DA2011
Machine Learning I

Lecture 3

Ms. Naethree Premnath
05th October 2025

Today you will learn…

- Multiple Linear Regression
  - Uses of Multiple Linear Regression
  - Model Selection Metrics for Feature Selection
  - Best Subset Selection
  - Forward Selection
  - Backward Elimination
- Overfitting
- Model Validation

Multiple Linear Regression

- Extension of simple linear regression
- Predicts a continuous outcome (Y) using multiple predictors ($X_1,\ X_2, \dots, X_n$)
- Equation: $\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n + \varepsilon}$

  where $\boldsymbol{\beta_j}$: the average effect on Y of a one unit increase in $\boldsymbol{X_j}$, holding all other predictors fixed.
- The prediction model (fitted line) is: $\boldsymbol{\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta_1} X_1 + \widehat{\beta_2} X_2 + \widehat{\beta_3} X_3 + \cdots + \widehat{\beta_n} X_n}$
- Example: Predicting house price using size, age, rooms, location

Uses of Multiple Linear Regression

- Captures more complex relationships

E.g. A student’s exam score can be predicted not only by study hours but also by attendance, prior GPA, and sleep patterns.

- Improves prediction accuracy

E.g. Predicting house prices with just square footage may be inaccurate, but adding location, age, and number of rooms makes predictions more precise.

- Allows testing of relative importance of predictors

E.g. In predicting sales revenue, MLR can show whether advertising spend, product price, or customer satisfaction has the strongest impact.

Feature Selection

- Independent variables must be selected, maximizing the accuracy while minimizing the number of variables used. (Avoid overfitting & underfitting)

- To select the independent variables there are multiple approaches:

➢ Best subset selection
➢ Forward selection
➢ Backward selection

Model Selection Metrics for Feature Selection

1) Adjusted R^2

$$Adjusted\ R^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

2) AIC (Akaike Information Criterion)

$$AIC = -2\ln(L) + 2p$$

3) BIC (Bayesian Information Criterion)

$$\mathrm{B}IC = -2\ln(L) + pln(n)$$

n = sample size
p = number of parameters
L = Likelihood of the model

Problems Without Feature Selection

- Multiple regression may overfit, capturing noise instead of true patterns.

E.g. Predicting exam scores by including irrelevant variables like students’ favorite color.

- Multicollinearity can occur producing unstable coefficients and inflated standard errors.

E.g. Using both ‘house size’ and ‘number of rooms’ which provide nearly the same information.

- The model may suffer from reduced interpretability, making it harder to explain results.

E.g. A model with 20+ predictors is difficult to interpret compared to one with only 3-4 key variables.

- Including irrelevant predictors increases computational inefficiency without improving accuracy.

E.g. Adding weather data when predicting house prices does not meaningfully improve the model.

Best Subset Selection

- Best subset selection involves fitting **all possible models** using different subsets of predictors and comparing them.

- For p predictors, the no.of possible subsets is: 2^p

1) Consider μ_0: null model (no predictors)

2) For k= 1,2….p , fit all $\binom{p}{k}$ that contain k predictors.

3) Pick the best (lowest RSS or largest R^2) among $\binom{p}{k}$ models and call it μ_k.

4) Select a single best model from among $\mu_0, \mu_{1,\ldots..}\mu_p$ using test MSE or any appropriate model selection metric.

Best Subset Selection Example

Let $p = 3$ (3 predictors)

Fit models for 0 predictors(μ_0),1 predictor(μ_1), 2 predictors(μ_2), 3 predictors(μ_3)

1) $p = 0$

$$\boldsymbol{Y} = \boldsymbol{\beta_0} + \boldsymbol{\varepsilon} \quad \text{--}\ \mu_0$$

