dk}$

$\frac{\partial z_{ik}}{\partial W_{dk}} = x_{id}$

Matrix form: $\frac{\partial \mathcal{L}}{\partial W} = \frac{1}{N} X^T (\hat{Y} - Y_{onehot})$

shape (D, N) shape (N, C)

$\Rightarrow$ result: shape (D, C), matching W.