
INSEA

28

Cross-validation

Machine learning classique: zero-to-hero

Idée: Effectuer plusieurs découpages et moyenner l’erreur de test

2. Choisir une liste de valeurs de C, par ex: [0.01, 0.05, 0.1, 1., 10]

Pour chaque C:

1. Optimiser sur $\mathbf{X}_{\text{train}}$ $\mathbf{Y}_{\text{train}}$ $\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \text{loss}(f_\theta(\mathbf{x}_i), y_i) + \frac{1}{C} \text{pénalité}(\theta)$

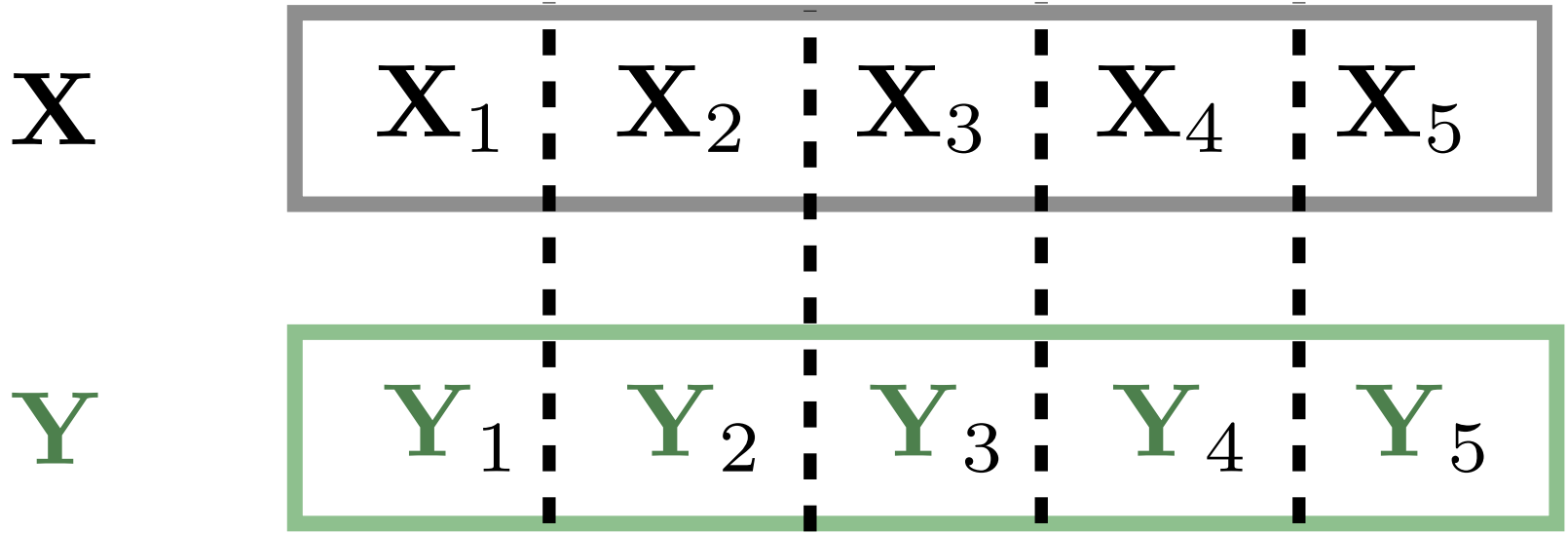
2. Évaluer l'erreur de prédiction sur $\mathbf{X}_{\text{test}}$ $\mathbf{Y}_{\text{test}}$