2) $p = 1$

$$\boldsymbol{Y} = \boldsymbol{\beta_0}^{(1)} + \boldsymbol{\beta_1}^{(1)}\boldsymbol{X_1} + \boldsymbol{\varepsilon}^{(1)}$$
$$\boldsymbol{Y} = \boldsymbol{\beta_0}^{(2)} + \boldsymbol{\beta_1}^{(2)}\boldsymbol{X_2} + \boldsymbol{\varepsilon}^{(2)}$$
$$\boldsymbol{Y} = \boldsymbol{\beta_0}^{(3)} + \boldsymbol{\beta_1}^{(3)}\boldsymbol{X_3} + \boldsymbol{\varepsilon}^{(3)}$$

Assume the below is the best model with 1 variable:

$$\boldsymbol{Y} = \boldsymbol{\beta_0}^{(2)} + \boldsymbol{\beta_1}^{(2)}\boldsymbol{X_2} + \boldsymbol{\varepsilon}^{(2)} \quad \text{--}\ \mu_1$$

Best Subset Selection Example

3) $p = 2$

$$\boldsymbol{Y} = \boldsymbol{\beta_0}^{(1)} + \boldsymbol{\beta_1}^{(1)}\boldsymbol{X_1} + \boldsymbol{\beta_2}^{(1)}\boldsymbol{X_2} + \boldsymbol{\varepsilon}^{(1)}$$
$$\boldsymbol{Y} = \boldsymbol{\beta_0}^{(2)} + \boldsymbol{\beta_1}^{(2)}\boldsymbol{X_1} + \boldsymbol{\beta_2}^{(2)}\boldsymbol{X_3} + \boldsymbol{\varepsilon}^{(2)}$$
$$\boldsymbol{Y} = \boldsymbol{\beta_0}^{(3)} + \boldsymbol{\beta_1}^{(3)}\boldsymbol{X_2} + \boldsymbol{\beta_1}^{(3)}\boldsymbol{X_3} + \boldsymbol{\varepsilon}^{(3)}$$

Assume the below is the best model with 2 variables:

$$\boldsymbol{Y} = \boldsymbol{\beta_0}^{(3)} + \boldsymbol{\beta_1}^{(3)}\boldsymbol{X_2} + \boldsymbol{\beta_1}^{(3)}\boldsymbol{X_3} + \boldsymbol{\varepsilon}^{(3)} \quad \text{--}\ \mu_2$$

4) $p = 3$

$$\boldsymbol{Y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\boldsymbol{X_1} + \boldsymbol{\beta_2}\boldsymbol{X_2} + \boldsymbol{\beta_3}\,\boldsymbol{X_3} + \boldsymbol{\varepsilon} \quad \text{--}\ \mu_3$$

Best Subset Selection Example

- Out of the best models $\mu_0, \mu_1, \mu_2, \mu_{3,}$ select the best.

- Can R^2 be used to compare between the models $\mu_0, \mu_1, \mu_2, \mu_{3,}$ to choose the final best model?

➢ No, because the model containing the largest no.of parameters (μ_3) will have the largest R^2.

Hence test MSE or any of the following metrics can be used

✓ AIC
✓ BIC
✓ Adjusted R^2
✓ Cross validation error

Forward Selection

- Instead of considering all 2^p models, forward selection starts with nothing and **adds predictors one at a time** based on which improves the model the most.

E.g. Consider p=3 (3 predictors)

1) Start with the null model

$$\boldsymbol{Y = \beta_0 + \varepsilon}$$

2) Add the best single predictor
Fit 3 models with 1 predictor each:

$$\boldsymbol{Y = {\beta_0}^{(1)} + {\beta_1}^{(1)} X_1 + \varepsilon^{(1)}}$$
$$\boldsymbol{Y = {\beta_0}^{(2)} + {\beta_1}^{(2)} X_2 + \varepsilon^{(2)}}$$
$$\boldsymbol{Y = {\beta_0}^{(3)} + {\beta_1}^{(3)} X_3 + \varepsilon^{(3)}}$$

