Model Assessment and Selection

FA590 Statistical Learning in Finance

Dr. Zonghao Yang

Stevens Institute of Technology

Learning Objectives

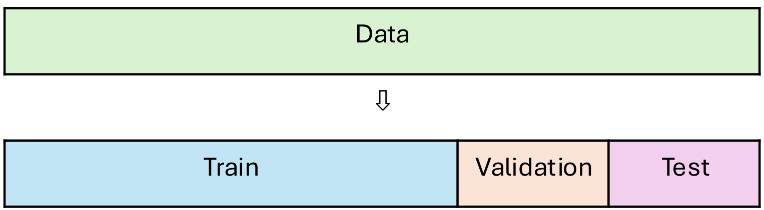
- Understand the key concepts and techniques used in model selection, including cross-validation methods
- Gain insight into handling temporal data in financial modeling, particularly expanding-window and rolling-window validation approaches
- Explore hyperparameter tuning and feature selection strategies to optimize model performance

Model Evaluation

- How the model really performs on unseen data
- Model evaluation can be used for different purposes
  - Model selection: Which model performs the best? How to determine the hyperparameters?
  - Feature selection: Which features should we include in the learning algorithms?
- To avoid overfitting, model evaluation is often performed on validation and test sets

Validation Set Approach

- Randomly divide the available set of observations into three parts: a training set, a validation set, and a test set
- The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set
- The hold-out test set can only be used after the model is finalized to test the true performance of the model
- The performance on the validation set can be used to select models and hyperparameters



Example: Ridge Regression

- Ridge regression

$$\hat{\beta}_0, \ldots, \hat{\beta}_p = \arg\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

  where λ is the hyperparameter for ridge regression
- Consider, for example, $\lambda \in \{0.0001, 0.001, 0.01, 0.05\}$
- For each specification, we fit the model using the training set, and evaluate the model using the validation set
- Suppose we use the R^2 as the evaluation metric; let us denote the R^2 of the validation set as $R^2_{val,\lambda=0.0001}, R^2_{val,\lambda=0.001}, \ldots$
- We select λ that optimizes the validation metric

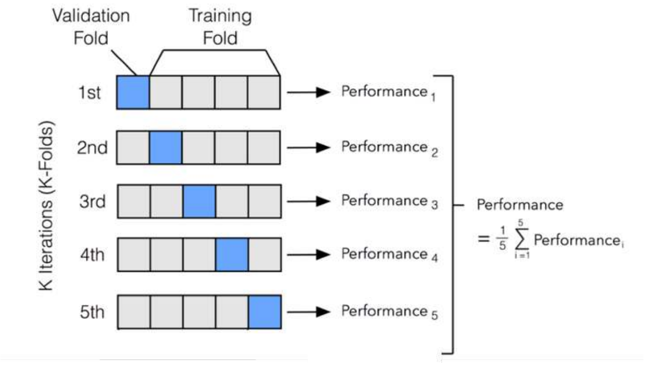
$$\lambda^* = \arg\max_{\lambda} R^2_{val,\lambda}$$

Drawbacks of the Validation Set Approach

- The validation estimate of the test error rate can be variable, depending on which observations are randomly chosen into the training/validation sets
- Only a subset of the available data are used for fitting the model

K-Fold Cross-Validation

- Rather than just keeping one part of the training data for validation, we can split the data into k folds
- We then train the model k times, each time by leaving out one of the folds for validation
- The final performance is the average performance across all k validation folds
- **K-Fold Cross-Validation**



Hyperparameter Selection: Example of Ridge Regression

- Ridge regression

$$\hat{\beta}_0, \ldots, \hat{\beta}_p = \arg\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Five-fold cross-validation: For each λ in $\{0.0001, 0.001, 0.01, 0.05\}$
  1. Use fold-1 as the validation set and the rest as the training set
  2. Denote the validation R^2 for fold-1 as $R^2_{fold-1,\lambda}$
  3. $\cdots$
  4. Use fold-5 as the validation set and the rest as the training set
  5. Denote the validation R^2 for fold-5 as $R^2_{fold-5,\lambda}$
  6. Calculate the cross-validation R^2 for λ: $R^2_{cv,\lambda} = \sum_{k=1}^{5} R^2_{fold-k,\lambda}/5$
- We select λ that optimizes the cross-validation metric

