QTM 347 Machine Learning

Lecture 13: Decision tree and bagging

Ruoxuan Xiong

Suggested reading: ISL Chapter 8

Regression tree

- **Two main steps** in constructing regression trees
  1. Partition the feature space into J **distinct and non-overlapping** regions, $R_1, R_2, \cdots, R_J$
  2. Make the **same** prediction for every observation in region R_j: Mean of the training observations in R_j

- Tree pruning to avoid overfitting, e.g., use cost complexity pruning
  - Solve the problem:

$$\min \sum_{j=1}^{|T|} \sum_{i \in R_j} \left(Y_i - \hat{Y}_{R_j}\right)^2 + \alpha |T|$$

Incorrect variation of cross-validation to select α

- **Cross-validation**: Split the training observations into 10 folds
  - For a range of values $\alpha_1, \alpha_2, \cdots, \alpha_m$, construct the corresponding sequence of trees are $T_1, T_2, \cdots, T_m$
    - **Tree structures are fixed** in the cross validation
  - For $k = 1, \cdots, 10$, using every fold except the kth
    - Make prediction for each region in each tree T_i
      - Prediction for each region vary with the hold-out fold k
    - For each tree T_i, calculate the RSS on the hold-out fold k
  - Select the optimal tree T_i that minimizes the average error across 10 folds

Correct variation of cross-validation to select α

- **Cross-validation**: Split the training observations into K folds
  - Hold out the kth fold, for a range of values $\alpha_1, \alpha_2, \cdots, \alpha_m$, construct the corresponding sequence of trees are $T_1^{(k)}, T_2^{(k)}, \cdots, T_m^{(k)}$
  - The **sequence of trees vary** with which fold is held out
  - Use tree $T_i^{(k)}$ to make prediction and calculate RSS on the hold-out fold k
  - Select the optimal parameter α that minimizes the average error across ten folds

Classification tree

- Classification trees work much like regression trees. Instead,

- In step 1, minimize the classification error rate (rather than RSS)

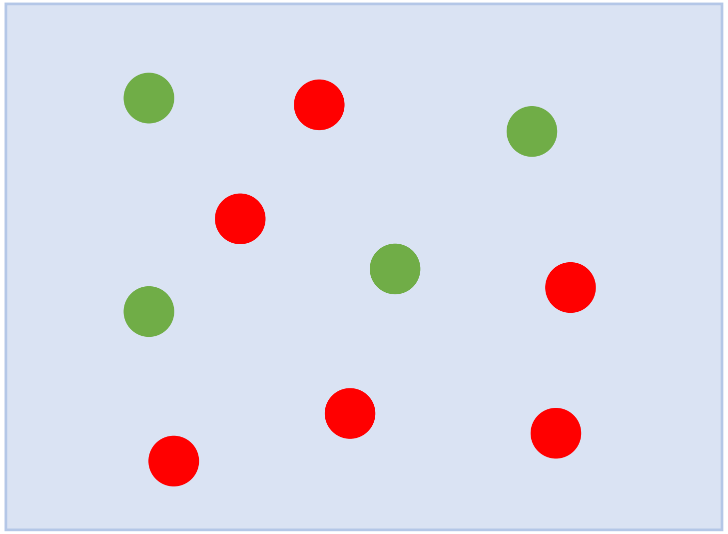
- In step 2, predict the response by **majority vote**, i.e. pick *the most common class* in every region

Classification losses: The $0-1$ loss

- The $0-1$ loss or misclassification rate in region m:

$$\sum_{i \in R_m} 1(Y_i \neq \hat{Y}_{R_m})$$

- Example:
  - $\hat{Y}_{R_m} = \text{red}$
  - $\sum_{i \in R_m} 1(Y_i \neq \hat{Y}_{R_m}) = 4$