Pick the one with the best performance (highest adjusted R^2). Call this μ_1

Forward Selection

3) Add the next predictor.
Suppose $\boldsymbol{X_2}$ was chosen first. Now test adding each of the remaining predictors:

$$\boldsymbol{Y = \beta_0{}^{(1)} + \beta_2{}^{(1)}X_2 + \beta_1{}^{(1)}X_1 + \varepsilon^{(1)}}$$
$$\boldsymbol{Y = \beta_0{}^{(2)} + \beta_2{}^{(2)}X_2 + \beta_3{}^{(2)}X_3 + \varepsilon^{(2)}}$$

Pick whichever improves the model most.
That becomes μ_2

4) Add the last predictor.
Finally, test whether adding the last remaining predictor improves the model enough (based on your stopping criterion).
If yes, full model: $\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3\, X_3 + \varepsilon}$ --- μ_3
If not, stop at μ_2.

Backward Elimination

Backward Elimination starts with the **full model** and removes predictors one by one.

1) Start with the full model

$$\boldsymbol{Y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \boldsymbol{X_1} + \boldsymbol{\beta_2} \boldsymbol{X_2} + \boldsymbol{\beta_3}\, \boldsymbol{X_3} + \boldsymbol{\varepsilon} \qquad \text{---}\ \mu_3$$

2) Remove the least useful predictor

Fit the 3 possible **2-predictor models** (each one drops one variable):

$$\boldsymbol{Y} = \boldsymbol{\beta_0}^{(1)} + \boldsymbol{\beta_1}^{(1)} \boldsymbol{X_1} + \boldsymbol{\beta_2}^{(1)} \boldsymbol{X_2} + \boldsymbol{\varepsilon}^{(1)}$$
$$\boldsymbol{Y} = \boldsymbol{\beta_0}^{(2)} + \boldsymbol{\beta_1}^{(2)} \boldsymbol{X_1} + \boldsymbol{\beta_3}^{(2)} \boldsymbol{X_3} + \boldsymbol{\varepsilon}^{(2)}$$
$$\boldsymbol{Y} = \boldsymbol{\beta_0}^{(3)} + \boldsymbol{\beta_2}^{(3)} \boldsymbol{X_2} + \boldsymbol{\beta_3}^{(3)} \boldsymbol{X_3} + \boldsymbol{\varepsilon}^{(3)}$$

Pick the one with the best performance (highest adjusted R^2). Call this μ_2

Backward Elimination

3) Remove another predictor

Assume that $\boldsymbol{Y} = \boldsymbol{\beta_0}^{(\mathbf{1})} + \boldsymbol{\beta_1}^{(\mathbf{1})}\boldsymbol{X_1} + \boldsymbol{\beta_2}^{(\mathbf{1})}\boldsymbol{X_2} + \boldsymbol{\varepsilon}^{(\mathbf{1})}$ from previous step was μ_2.
Fit the 2 possible **1-predictor models**:

$$\boldsymbol{Y} = \boldsymbol{\beta_0}^{(\mathbf{1})} + \boldsymbol{\beta_1}^{(\mathbf{1})}\boldsymbol{X_1} + \boldsymbol{\varepsilon}^{(\mathbf{1})}$$
$$\boldsymbol{Y} = \boldsymbol{\beta_0}^{(\mathbf{2})} + \boldsymbol{\beta_2}^{(\mathbf{2})}\boldsymbol{X_2} + \boldsymbol{\varepsilon}^{(\mathbf{2})}$$

Pick the one with the best performance (highest adjusted R^2). Call this μ_1.

4) Stop when appropriate.
You can continue until you reach the null model. Usually, you stop removing predictors when further removals make the model worse.

