Project in Data Intensive Systems

4DV652

Lab Lecture 5

Welf Löwe

Agenda

- Tree-based regression and classification
- Lab 5 task descriptions

Regression

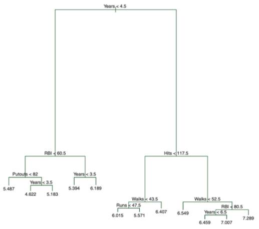
- Response Y as a function of predictors $X_1, \ldots, X_p$

$$f(X): X_1 \times \cdots \times X_p \to Y$$

  - X_i are possibly linear or nonlinear terms, predictors are quantitative variables, dummy variables for each class of the categorical predictors
- Learning:
  - estimate $f(X)$ such that $\hat{f}(X) = \hat{y}$ is a good prediction of Y, i.e., low *MSE*, given predictor values X=[$x_1, \ldots, x_p$]
- Predicting:
  - apply $\hat{y} = \hat{f}(x_1, \ldots, x_p)$

Tree-based Regression

- Learning: Splits the predictor space into regions by maximizing the information gain in the training data
  - No special treatment of class predictors necessary
  - Use direct or indirectly computed predictors
- Predicting: the average values of the observed training data in a region



Classification

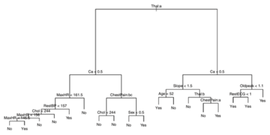
- Class Y as a function of predictors $X_1, \ldots, X_p$

$$f(X): X_1 \times \cdots \times X_p \to Y$$

  - X_i are quantitative or categorical predictors
- Learning:
  - estimate $f(X)$ such that $\hat{f}(X) = \hat{y}$ is a good prediction of Y, i.e., low *error rate*, high *precision/recall* , given predictor values X=[$x_1, \ldots, x_p$]
- Predicting:
  - apply $\hat{y} = \hat{f}(x_1, \ldots, x_p)$

Tree-based Classification

- Learning: Splits the predictor space into regions by maximizing the information gain in the training data
  - No special treatment of class predictors necessary
  - Use direct or indirectly computed predictors
- Classifying: the most frequently occurring class (mode) in training data in a region



Random Forrest

- Generalization of trees-based regression and classification
- Learning: Create several trees
  - using bagging: bootstrapping and learning tree models for each set
  - using a subsets of predictors for each of these models
  - both helps creating models with more or less (un-) correlated errors
- Predict: ensemble technique
  - Regression: the mean of the individual tree's predictions
  - Classification: the mode of the individual tree's classes
- Resampling technique that protects against overfitting

7

Tree Boosting

- Generalization of trees-based regression and classification
- Learning:
  - Regression: Create a series of trees (using bagging) each predicting the residuals from the sum of the shrunken (with weight $\lambda \in [0,1]$) previous trees
  - Classification: Create a series of trees (using bagging) each grown based on weighted training errors where weights are calculated based on the error of the previous trees
- Predict: pipeline technique
  - Predict the weighted (with weight λ) sum of the individual tree's predictions
  - Classify by the sign of the weighted sum of the individual tree's $(-1, 1)$ predictions
- Yet another resampling technique that protects against overfitting

8

Agenda

- Tree-based regression and classification
- Lab 5 task descriptions

9

Lab assignment 5: tree-based approaches for regression and classification

- ML
  - Challenge the current champion regression with a tree-based approach
  - Challenge the current champion classification with a tree-based approach
- Software development
  - If applicable, implement and deploy the new champion regression and classification
- Reporting:
  - In a fifth notebook, document the iteration(s) over the ML process steps
- Deadline: 2023-03-01

10