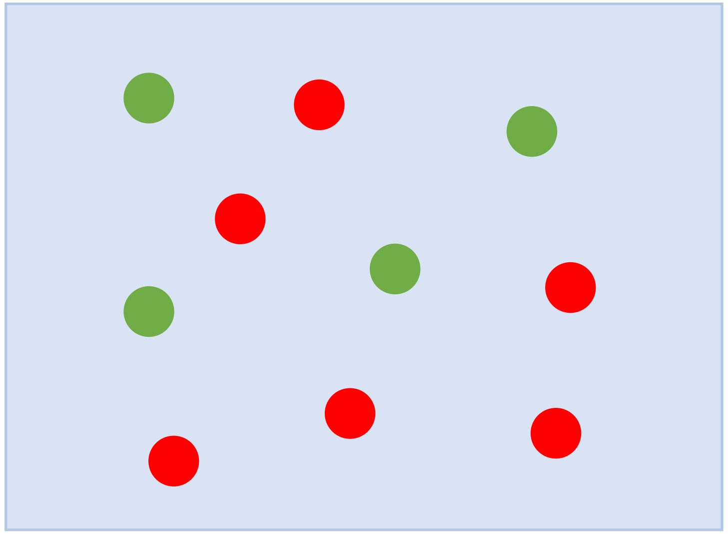
Region m

Classification losses: Gini index

- The Gini index in region m

$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

- p_{mk}: proportion of training observations in the mth region that are from kth class
- Example:
  - $\hat{p}_{m,\text{red}} = \frac{6}{10} = 0.6$
  - $\hat{p}_{m,\text{green}} = \frac{4}{10} = 0.4$
  - $G_m = 0.6(1 - 0.6) + 0.4(1 - 0.4) = 0.48$



Region m

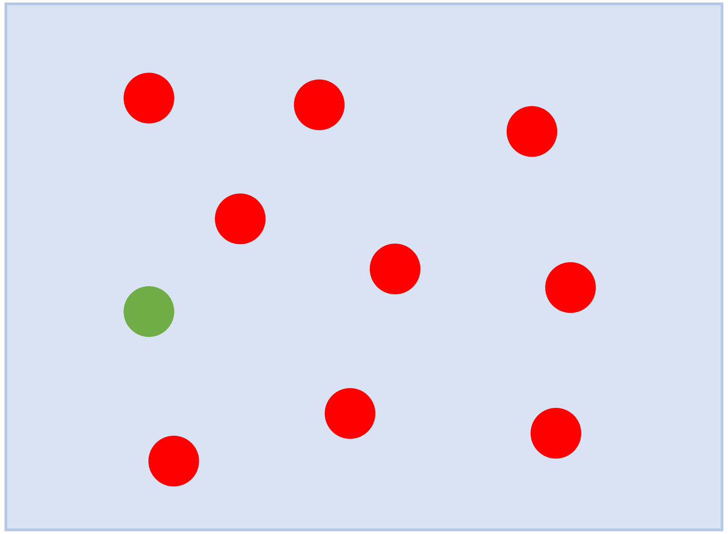
Classification losses: Gini index

- The Gini index in region m

$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

- p_{mk}: proportion of training observations in the mth region that are from kth class

- Example:
  - $\hat{p}_{m,\text{red}} = \frac{9}{10} = 0.9$
  - $\hat{p}_{m,\text{green}} = \frac{1}{10} = 0.1$
  - $G_m = 0.9(1 - 0.9) + 0.1(1 - 0.1) = 0.18$



Region m

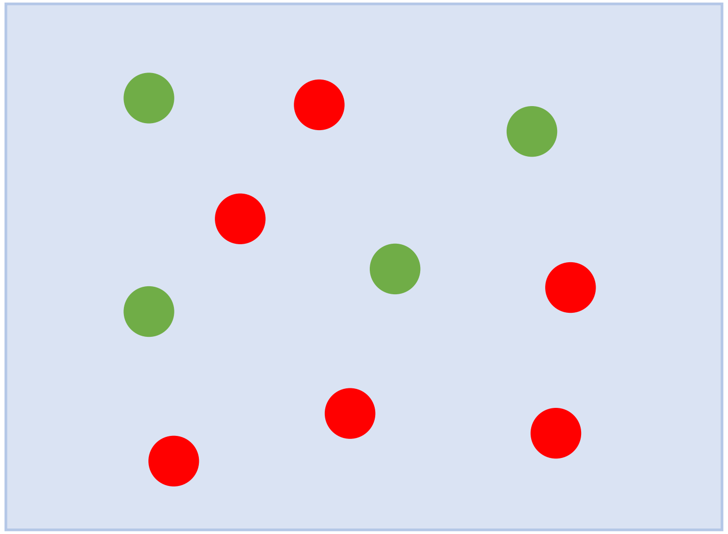
- G_m is a measure of node purity
  - G_m is small if all $\hat{p}_{mk}$'s are close to zero or one

Classification losses: Entropy

- The entropy in region m

$$D_m = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

- p_{mk}: proportion of training observations in the mth region that are from kth class
- Example:
  - $\hat{p}_{m,\text{red}} = \frac{6}{10} = 0.6$
  - $\hat{p}_{m,\text{green}} = \frac{4}{10} = 0.4$
  - $D_m = -0.6 \log 0.6 - 0.4 \log 0.4 = 0.673$



