Machine Learning and Applied Econometrics

Tree-Based Models

Machine Learning and Econometrics

- This introductory lecture is based on
  - Kevin P. Murphy, Machine Learning A Probabilistic Perspective, The MIT Press, 2017.
  - Darren Cook, Practical Machine Learning with H2O, O'Reilly Media, Inc., 2017.
  - Scott Burger, Introduction to Machine Learning with R: Rigorous Mathematical Analysis, O'Reilly Media, Inc., 2018.

Supervised Machine Learning

- Regression-based Methods
  - Generalized Linear Models
    - Linear Regression
    - Logistic Regression
  - Deep Learning (Neural Nets)
- Tree-based Ensemble Methods
  - Random Forest (Bagging: Bootstrap Aggregation)
    - Parallel ensemble to reduce variance
  - Gradient Boost Machine (Boosting)
    - Sequential ensemble to reduce bias

Tree-Based Models

- Random Forest (Bagging: Bootstrap Aggregation)
  - Parallel ensemble to reduce variance
- Gradient Boost Machine (Boosting)
  - Sequential ensemble to reduce bias

Trees

- Classification Tree

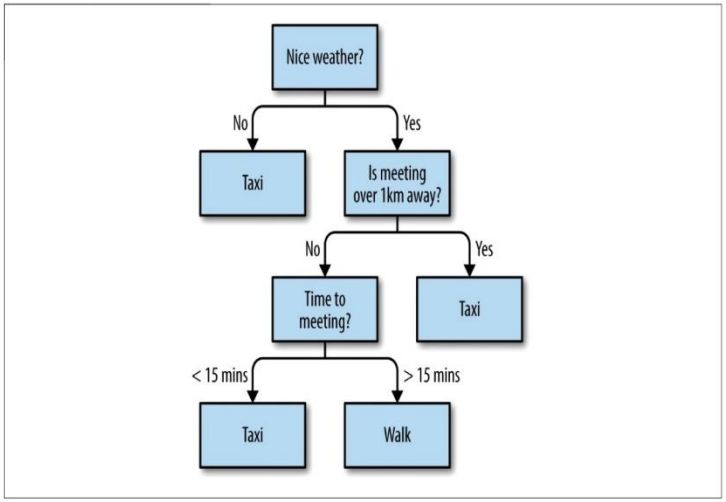


Figure 5-1. A classification tree: deciding whether to walk or catch a taxi

- Regression Tree

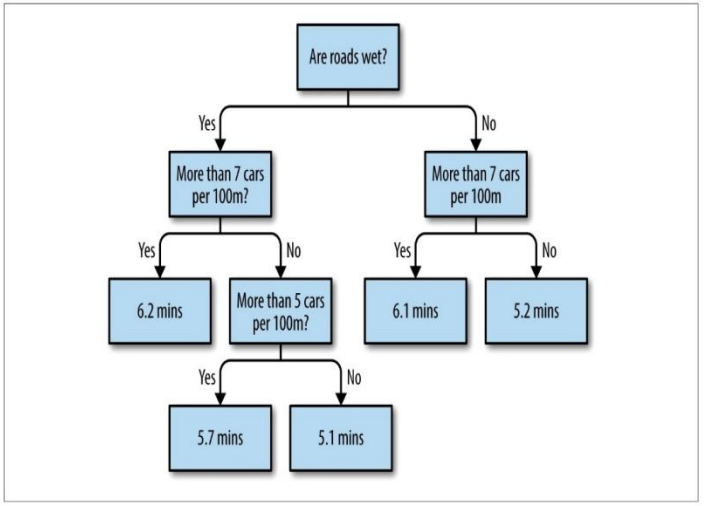
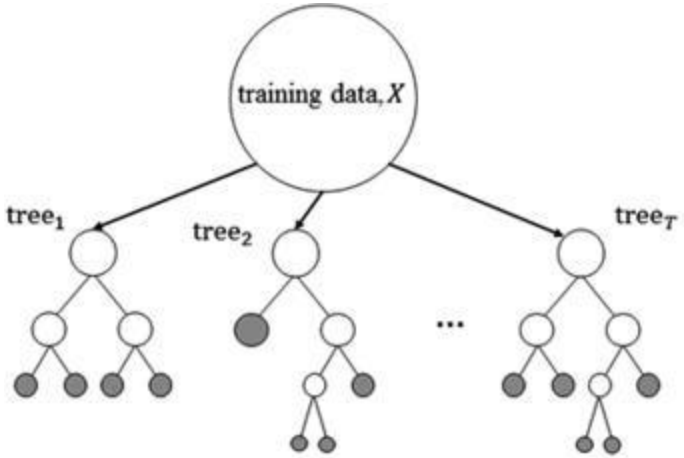