Feature Selection Methods Comparison

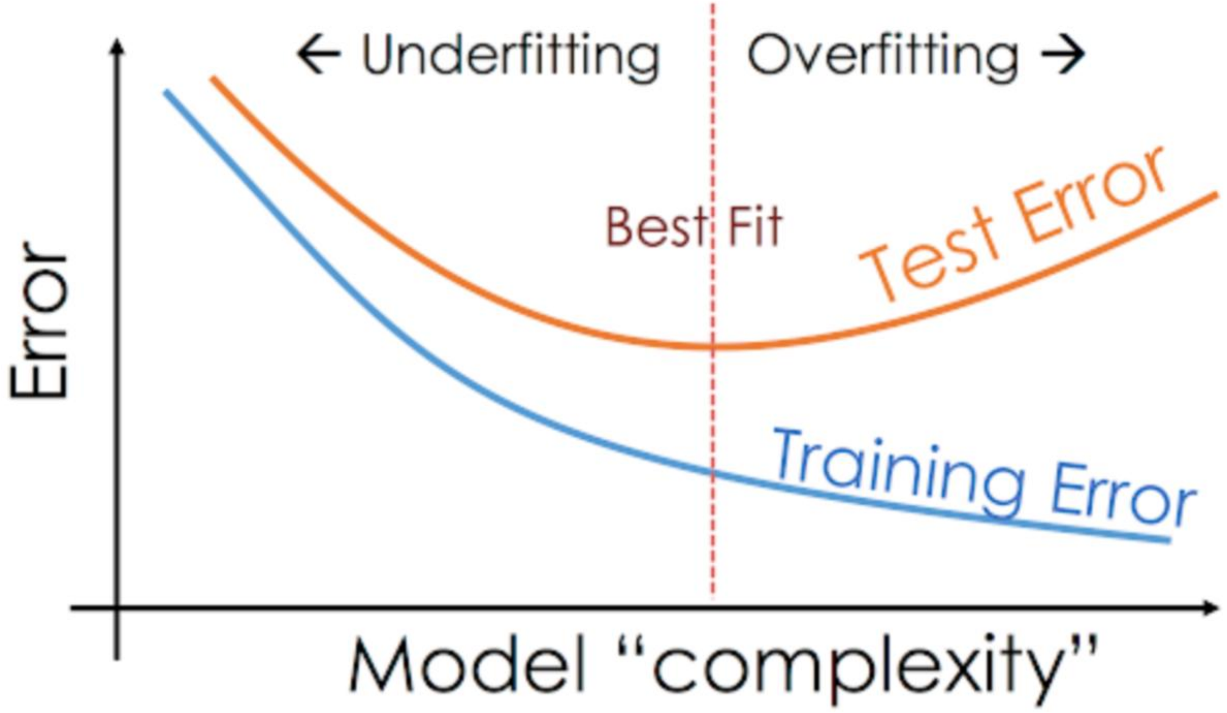
Breakout Room Group Activity

1. Compare between the three selection methods in your breakout rooms.
2. Discuss advantages & disadvantages of each method.
3. Fill the table given in this sheet.

(10 mins)

Overfitting

Overfitting occurs when the machine learning model gives accurate predictions for training data but not for new data. Hence, predictions of overfitting models cannot be trusted.



Reasons for Overfitting

- **Too Many Features (High Dimensionality)**

E.g. Using 100 predictors for only 200 observations in multiple regression.

- **Model Complexity**

E.g. Adding unnecessary higher-order terms in polynomial regression.

- **Small Training Dataset**

E.g. Fitting a regression line on only 20 data points with 10 predictors.

- **Too Many Parameters Relative to Observations**

E.g. Logistic regression with 50 predictors but only 100 samples. Leads to unstable estimates and high variance.

- **Noise in Data**

E.g. Outliers or random measurement errors. Complex model tries to explain noise which reduces accuracy on clean data.

- **Insufficient Regularization**

E.g. Using plain linear regression when ridge or lasso would control large coefficients.

- **Improper Validation**

E.g. Evaluating only on training set instead of cross-validation. Model looks ‘perfect’ on training but collapses on test data.

Model Validation

Goal: Ensure the model performs well on new, unseen data.
Prevents overfitting (good fit to training data but poor generalization).

