Lecture 8

~~Start with a review: Quiz 1~~

Review of bagging and random forests

Implementing a random forest for linear outcome

If time, we will start conditional logistic regression

Lecture 7 Review:

- CART: Classification And Regression Trees
  - Made the link between the CART and a regression model
  - Examples with both linear and binary outcome
  - Compared CART to parametric model using AUC ROC

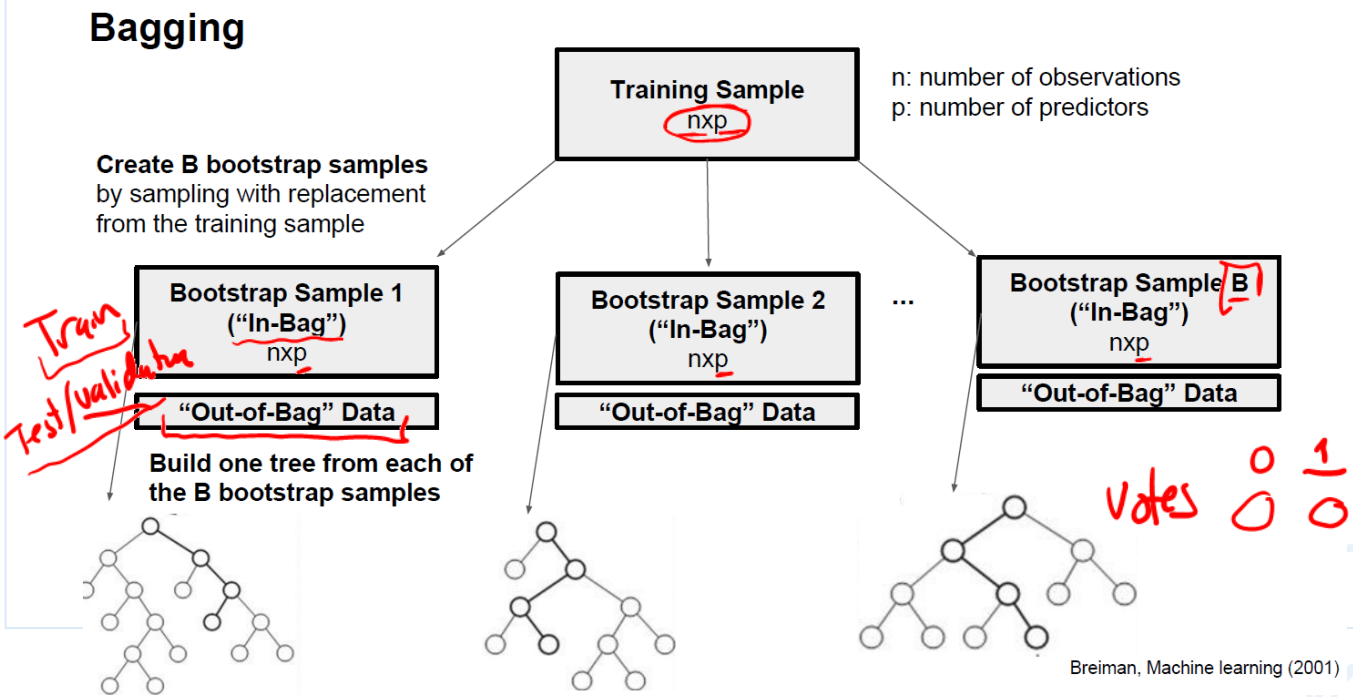
Binary outcomes

linear outcomes → MSE of prediction

- Random forests
  - One issue with CART is sensitivity of predictions to small perturbations in X
  - CART results are highly variable
  - Ensemble learners: predictions based on averaging over many realizations of CART
  - Ensemble learners: improve accuracy and precision