4. Pour chaque C, calculer l’erreur de prédiction moyenne

5. Choisir le C avec l’erreur de prédiction moyenne la plus petite

3. Pour chaque **k** in [1, 2, 3, 4, 5], créer un découpage train/test

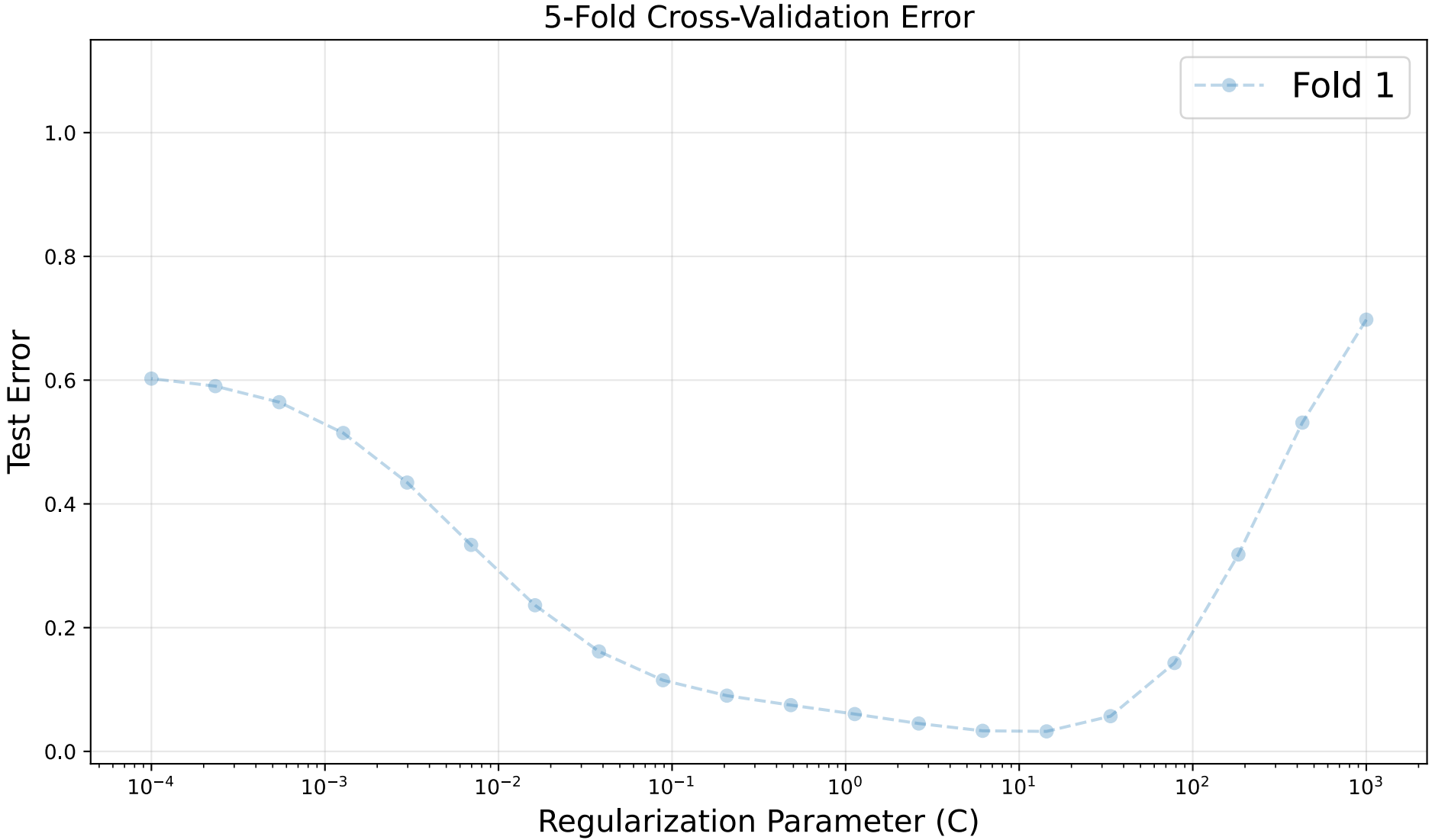
1. Couper le dataset en 5 parties (folds)