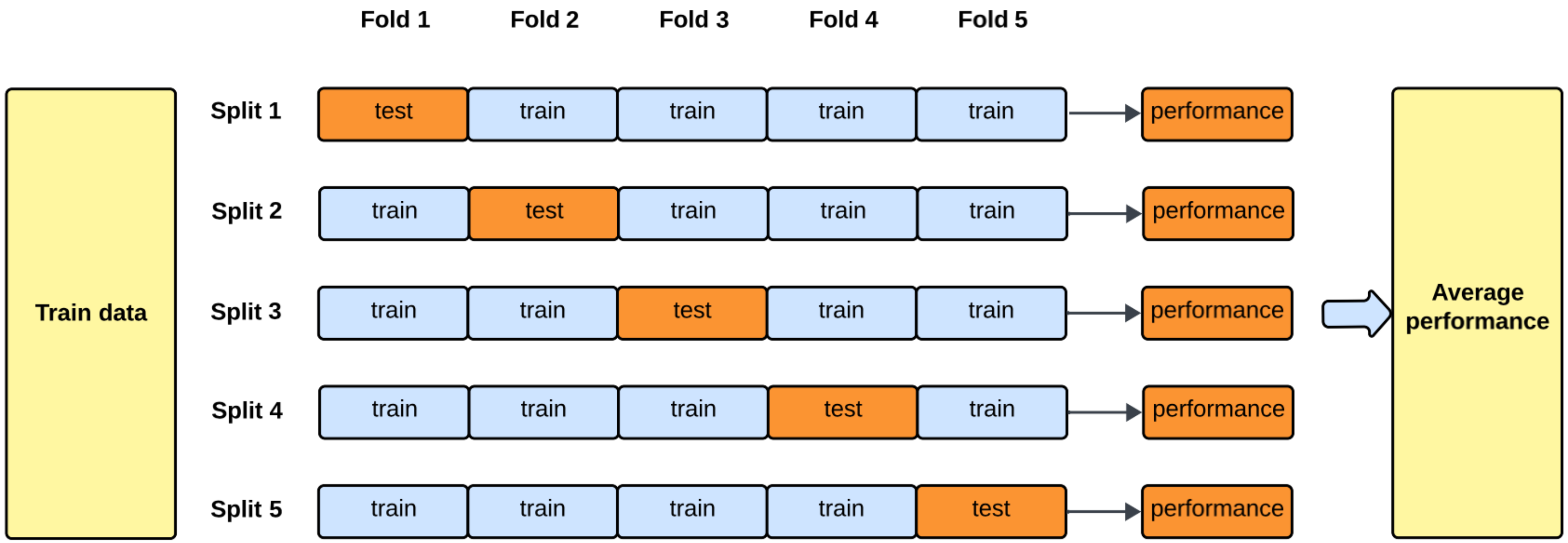
1. Cross-Validation (CV)

- Split the training data into k roughly equal parts (folds).
- Hold out 1 fold as test set, train on the remaining k-1.
- Compute prediction error on the test fold.
- Repeat for each fold.
- Take the average error across all folds.

- Common choices:

➢ k = 5 or 10
➢ LOOCV (Leave-One-Out CV): each observation is its own test set

Cross Validation



Train data

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	
Split 1	test	train	train	train	train	performance
Split 2	train	test	train	train	train	performance
Split 3	train	train	test	train	train	performance
Split 4	train	train	train	test	train	performance
Split 5	train	train	train	train	test	performance