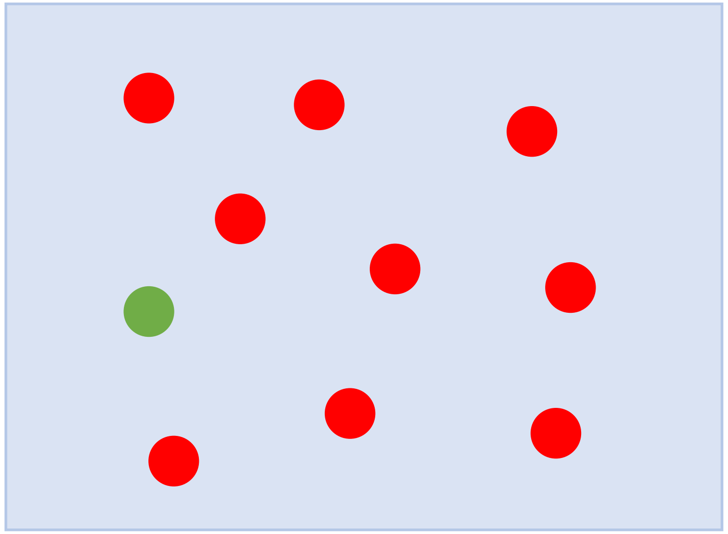
Region m

Classification losses: Entropy

- The entropy in region m

$$D_m = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

- p_{mk}: proportion of training observations in the mth region that are from kth class
- Example:
  - $\hat{p}_{m,\text{red}} = \frac{9}{10} = 0.9$
  - $\hat{p}_{m,\text{green}} = \frac{1}{10} = 0.1$
  - $D_m = -0.9 \log 0.9 - 0.1 \log 0.1 = 0.461$
- D_m is another measure of purity
  - D_m is small if all $\hat{p}_{mk}$'s are close to zero or one



Region m

Classification losses

- The 0–1 loss or misclassification rate in region m (prune tree)

$$\sum_{i \in R_m} 1(Y_i \neq \hat{Y}_{R_m})$$

- The Gini index in region m (evaluate the quality of a split)

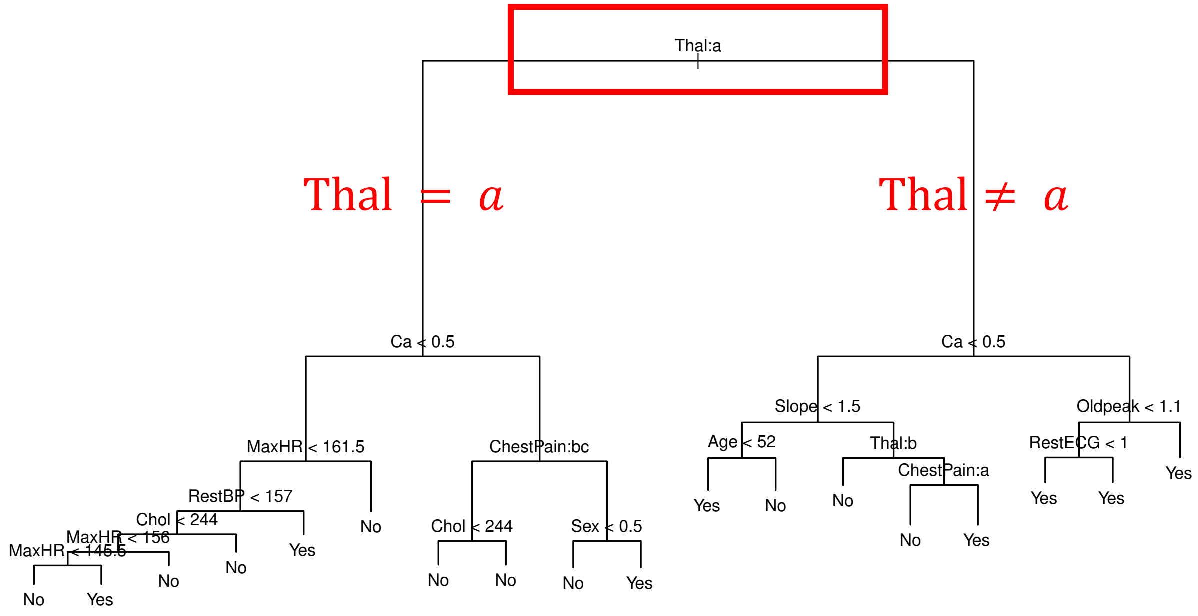
$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

