STATS 769

Ensemble Methods

Yong Wang
Department of Statistics
The University of Auckland

2021

Ensemble Methods

- An ensemble method is to build a large number of simple models in some way and combine them to obtain a single and potentially much more powerful model.
- Useful for many models, but particularly so for tree-based models.
- Main ensemble methods
    - Bagging
    - Random Forests
    - Boosting

Bagging

- Bagging stands for "Bootstrap aggregation".
- A bootstrap sample is a random sample $(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$ drawn independently with replacement from observations $(x_1, y_1), \ldots, (x_n, y_n)$.
  - It has the same size n, but with duplicated observations.
- Consider some estimator $\widehat{f}$ (which produces a model given a data set).
- Generate B bootstrap samples, and build a model $\widehat{f}_b^*$ for the bth bootstrap sample.
- To predict the value of y given x, use

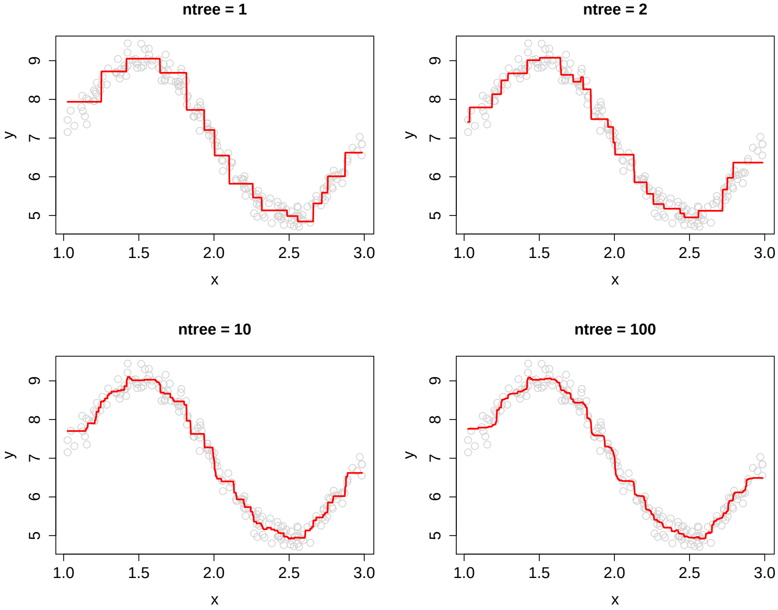
$$\widehat{f}^{\mathrm{bag}}(x) = \frac{1}{B} \sum_{b=1}^{B} \widehat{f}_b^*(x)$$

for regression, and the majority rule for classification.

Out-of-Bag Errors

- To estimate prediction errors, no need to resort to cross-validation.
- Each bootstrap sample uses about $\frac{2}{3}$ of the original observations (with duplications).
- The remaining, out-of-bag observations can thus be used to provide test errors, which are known as out-of-bag (OOB) estimates of errors.
- For B sufficiently large, every original obvervation will be out-of-bag at least once, and the averaged out-of-bag test error is the estimate of its individual PE.
- The overall PE is the mean of n individual PEs.

Bagging (`nodesize = 20`): Number of Trees



Random Forests

- It is an improvement upon Bagging with trees, with a small tweak.
- While searching for an optimal split, the standard tree-building approach is to consider all p predictors as candidates.
- However, a Random Forest only considers m randomly chosen predictors as candidates for each search for an optimal split.
- Typically, $m \approx \sqrt{p}$.
- It is Bagging, if $m = p$.
- This gives weaker predictors a chance to be used in tree building.
- As a result, the effects of these predictors are also well taken into account.

Boosting

- The fundamental idea of Boosting is repeatedly fit a primary model to residuals.
- Where the current model does not fit well to the residuals will get “boosted” next time.
- Residuals are progressively be explained off with more fitted primary models included.
- The primary model used is often chosen to be simple.
- Unlike Bagging and Random Forests, Boosting can overfit the model.

