SMART Program Proposal

Namhwa Lee

Statistical Learning

- Y: A quantitative response, and $X_1, \ldots, X_p$: p different predictors.
- Assume some relationship between Y and $X = (X_1, \ldots, X_p)$:

$$Y = f(X) + \epsilon$$

where f is some fixed but unknown function of X, and ϵ is a random error, independent of X and has mean zero.

- Goal: Estimate f based on the observed data ($\hat{f}$: Estimate of f).

Inference v.s. Prediction

- Why estimate f?
- `Inference`: Understand the way that Y is affected as X change.
  - We need to know the exact form of $\hat{f}$
  - Should not be a black box.
- `Prediction`: Accurately predict new or future data based on a given or past dataset.
  - Do not need to know the exact form of $\hat{f}$ if it yields accurate predictions for Y.
  - Can be a black box.

Prediction Accuracy

- $\hat{Y}$: Predicted value of Y.
- The prediction accuracy of $\hat{Y}$ depends on:
  - Reducible error: Can be reduced by improving the accuracy of $\hat{f}$.
  - Irreducible error: Due to the random error ϵ which cannot be predicted by X.

$$\begin{aligned} E(Y-\hat{Y})^2 &= E\big[f(X)+\epsilon-\hat{f}(X)\big]^2 \\ &= \big[f(X)-\hat{f}(X)\big]^2 + Var(\epsilon) \end{aligned}$$

Prediction Accuracy v.s. Interpretability

- Flexible model: More accurate prediction but difficult to interpret.
- Restrictive model: Easy to interpret but less accurate prediction.
- Trade-off relationship between prediction accuracy and interpretability
- If our goal is inference → Restrictive model

  If our goal is prediction → Flexible model

  (Not always true because flexible models might have overfitting problem).

Assessing Model Prediction Accuracy

- We might have multiple models for prediction, and want to choose the best model.
- How to evaluate the performance of models with respect to prediction accuracy?

  → If the model has better prediction of Y, it will predict better.
- How to select the best model?

  → (Regression) Smaller mean squared error (MSE)

  → (Classification) Smaller misclassification rate

Project Overview

Topic: Statistical Data Mining with R.

Description: In statistical data analysis, prediction and inference are both based on data, but they have different purposes:

- `Inference`: The primary focus is on interpreting the model we fit.
- `Prediction`: The primary goal is to accurately predict new or future data based on a given or past dataset.

Throughout the program, we will learn how to employ statistical data mining methods (e.g. regression, and tree-based techniques) using R. Additionally, we will learn how to evaluate or assess the model performance based on its prediction error using data splitting method. Finally, if time allows, we will do a small data analysis project to apply what we have learned during the program.

Project Overview

Prerequisites: STAT 170A/B

Reference: James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.

https://link.springer.com/content/pdf/10.1007/978-1-0716-1418-1.pdf

Project Goal

1. **Distinguish Between Inference and Prediction:**
   - Develop a clear understanding of the difference between inference and prediction.
2. **Model Assessment and Selection for Prediction Model:**
   - Learn a technique/criteria for assessing and selecting models when the goal is to predict new outcomes.
3. **Application of Regression and Tree-Based Methods in R:**
   - Develop data mining skills by employing regression (linear/logistic, etc.) and tree-based methods (regression/classification tree) using R.

Project Goal

4. **Evaluation of Prediction Methods:**
   - Evaluate the performance of each method by calculating prediction errors or mis-classification rates.
5. **Data Analysis Project:**
   - Apply multiple prediction methods in a real data, and determine the final model.

Tentative Weekly Schedule

Week1: Initial Meeting.Topic Introduction (Prediction v.s. Inference)

Week2: Model Assessment for Prediction Model. Data-Splitting Method

Week3: Linear / Logistic Regression for Prediction

Week4: Tree-based Methods for Prediction

Week5: Application to Real Data.

Week6: Progress Report1

Week7: Progress Report2

Week8: Finalize the Project and Report.

Week9: Presentation Prep.

Week10: Presentation