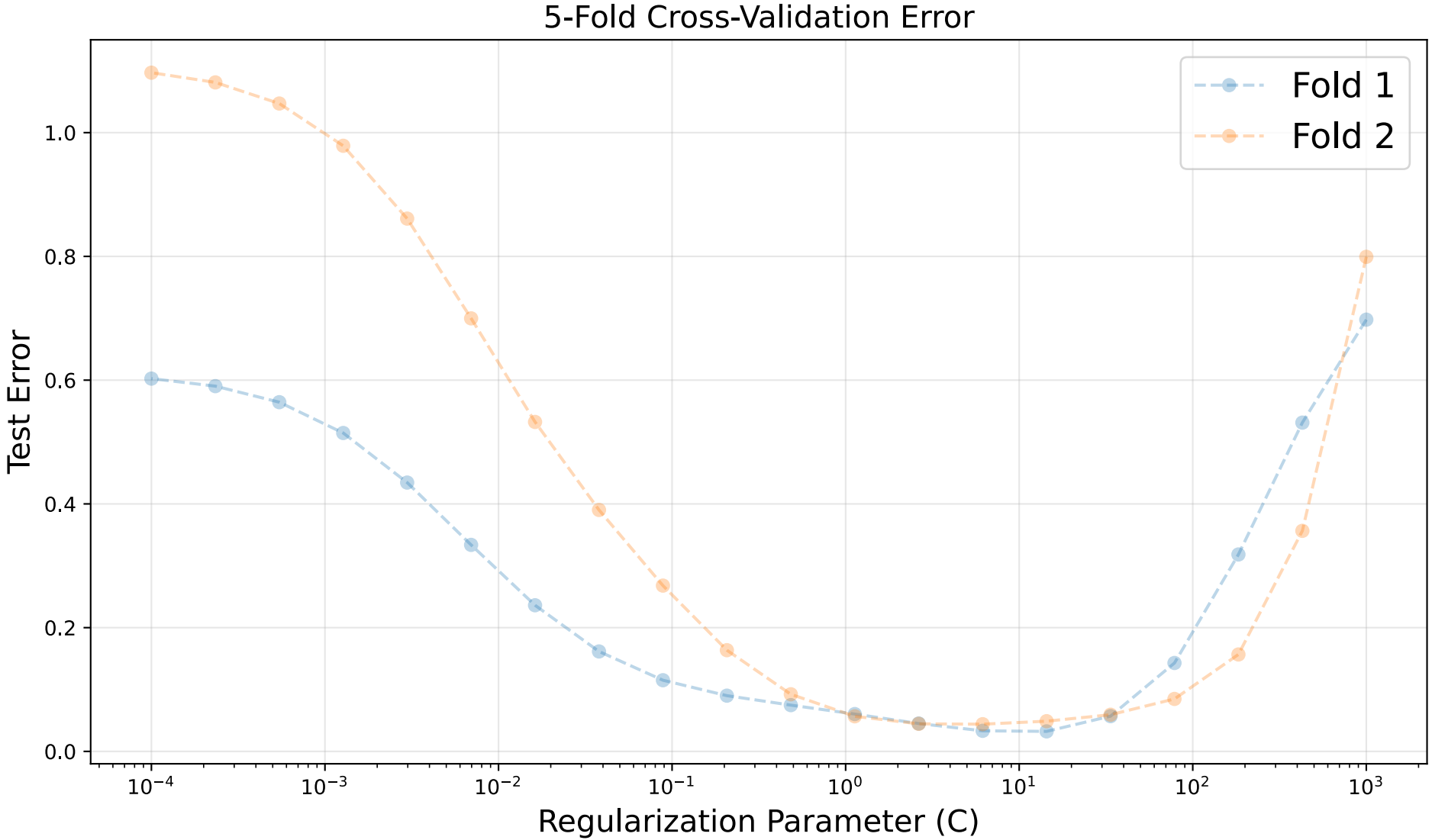


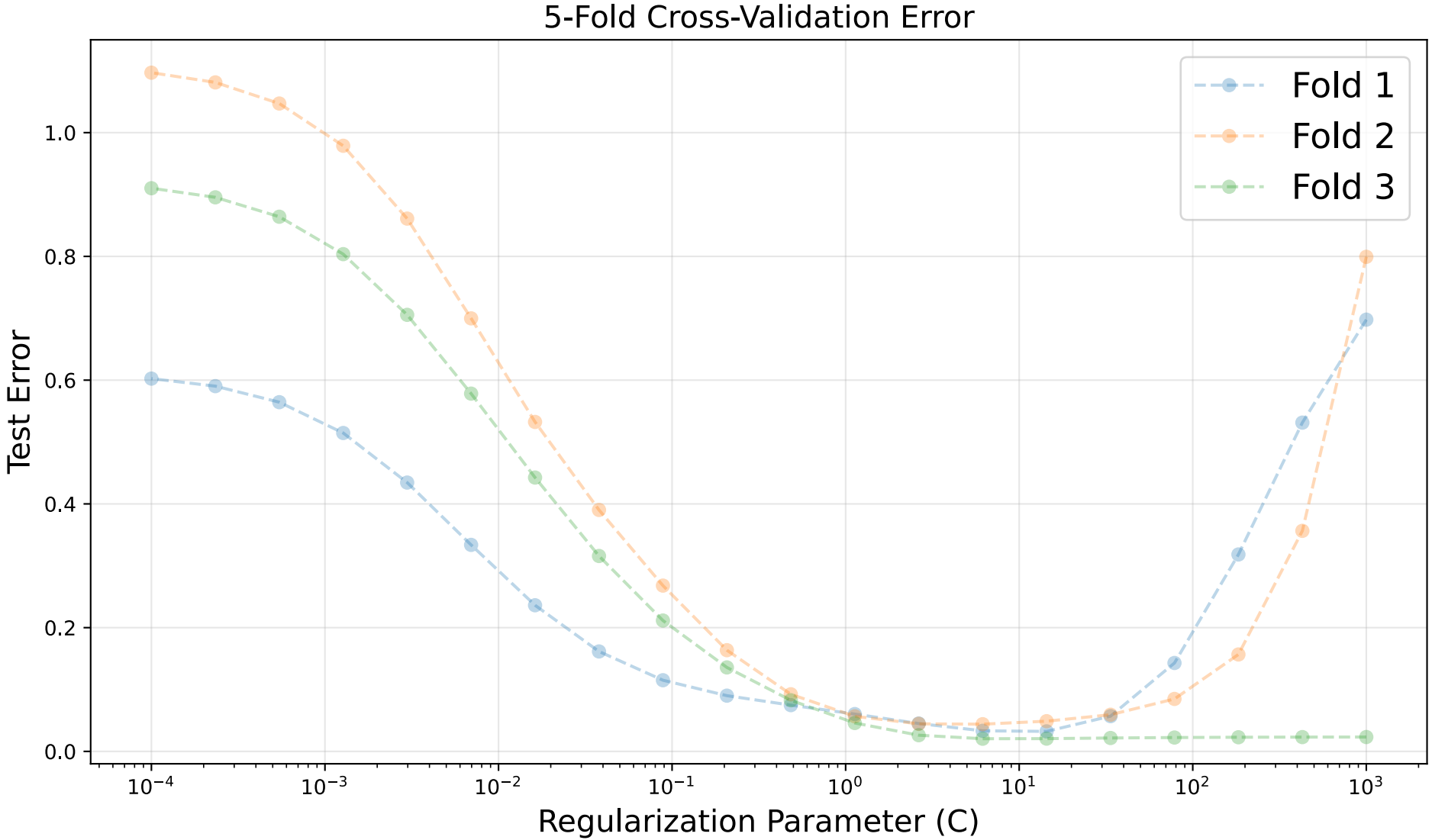
C’est l’erreur de validation croisée

