24/01/25

K-fold cross validation

sample	x_1	x_2	Y
1	2	-1	1
2	0.5	1.2	0
3	1	2	1
4	-3	-2	1
5	4	0.1	0

3-fold cross validation

Acc

std

$F_1\{\theta_1, \theta_2\} = \{1.8, 2.8\}$

$F_2\{\theta_1, \theta_2\} = \{2.1, 3.1\}$

$F_3\{\theta_1, \theta_2\} = \{1.9, 4\}$

sample	x_1	x_2	Y
1	2	-1	1
2	0.5	1.2	0
3	1	2	1
4	-3	-2	1
5	4	0.1	0

Test

Fold 1

Train data #{1,2,3,4}

Test data #{5}

$$\hat{y}_1 = g(\theta^T \cdot x_1) = \frac{1}{1+e^{(-\theta x)}}$$

$$z = \theta x_1 = (1.8 \times 2) + (2.8 \times -1)$$

$$3.6 - 2.8 = 0.8$$

$$g(z_1) = \frac{1}{1+e^{(0.8)}} = 0.31 \quad , \hat{y}_1 = 0$$

$$z_2 = (1.8 \times 0.15) + (2.8 \quad 1.2)$$

$$0.27 + 3.36 = 3.63$$

$$g(z_2) = \frac{1}{1+e^{(z_2)}} = 0.025 \qquad \hat{y}_2 = 0$$

$$z_3 = (1.8 \times 1) + (2.8 \times 2)$$

$$= 1.8 + 5.6 = 7.4$$

$$g(z_3) = 0.0006$$

$$\hat{y}_3 = 0$$

$$z_4 = (1.8 \times -3) + (2.8 \times -2)$$
$$= -5.4 - 3.6$$
$$= -9$$

$$g(z_3) = \frac{1}{1+e^{(-9)}} = 0.99$$

$$\hat{y} = 1$$

$\hat{y}$	y
0	1
0	0
0	1
1	1

Training dataset

$$\boxed{\text{Accuracy}_1 = \frac{2}{4} = 50\%}$$

Accuracy should be calculated on test data

Accuracy on test data

$$\theta x_5 = (1.8 \times 4) + (2.8 \times 0.1)$$
$$= 7.2 + 0.28$$
$$= 7.48$$

$$g(z_5) = 0.0005$$
$$\hat{y} = 0$$

$$\boxed{\text{Accuracy}_1 = \frac{1}{1} \times 100 = 100\%}$$

Fold 2

Train data # {3, 4, 5}

Test # {1, 2}

$$\theta^T x = \begin{bmatrix} 2.1 \\ 3.1 \end{bmatrix}$$

sample	x_1	x_2	y
1	2	-1	1
2	0.5	1.2	0
3	1	2	1
4	-3	-2	1
5	4	0.1	0

Test

$$\theta^T x' = (2.1 \times 2) + (3.1 \times -1)$$
$$= 4.2 - 3.1$$
$$= 1.1$$

$$g(z_1) = \frac{1}{1+e^{(1.1)}} = 0.249$$

$\hat{y} = 0$

$$\theta^T x^2 = (2.1 \times 0.5) + (3.1 \times 1.2)$$
$$= 1.05 + 3.72$$
$$= 4.77$$

$$g(z_2) = \frac{1}{1+e^{(4.77)}} = 0.008 \qquad \hat{y}_2 = 0$$

$$\boxed{\text{Accuracy}_2 = \frac{1}{2} \times 100 = 50\%}$$

Fold 3

Train # $\{1, 2, 5\}$

Test # $\{3, 4\}$

$F_3 \{\theta_1, \theta_2\} = \{1.9, 4\}$

sample	x_1	x_2	y
1	2	-1	1
2	0.5	1.2	0
3	1	2	1
4	-3	-2	1
5	4	0.1	0

$$z_3 = \theta x^3 = (1.9 \times 1) + (4 \times 2)$$
$$= 1.9 + 8$$
$$= 9.9$$

$$g(z_3) = \frac{1}{1+e^{(z_3)}} = 0.0001$$

$$\hat{y}_3 = 0$$

$$z_4 = \theta^T x^4 = (1.9 \times -3) + (4 \times -2)$$
$$= 5.7 - 8 = -3.7$$

$g(z_3) = 0.97$

$\hat{y}_4 = 1$

$$\boxed{\text{Accuracy}_3 = \frac{1}{2} = 50\%}$$

Average accuracy $= \left(\frac{1+0.5+0.5}{3}\right) \times 100$

$= \frac{2}{3} \times 100 = 75\%$

When $K = n$; n = no. of samples

LOOCV → Leave one out cross-validation
- ↳ Computationally expensive / complex
- ↳ Use for small dataset ($n < 100$)
- ↳ To gain more confidence on model performance

Techniques to train a model

1) Train-test split — ($n > 100$) Reason → chances of having ideal distribution
2) Cross-validation
   - k-fold
   - LOOCV — ($n < 100$)