

H

X

L

R

R

ff

R

R

H

J

R

a

P

$S = \{X^\mu\}_{\mu=1}^N$ $X^\mu = (x_1^\mu, x_2^\mu, \dots, x_d^\mu)^T \in \mathbb{R}^d$

k

$$EMSE = \sum_{ij\mu} M_i^\mu (x_j^\mu - w_{ij})^2 = \sum_{\mu} \|X^\mu - W_{c(\mu)}\|^2$$

$W = (W_1, W_2, \dots, W_k)$ $W_i = (w_{i1}, w_{i2}, \dots, w_{id})^T \in \mathbb{R}^d$

$$M_i^\mu = \begin{cases} 1 & \text{if } c(\mu) = i \\ 0 & \text{otherwise} \end{cases} \quad \|X^\mu - W_{c(\mu)}\| = \min_j \|X^\mu - W_j\|$$

$c(\mu)$ is the class label of X^μ μ^{th}

$$E_{MSE} \mathbf{W} = \frac{1}{k} \sum_{\mu=1}^k \mathbb{E} \left[\sum_{c=1}^C \delta_{c(\mu)}^2 \left\| \sum_{i=1}^n X_i^\mu - W_c \right\|^2 \right]$$

$$E_1 \mathbf{W} = \sum_{\mu} \left\| X^\mu - W_{c(\mu)} \right\|^2, \quad E_2 \mathbf{W} = \frac{1}{P} \sum_{\mu, i \neq c(\mu)} \left\| X^\mu - W_i \right\|^{-P}$$

where $\mathbf{W} = \text{vec} \{W_1, W_2, \dots, W_n\}$ and P is a positive integer. The first term E_1 is the mean squared error, and the second term E_2 is a regularization term.

The optimization problem is to find \mathbf{W} that minimizes the cost function $E(\mathbf{W}) = E_1 \mathbf{W} + \eta E_2 \mathbf{W}$, where η is a regularization parameter.

The gradient of the cost function with respect to \mathbf{W} is given by $\nabla_{\mathbf{W}} E = \nabla_{\mathbf{W}} E_1 + \eta \nabla_{\mathbf{W}} E_2$. The gradient of E_1 is $\nabla_{\mathbf{W}} E_1 = \sum_{\mu} \delta_{c(\mu)}^2 (X^\mu - W_{c(\mu)})$, and the gradient of E_2 is $\nabla_{\mathbf{W}} E_2 = -\frac{1}{P} \sum_{\mu, i \neq c(\mu)} \left\| X^\mu - W_i \right\|^{-P-2} (X^\mu - W_i)$.

$$\left\| X^\mu - W_i \right\|^2 = (X^\mu - W_i)^T \Sigma_i^{-1} (X^\mu - W_i),$$

where Σ_i is the covariance matrix of the data points X^μ for class i . The optimization problem can be written as $\min_{\mathbf{W}} E(\mathbf{W})$, where $E(\mathbf{W}) = \sum_{\mu} \delta_{c(\mu)}^2 \left\| X^\mu - W_{c(\mu)} \right\|^2 + \eta \sum_{\mu, i \neq c(\mu)} \left\| X^\mu - W_i \right\|^{-P}$.

The gradient of the cost function with respect to \mathbf{W} is given by $\nabla_{\mathbf{W}} E = \nabla_{\mathbf{W}} E_1 + \eta \nabla_{\mathbf{W}} E_2$. The gradient of E_1 is $\nabla_{\mathbf{W}} E_1 = \sum_{\mu} \delta_{c(\mu)}^2 (X^\mu - W_{c(\mu)})$, and the gradient of E_2 is $\nabla_{\mathbf{W}} E_2 = -\frac{1}{P} \sum_{\mu, i \neq c(\mu)} \left\| X^\mu - W_i \right\|^{-P-2} (X^\mu - W_i)$.

$$\frac{\partial E}{\partial W_j} = \sum_{\mu} \delta_{c(\mu)}^j \Sigma_j^{-1} (X^\mu - W_j) - \sum_{\mu, j} \delta_{c(\mu)}^j \left\| X^\mu - W_j \right\|^{-P-2} \Sigma_j^{-1} (X^\mu - W_j)$$

The optimization problem can be solved using gradient descent. The update rule for \mathbf{W} is $\mathbf{W}^{(t)} = \mathbf{W}^{(t-1)} - \eta \nabla_{\mathbf{W}} E(\mathbf{W}^{(t-1)})$, where η is the learning rate.

$$W_j^{(t)} = W_j^{(t-1)} - \eta \Delta W_j,$$

where $\Delta W_j = -\frac{\partial E}{\partial W_j}$. The learning rate η should be chosen such that the cost function decreases monotonically.

X^μ

$$\frac{\partial E_1}{\partial W_j} \begin{cases} -\Sigma_j^{-1} X^\mu - W_j, & j = c(\mu) \\ , & \text{else} \end{cases}$$

$$\frac{\partial E_2}{\partial W_j} \begin{cases} , & j = c(\mu) \\ \left\| \|X^\mu - W_i\|^{-P-2} \Sigma_j^{-1} X^\mu - W_j \right\|, & \text{else} \end{cases}$$

Σ_j

$$\Sigma_j^{(t)} \begin{cases} -\eta' \Sigma_j^{(t-1)} & \eta' \sum_\mu X^\mu - W_{c(\mu)} \quad X^\mu - W_{c(\mu)}^T, & j = c(\mu) \\ \Sigma_j^{(t-1)}, & \text{else} \end{cases}$$

.

$$\Sigma_j^{(t)} \begin{cases} -\eta' \Sigma_j^{(t-1)} & \eta' X^\mu - W_{c(\mu)} \quad X^\mu - W_{c(\mu)}^T, & j = c(\mu) \\ \Sigma_j^{(t-1)}, & \text{else} \end{cases}$$

X^μ $\mathbb{R} \oplus \mathbb{L}$ \mathbb{L}

. $\mathbb{R} \oplus \mathbb{L}$ \mathbb{L}

. $\mathbb{R} \oplus \mathbb{L}$ \mathbb{L}

. $\mathbb{R} \oplus \mathbb{L}$ \mathbb{L}

. $\mathbb{R} \oplus \mathbb{L}$ \mathbb{L}

. $\mathbb{R} \oplus \mathbb{L}$ \mathbb{L}

. $\mathbb{R} \oplus \mathbb{L}$ \mathbb{L}

. $\mathbb{R} \oplus \mathbb{L}$ \mathbb{L}

. $\mathbb{R} \oplus \mathbb{L}$ \mathbb{L}

. $\mathbb{R} \oplus \mathbb{L}$ \mathbb{L}

3.1 On Simulated Data Sets

. d $\mathbb{R} \oplus \mathbb{L}$ \mathbb{L}

