

Exercises

§11.4,11.7: Cross-Validation

Python: There is an optional starter script that supports Python-based problem(s) on this sheet. Use it as scaffolding: read the comments, complete the lines marked **TODO**, and compare the output with the mathematics requested in the problem(s).

Problem 1 A dataset contains $P = 100$ observations and uses $K = 5$ equal folds. (a) How many observations are in each validation fold? (b) How many observations train each fold model? (c) How many separate model fits are required for one hyperparameter value? (d) If 12 hyperparameter values are compared, how many model fits are required?

Problem 2 Suppose three folds have sizes 4, 4, 2 and fold MSE values 1, 2, 10. (a) Compute the unweighted average of the three fold MSE values. (b) Compute the pooled CV MSE $\sum_k (|F_k|/P)E_k$. (c) Explain why the two values differ. Which one gives every observation equal weight?

Problem 3 In 4-fold CV, an observation belongs to fold F_3 . Explain exactly which observations may be used to: (a) compute standardization means and scales for the prediction of this observation; (b) fit the model coefficients; (c) evaluate the loss for this observation. Why would computing one global standardizer before cross-validation be leakage?

Problem 4 A binary classification dataset contains 950 negative observations and 50 positive observations. Explain why stratified folds may be preferable to unconstrained random folds. If $K = 5$ and the classes can be divided evenly, approximately how many positive and negative observations should each fold contain?

Problem 5 A medical dataset contains four measurements from each patient. Explain why ordinary random row-wise K -fold cross-validation may overstate performance. Describe a safer fold construction. What should the basic observational group be?

Problem 6 A time series records monthly observations from 2010 through 2025. The goal is to predict future months. Why can ordinary randomly shuffled cross-validation be inappropriate? Describe the basic logic of a time-respecting validation scheme.

Problem 7 Use the running dataset:

```
x = np.array([
    [-3., -2., -1., 0.,
     1., 2., 3., 4.]
])

y = np.array([
    [9.2, 3.9, 1.1, 0.2,
     0.9, 4.1, 8.8, 16.2]
])
```

with random seed 0 and 4 folds. Use cross-validation to compare polynomial degrees 1,2,3,4. Report the four CV MSE values and the selected degree.

Problem 8 Explain the complete workflow for choosing a ridge parameter λ using K -fold