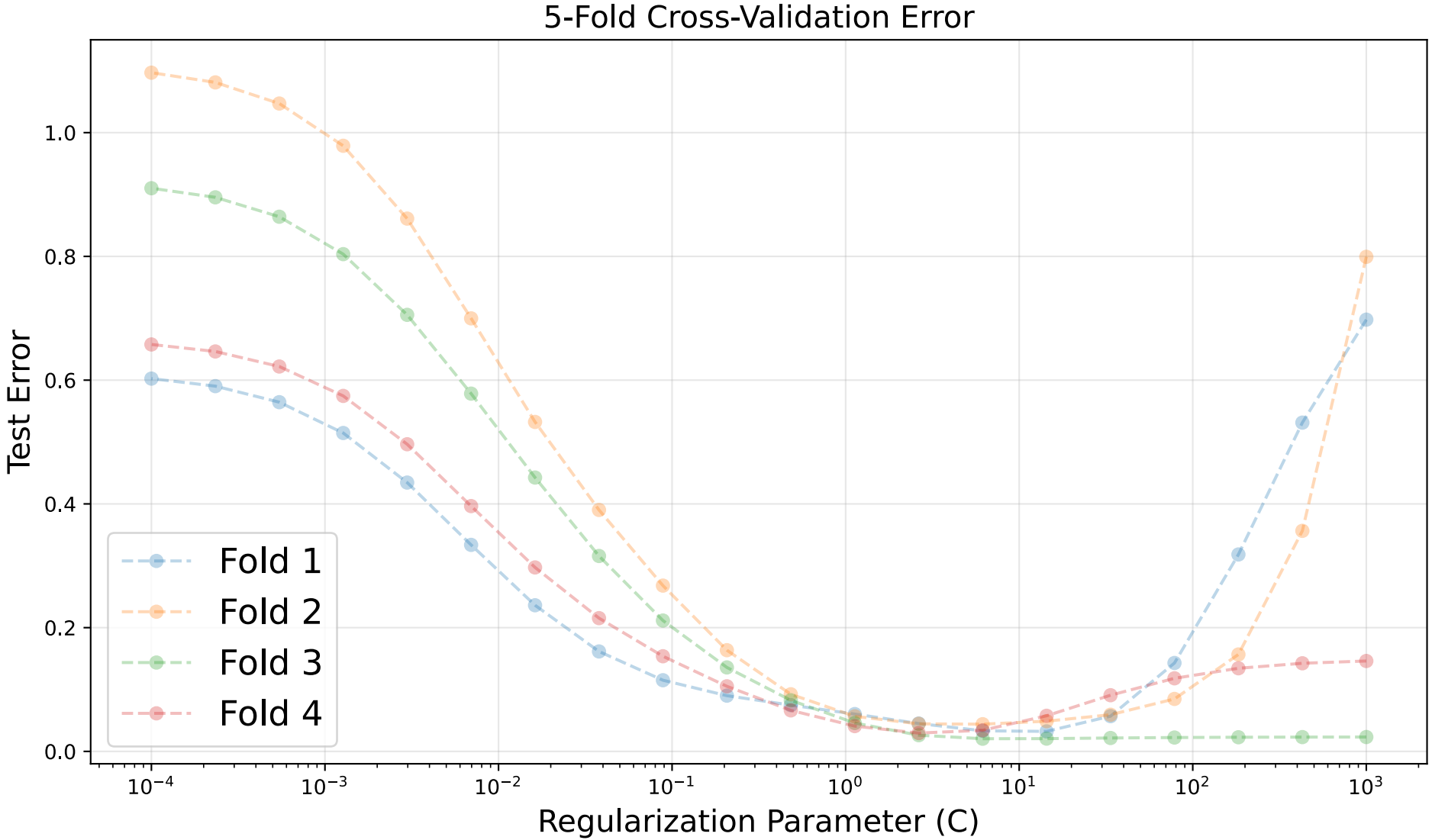
5-Fold cross validation

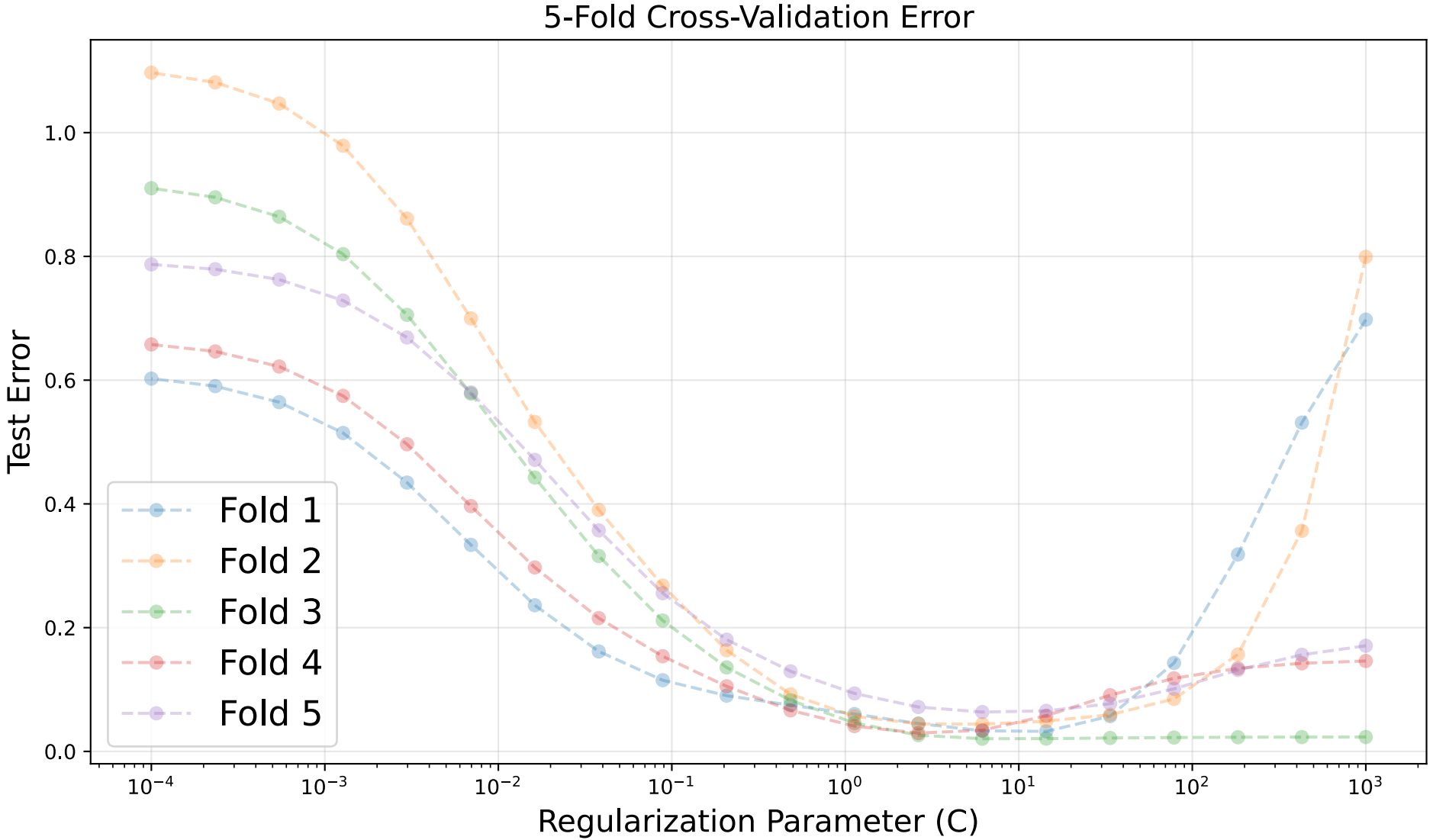
$$\mathbf{X}_{\text{test}}, \mathbf{Y}_{\text{test}} = \mathbf{X}_k, \mathbf{Y}_k \qquad \mathbf{X}_{\text{train}}, \mathbf{Y}_{\text{train}} = [\mathbf{X} \text{ sans } \mathbf{X}_k], \ldots, [\mathbf{Y} \text{ sans } \mathbf{Y}_k]$$

