7. Predictive Modelling Procedure

Statistical Modelling & Machine Learning

Jaejik Kim

Department of Statistics
Sungkyunkwan University

STA3036

Predictive Modelling Procedure

1. Data pre-processing:
    - Imputation for missing values.
    - Create important input variables (as many as possible).
    - If you miss important X variables, there will be no chance to have good predictive models.
    - Transformation of variables.

2. Exploring data:
    - Cluster analysis: Groups of data, properties of groups.
    - Dimension reduction: Visualization of data.

3. Filtering variables:
    - If p is very large (relatively to n), filter input variables in terms of both linearity and nonlinearity.

Predictive Modelling Procedure

4. Predictive modelling (suggestion):

    - If your main goal is interpretation,
        - Consider linear models with variable selection (e.g., linear regression, logistic regression, etc.)
        - If p is large, you can consider penalization methods such as lasso, SCAD, and MCP, etc.
        - If you want to consider nonlinear relationships, try the generalized additive model (GAM) to identify the functional relationship between Y and individual input variables. Based on the GAM, try to build a parametric regression model.

    - If your main goal is prediction,
        - If you have no idea about your data, consider complex models such as boosting, random forests, and SVM, etc.
        - If you have no idea about the best set of important input variables, the random forests is recommended.

Predictive Modelling Procedure

4. Predictive modelling (suggestion):
    - If your main goal is prediction, (continued)
        - For the most models, it is very important to reduce the number of dimensions through variable selection or dimension reduction techniques.
        - If you have information about your data and you are familiar with statistical modelling, try to build a model using the data modelling techniques (Remind ARGO!).
        - Find the best model.

5. Summary and conclusion from your model.
    - Use appropriate tables and graphs!
    - Effective visualization.