Boosting for Regression

- Regression models can be some small regression trees, e.g., those with d splits (thus $d+1$ leaf nodes).
- Algorithm:
  1. Choose a small value for $\lambda > 0$, and set $r_i = y_i$ for all i.
  2. For $b = 1, \ldots, B$, repeat:
     - Fit a regression model $\widehat{f}_b(x)$ to data $\{(x_i, r_i)\}_{i=1}^n$.
     - Update residuals: $r_i = r_i - \lambda \widehat{f}_b(x_i)$ for all i.
  3. Output

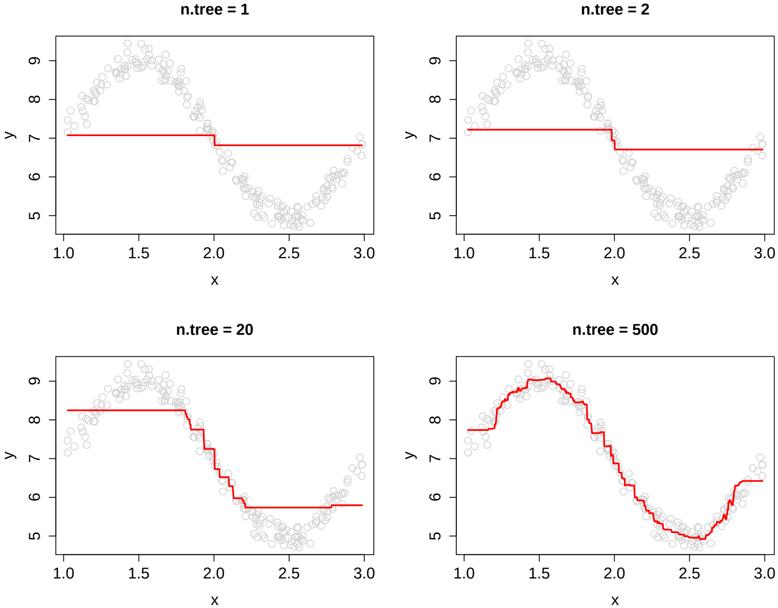
$$\widehat{f}^{\text{boost}}(x) = \lambda \sum_{b=1}^{B} \widehat{f}_b(x).$$

Boosting for Classification

- Suppose $Y \in \{-1, 1\}$, i.e., a two-class problem.
- A classifier $\widehat{f}(x)$ returns either -1 or 1 here.
- Algorithm AdaBoost.M1:
  1. Initialise observation weights $w_i = 1/n$ for all i.
  2. For $b = 1, \ldots, B$, repeat:
     - Fit a classifier $\widehat{f}_b(x)$ to data $\{(x_i, y_i)\}_{i=1}^n$ with weights w_i.
     - Compute $e_b = \sum_{i=1}^n w_i I[y_i \neq \widehat{f}_b(x_i)]$.
     - Compute $\lambda_b = \log[(1 - e_b)/e_b]$.
     - Set $w_i = w_i \exp\{\lambda_b I[y_i \neq \widehat{f}_b(x_i)]\}$ for all i.
     - Set $w_i = w_i/(\sum_l w_l)$ for all i.
  3. Output

$$\widehat{f}^{\text{boost}}(x) = \operatorname{sign}\left[\sum_{b=1}^{B} \lambda_b \widehat{f}_b(x)\right].$$

Boosting ($d = 1$): Number of Trees



Pros and Cons of Ensemble Methods

Pros

- Higher prediction accuracy

Cons

- Lower interpretability

Parallell Computing with Ensemble Methods

- Bagging and Random Forests can be easily run in parallel.
- But not for Boosting. There is no random sample drawn.

Recommended Readings

ISLv2 (basics):

- Section 8.2
- Labs: Sections 8.3.3, 8.3.4

ESL (advanced):

- Sections 8.7, 10.1, 10.9, 15.1–15.3