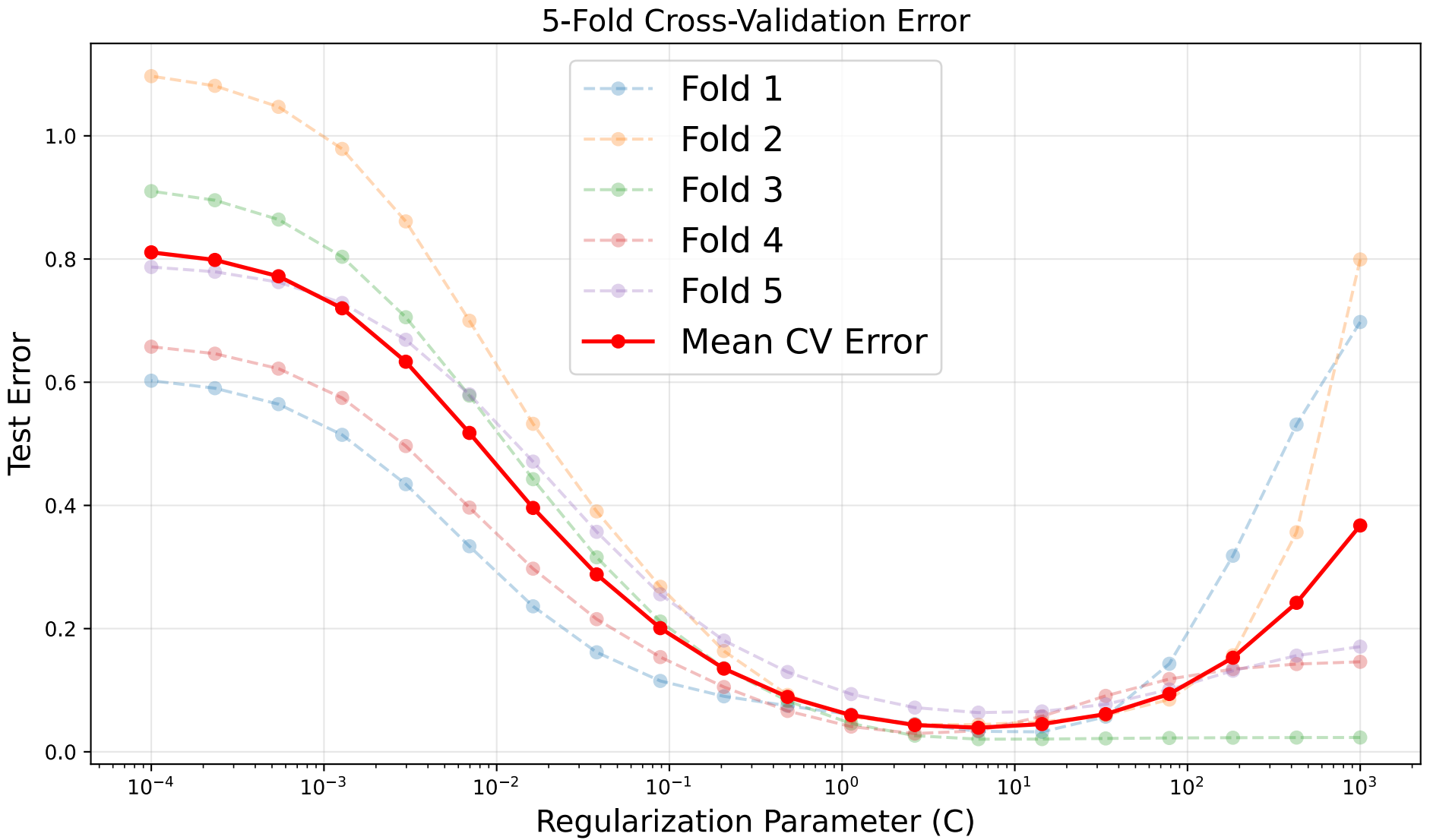

5-Fold Cross-Validation Error
Fold 1
Test Error
1.0
0.8
0.6
0.4
0.2
0.0
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
Regularization Parameter (C)


5-Fold Cross-Validation Error
Fold 1
Fold 2
Test Error
1.0
0.8
0.6
0.4
0.2
0.0
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
Regularization Parameter (C)


5-Fold Cross-Validation Error
Fold 1
Fold 2
Fold 3
Test Error
0.0
0.2
0.4
0.6
0.8
1.0
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
Regularization Parameter (C)


5-Fold Cross-Validation Error
Test Error
Regularization Parameter (C)
Fold 1
Fold 2
Fold 3
Fold 4
0.0
0.2
0.4
0.6
0.8
1.0
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}


5-Fold Cross-Validation Error
Test Error
Regularization Parameter (C)
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
0.0
0.2
0.4
0.6
0.8
1.0
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}

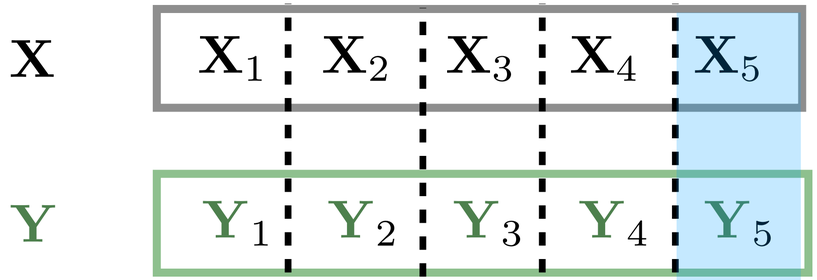

5-Fold Cross-Validation Error
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Mean CV Error
Test Error
Regularization Parameter (C)
0.0
0.2
0.4
0.6
0.8
1.0
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}