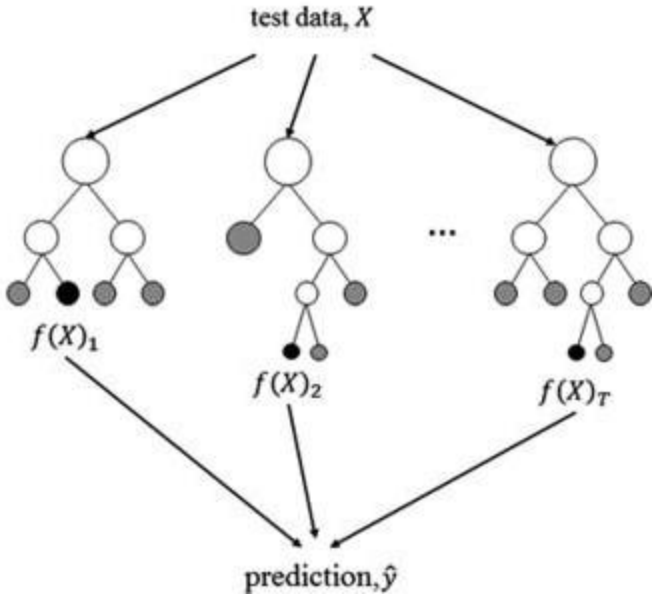
Figure 5-2. A regression tree: estimating how long a car journey will take

Random Forest

- Random Forest is a bagging (bootstrap aggregation) of trees.
- Given a set of data, each of these trees in the forest is a weak learner built on a subset of rows (data observations) and columns (features or variables).
- More trees will reduce the variance, which may be processed in parallel.

Random Forest





Random Forest Modeling with H2O

- Basic Model
  - `h2o.randomForest (x, y, training_frame, model_id = NULL, seed = -1, …)`
- Model Specification Options
  - `ntrees = 50, max_depth = 20, mtries = -1,`
  - `sample_rate = 0.632,`
  - `sample_rate_per_class = NULL, col_sample_rate_change_per_level = 1, col_sample_rate_per_tree = 1,`
  - `min_rows = 1, nbins = 20,`
  - `nbins_top_level = 1024, nbins_cats = 1024,`

Random Forest Modeling with H2O

- Model Specification Options (Continued)
  - `distribution = c("AUTO", "bernoulli", "multinomial", "gaussian", "poisson", "gamma", "tweedie", "laplace", "quantile", "huber"),`
  - `histogram_type = c("AUTO", "UniformAdaptive", "Random", "QuantilesGlobal", "RoundRobin"),`
  - `checkpoint = NULL,`

Random Forest Modling with H2O

- Cross-Validation Parameters
  - `validation_frame = NULL,`
  - `nfolds = 0, seed = -1,`
  - `keep_cross_validation_models = TRUE,`
  - `keep_cross_validation_predictions = FALSE,`
  - `keep_cross_validation_fold_assignment = FALSE,`
  - `fold_assignment = c("AUTO", "Random", "Modulo", "Stratified"),`
  - `fold_column = NULL,`

Random Forest Modeling with H2O

- Early Stopping
  - `stopping_rounds = 0,`
  - `stopping_metric = c("AUTO", "deviance", "logloss", "MSE", "RMSE", "MAE", "RMSLE", "AUC", "lift_top_group", "misclassification", "mean_per_class_error", "custom", "custom_increasing"),`
  - `stopping_tolerance = 0.001,`
  - `max_runtime_secs = 0,`

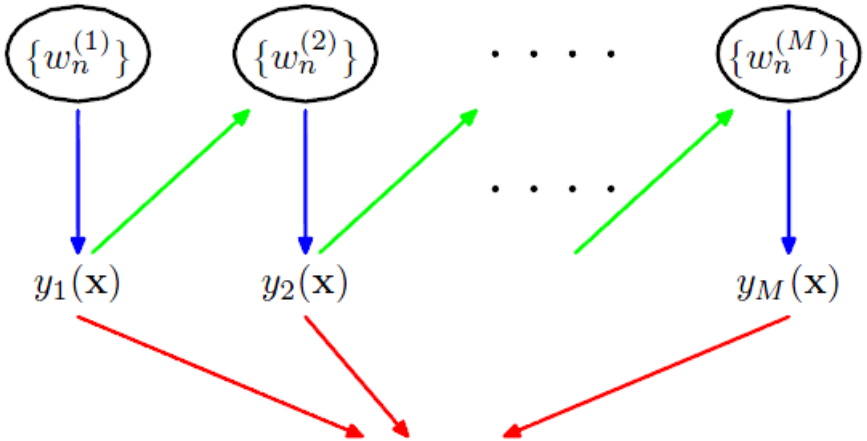
Random Forest Modeling with H2O

- Other Important Control Parameters
  - `balance_classes = FALSE,`
  - `class_sampling_factors = NULL,`
  - `max_after_balance_size = 5,`
  - `max_hit_ratio_k = 0,`
  - `min_split_improvement = 1e-05`
  - `binomial_double_trees = FALSE,`
  - `col_sample_rate_change_per_level = 1,`
  - `col_sample_rate_per_tree = 1,`

Gradient Boosting Machine

- Gradient Boosting Machine (GBM) is a forward learning ensemble method. It combines gradient-based optimization and boosting.
  - Gradient-based optimization uses gradient computations to minimize a model's loss function in terms of the training data.
  - Boosting additively collects an ensemble of weak models to create a robust learning system for predictive tasks.