$$\lambda^* = \arg\max_{\lambda} R^2_{cv,\lambda}$$

Model Selection

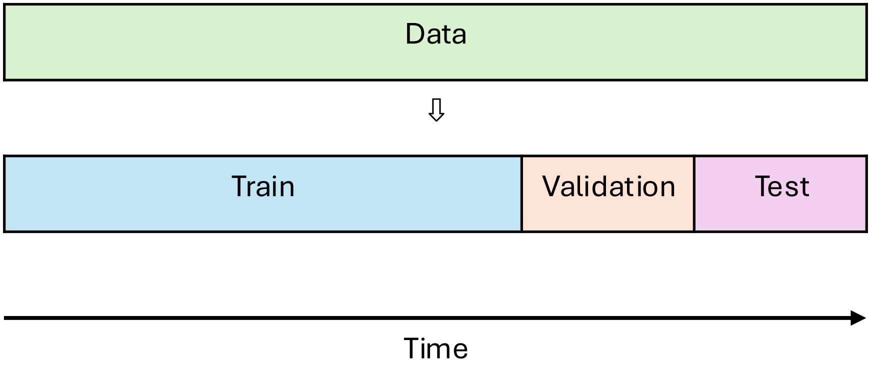
- Consider linear regression, decision tree, random forest, and neural network
- For each model $\mathcal{M}$
  1. Use fold-1 as the validation set and the rest as the training set
  2. Denote the validation R^2 for fold-1 as $R^2_{fold-1,\mathcal{M}}$
  3. $\cdots$
  4. Use fold-5 as the validation set and the rest as the training set
  5. Denote the validation R^2 for fold-5 as $R^2_{fold-5,\mathcal{M}}$
  6. Calculate the cross-validation R^2 for $\mathcal{M}$: $R^2_{cv,\mathcal{M}} = \sum_{k=1}^{5} R^2_{fold-k,\mathcal{M}}/5$
- We select $\mathcal{M}$ that optimizes the cross-validation metric

$$\mathcal{M}^* = \arg\max_{\mathcal{M}} R^2_{cv,\mathcal{M}}$$

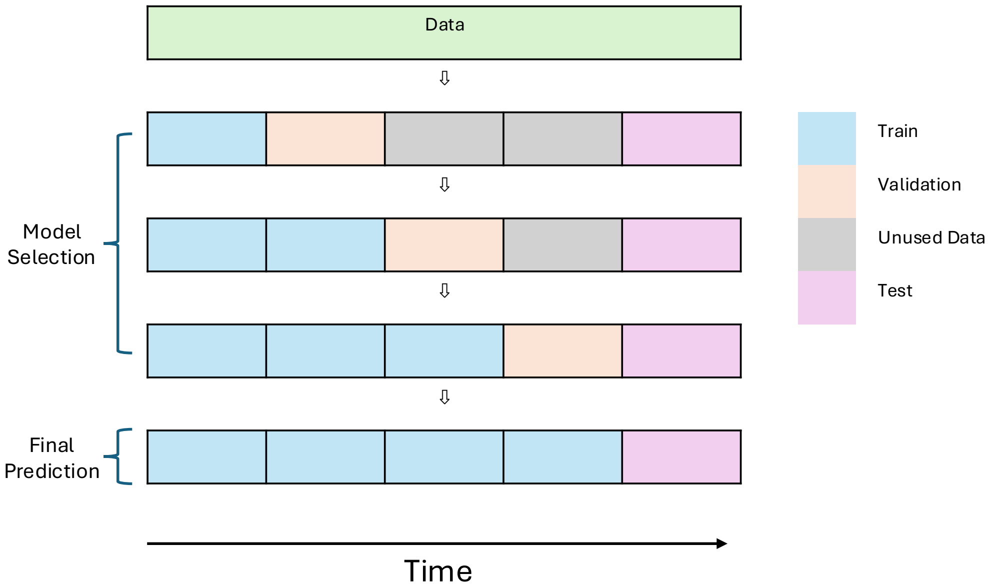
TEMPORAL DATA

- **Temporal Data** : data that is associated with time
- In Finance, we often encounter data with temporal information
  - Loan returns, stock prices, bond yields, etc.
- **Look-Ahead Bias** One cannot use future information to train and predict the past
- In this case, data splitting by randomly drawing samples for training, validation, and test is not appropriate
- Same idea: we need to partition the data in some way to test the model performance on unseen data
  - Simple time split validation
  - Expanding/rolling window approaches

Simple Time Split Validation



Expanding Window



Rolling Window