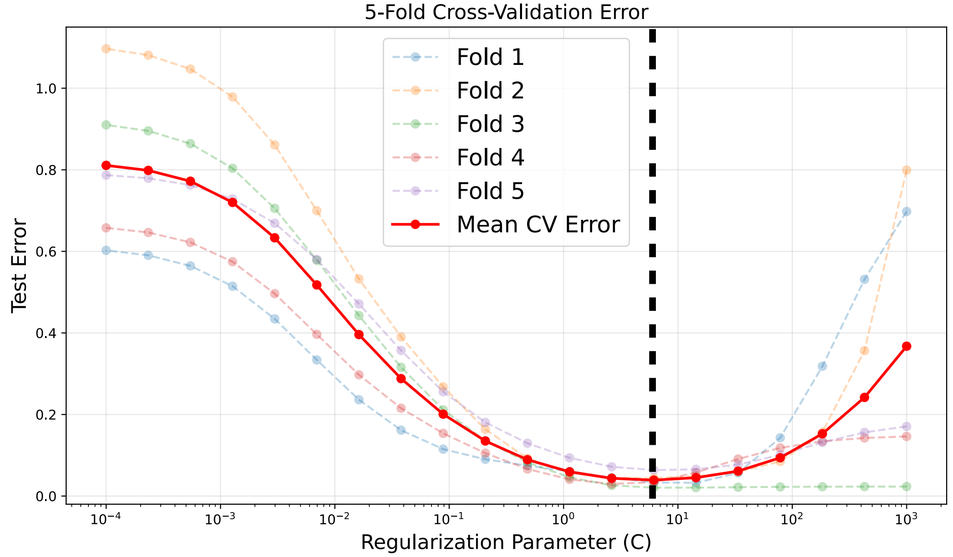
C optimal

Cross-validation

Idée: Effectuer plusieurs découpages et moyenner l'erreur de test

1. Couper le dataset en 5 parties (folds)





C optimal

2. Choisir une liste de valeurs de C, par ex: [0.01, 0.05, 0.1, 1., 10]

3. Pour chaque **k** in [1, 2, 3, 4, 5], créer un découpage train/test

$$\mathbf{X}_{\text{test}}, \mathbf{Y}_{\text{test}} = \mathbf{X}_k, \mathbf{Y}_k \qquad \mathbf{X}_{\text{train}}, \mathbf{Y}_{\text{train}} = [\mathbf{X} \text{ sans } \mathbf{X}_k], \ldots, [\mathbf{Y} \text{ sans } \mathbf{Y}_k]$$

Pour chaque C:

1. Optimiser sur $\mathbf{X}_{\text{train}}$ $\mathbf{Y}_{\text{train}}$ $\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \text{loss}(f_\theta(\mathbf{x}_i), y_i) + \frac{1}{C} \text{pénalité}(\theta)$

2. Évaluer l'erreur de prédiction sur $\mathbf{X}_{\text{test}}$ $\mathbf{Y}_{\text{test}}$

4. Pour chaque C, calculer l'erreur de prédiction moyenne

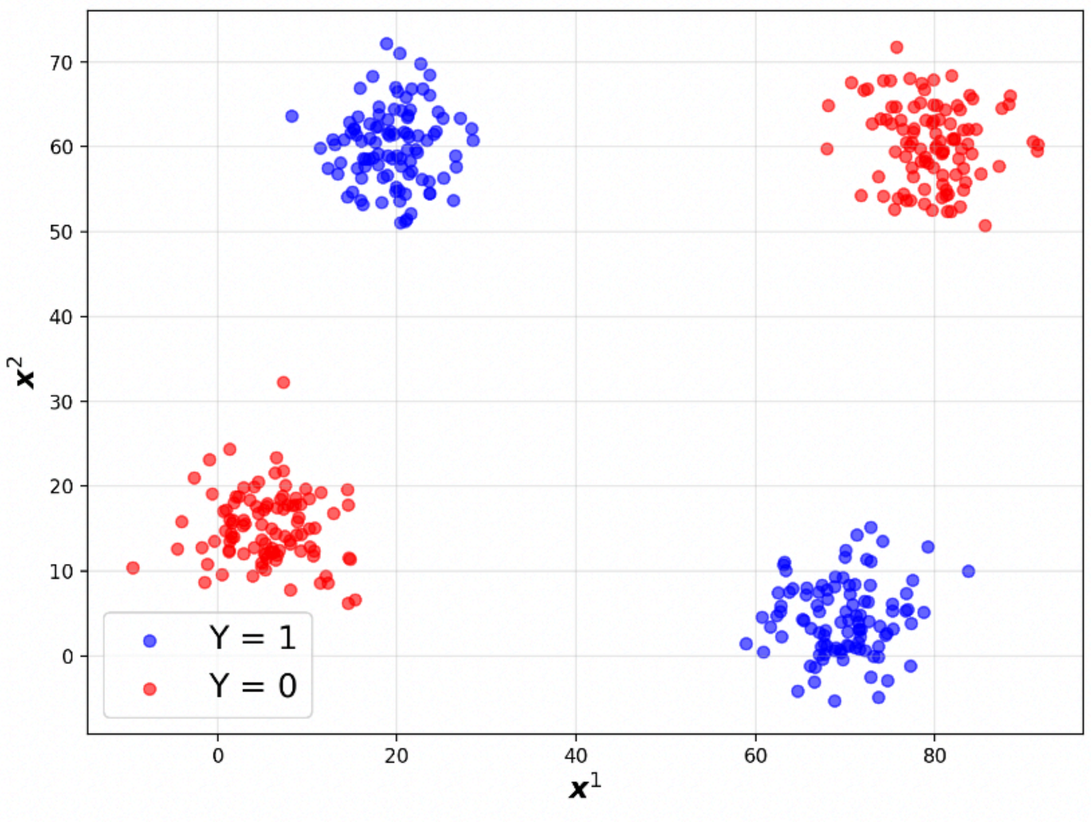
5. Choisir le C avec l'erreur de prédiction moyenne la plus petite

C'est l'erreur de validation croisée

5-Fold cross validation

INSEA

Et si les données ressemblent à ceci ?



INSEA