Boosting



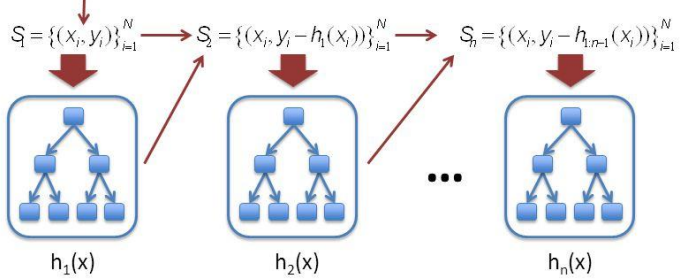
$$Y_M(\mathbf{x}) = \text{sign}\left(\sum_m^M \alpha_m y_m(\mathbf{x})\right)$$

Gradient Boosting Machine

Gradient Boosting (Simple Version)

(Why is it called "gradient"?) (For Regression Only)
(Answer next slides.)

$S = \{(x_i, y_i)\}_{i=1}^{N}$ $\quad h(x) = h_1(x) + h_2(x) + \ldots + h_n(x)$



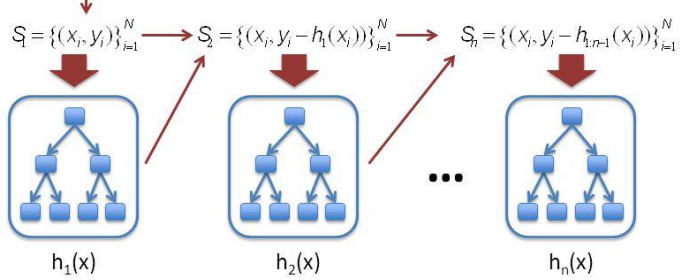
http://statweb.stanford.edu/~jhf/ftp/trebst.pdf

24

Gradient Boosting (Full Version)

(Instance of Functional Gradient Descent) (For Regression Only)

$S = \{(x_i, y_i)\}_{i=1}^{N}$ $\quad h_{1:n}(x) = h_1(x) + \eta_2 h_2(x) + \ldots + \eta_n h_n(x)$



http://statweb.stanford.edu/~jhf/ftp/trebst.pdf ← See reference for how to set η

33

Gradient Boosting with H2O

- Basic Model
  - `h2o.gbm (x, y, training_frame, model_id = NULL, seed = -1, …)`
- Model Specification Options
  - `ntrees = 50, max_depth = 5, min_rows = 10,`
  - nbins = 20, nbins_top_level = 1024, nbins_cats = 1024,
  - `learn_rate = 0.1, learn_rate_annealing = 1,`
  - `sample_rate = 1, sample_rate_per_class = NULL, col_sample_rate = 1, col_sample_rate_change_per_level = 1, col_sample_rate_per_tree = 1, max_abs_leaf, node_pred = Inf, …)`

Gradient Boosting with H2O

- Model Specification Options (Continued)
  - `distribution = c("AUTO", "bernoulli", "quasibinomial", "multinomial", "gaussian", "poisson", "gamma", "tweedie", "laplace", "quantile", "huber"),`
  - `quantile_alpha = 0.5,`
  - `tweedie_power = 1.5,`
  - `huber_alpha = 0.9,`
  - `checkpoint = NULL`

Gradient Boosting with H2O

- Cross-Validation Parameters
  - `validation_frame = NULL,`
  - `nfolds = 0, seed = -1,`
  - keep_cross_validation_models = TRUE,
  - `keep_cross_validation_predictions = FALSE,`
  - `keep_cross_validation_fold_assignment = FALSE,`
  - `fold_assignment = c("AUTO", "Random", "Modulo", "Stratified"),`
  - `fold_column = NULL,`

Gradient Boosting with H2O

- Early Stopping
  - `stopping_rounds = 0,`
  - `stopping_metric = c("AUTO", "deviance", "logloss", "MSE", "RMSE", "MAE", "RMSLE", "AUC", "lift_top_group", "misclassification", "mean_per_class_error", "custom", "custom_increasing"),`
  - `stopping_tolerance = 0.001,`
  - `max_runtime_secs = 0,`

Gradient Boosting with H2O

- Other Important Control Parameters
  - `min_split_improvement = 1e-05`
  - `histogram_type = c("AUTO", "UniformAdaptive", "Random", "QuantilesGlobal", "RoundRobin")`