Average performance

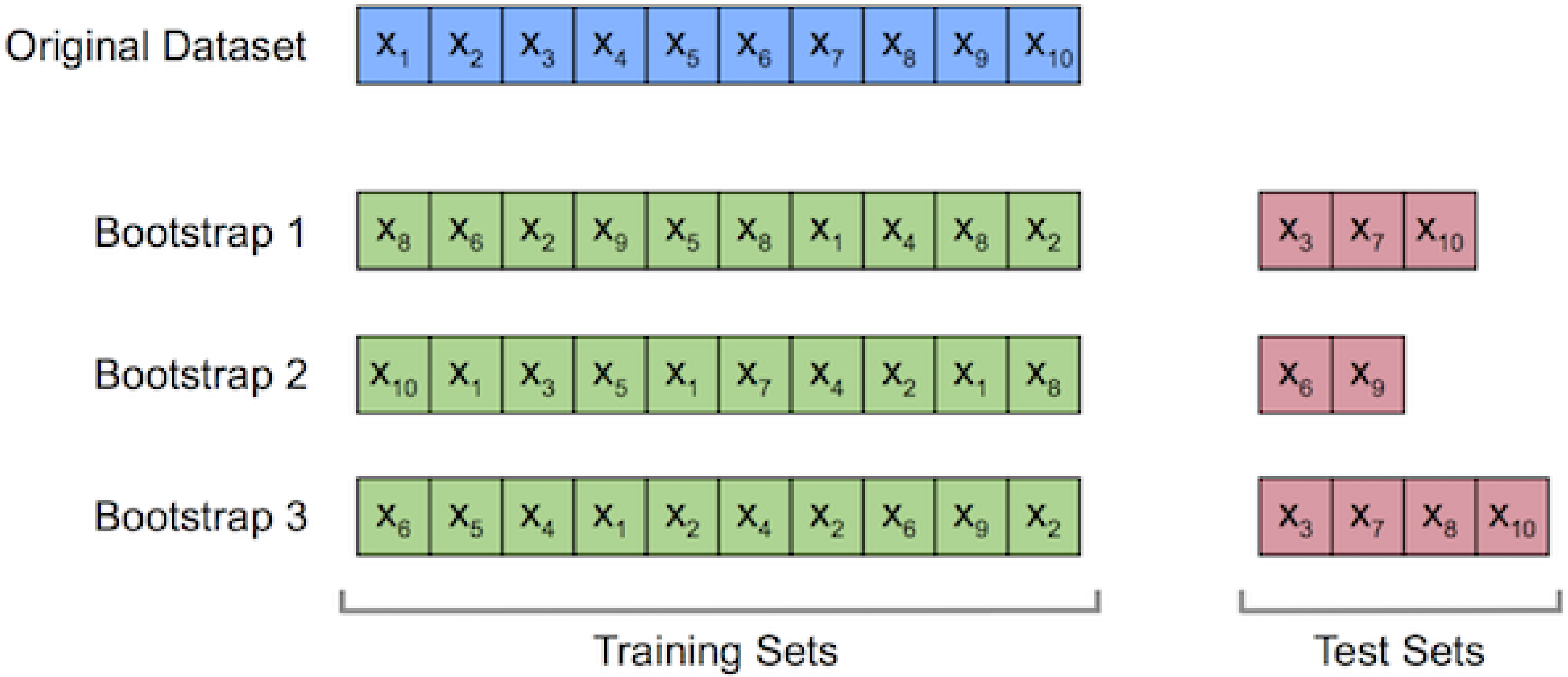
Model Validation

2. Bootstrapping

- Repeatedly draw samples with replacement from the original dataset.
- For each bootstrap sample, train the model.
- Evaluate its performance.
- Repeat B times (e.g., 500 or 1000 samples).

- Use results to estimate:
  - ➢ Prediction error
  - ➢ Stability of coefficients (variance of estimates)

Bootstrapping



	Training Sets	Test Sets
Original Dataset	x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10}	
Bootstrap 1	x_8 x_6 x_2 x_9 x_5 x_8 x_1 x_4 x_8 x_2	x_3 x_7 x_{10}
Bootstrap 2	x_{10} x_1 x_3 x_5 x_1 x_7 x_4 x_2 x_1 x_8	x_6 x_9
Bootstrap 3	x_6 x_5 x_4 x_1 x_2 x_4 x_2 x_6 x_9 x_2	x_3 x_7 x_8 x_{10}

Thank you
See you next week!