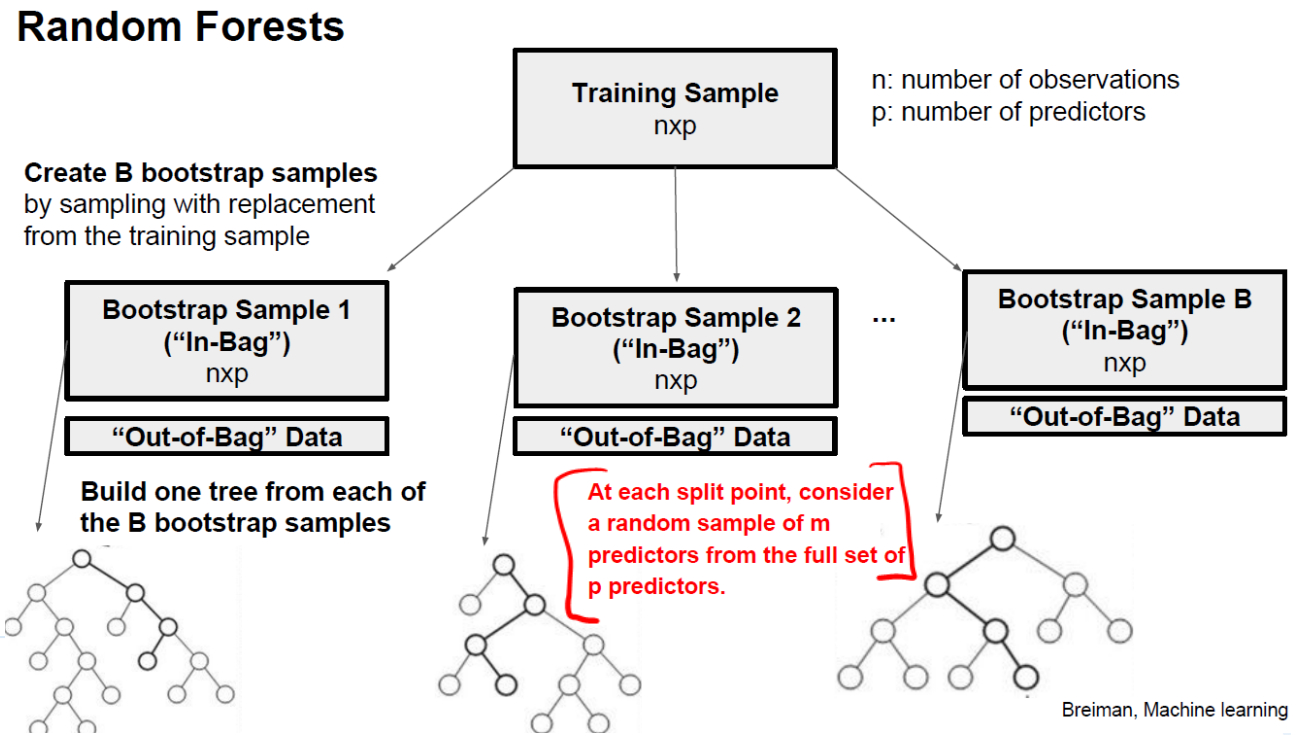
Ensemble Methods

- Bagging (**Bootstrap** **Aggregating**)
- Random Forests



Ensemble Methods

- Bagging (**Bootstrap** **Aggregating**)
- Random Forests



Breiman, Machine learning (2001)

Out-of-bag samples

- Allows for internal cross-validation / evaluation of the quality of the ensemble method
- For a random forest constructed with B trees,
  - Each observation in the training data appears in
    - Roughly 2/3 of the trees / bootstrap samples
    - Roughly 1/3 of out-of-bag samples
  - Predicted values for observations in the training data are based on predictions made from the trees where the observation was an out-of-bag observation
    - Linear/continuous outcome: average prediction across all out-of-bag trees
    - Binary/categorical outcome: proportion of votes received for each level of the outcome across all out-of-bag trees, i.e. Pr(Y=y) for all y = 0, 1, …, and classification assigned as the y that yields the largest Pr(Y=y)
- Out-of-bag error is used to evaluate the tuning parameters for the ensemble learner
  - Linear/continuous outcome: mean squared error computed with out-of-bag prediction
  - Binary/categorical outcome: misclassification error computed with out-of-bag classification

Construction of random forest

- Parameters that we control
  - Number of variables considered at each split, m
    - Classification tree: floor square-root p
    - Regression tree: floor p/3
  - Number of trees

- Implementation:
  - To find m: set number of trees large (e.g. 500), identify minimum out-of-bag error for m = 1, 2, …, beyond default.
  - After finding m: check to see if your forest is sensitive to number of trees by plotting MSE or out-of-bag misclassification error as a function of number of trees.

Example: Predict log(expenditures + 1) in NMES

- See handout

Review of logistic regression assumptions

- And solutions to violations

- Mean model is correctly specified
  - Plot average predicted vs. observed proportions within quintiles or deciles of predicted values
  - Plot average predicted vs. observed proportions as a function continuous exposure
  - Summary tables of average predicted vs. observed proportions by level of categorical exposure
  - SOLUTION: change your mean model

- Observations are independent
  - More on this next Tuesday

Review of logistic regression assumptions

- Variance is correctly specified
  - Logistic model assumes: Var(Y) = p(1-p)
  - Under or over-dispersion
  - Compute Var(Y) and compare with predicted variance, overall or by select variables
  - SOLUTION:
    - Bootstrap
    - GLM: family = “quasibinomial” assumes Var(Y) = ϕ x p x (1-p) where ϕ = 1/(n-k) sum of squared Pearson residuals

- There are no “influencial” observations
  - DFFITS or DFBETAS