- The entropy in region m (evaluate the quality of a split)

$$D_m = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

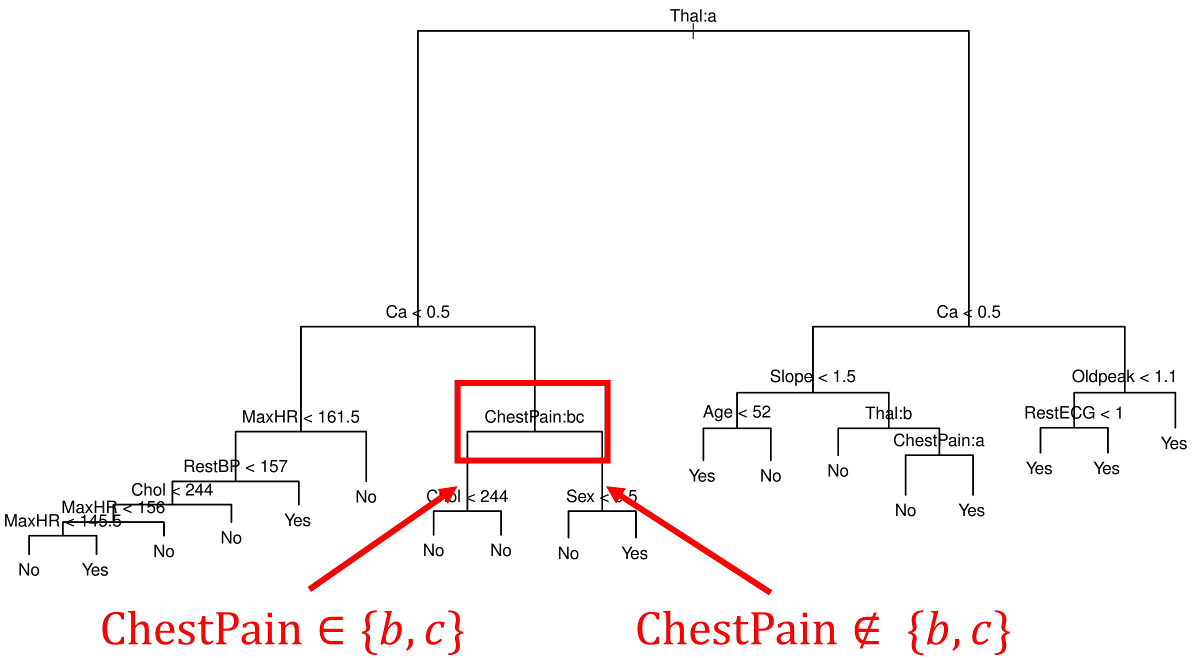
Example: Predicting heart disease

- Predict whether a patient with chest pain has heart disease based on Age, Sex, Chol (a cholesterol measure), and other heart and lung function measures
- Some predictors are qualitative
  - Thal (Thallium stress test)
  - ChestPain
  - Sex



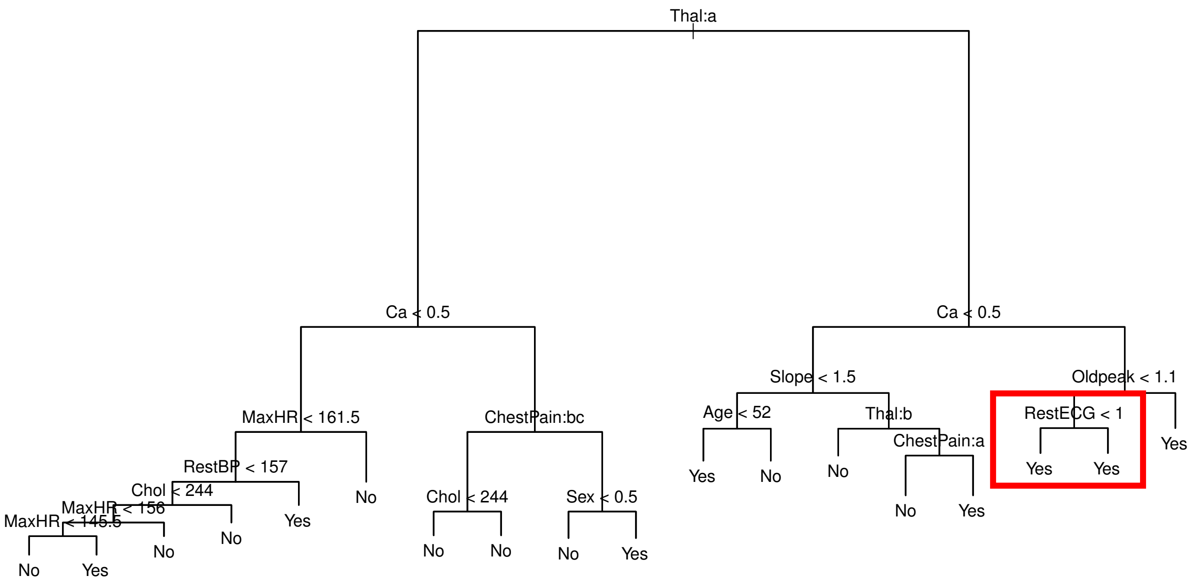
Example: Predicting heart disease

- Predict whether a patient with chest pain has heart disease based on Age, Sex, Chol (a cholesterol measure), and other heart and lung function measures
- Some predictors are qualitative
  - Thal (Thallium stress test)
  - ChestPain
  - Sex



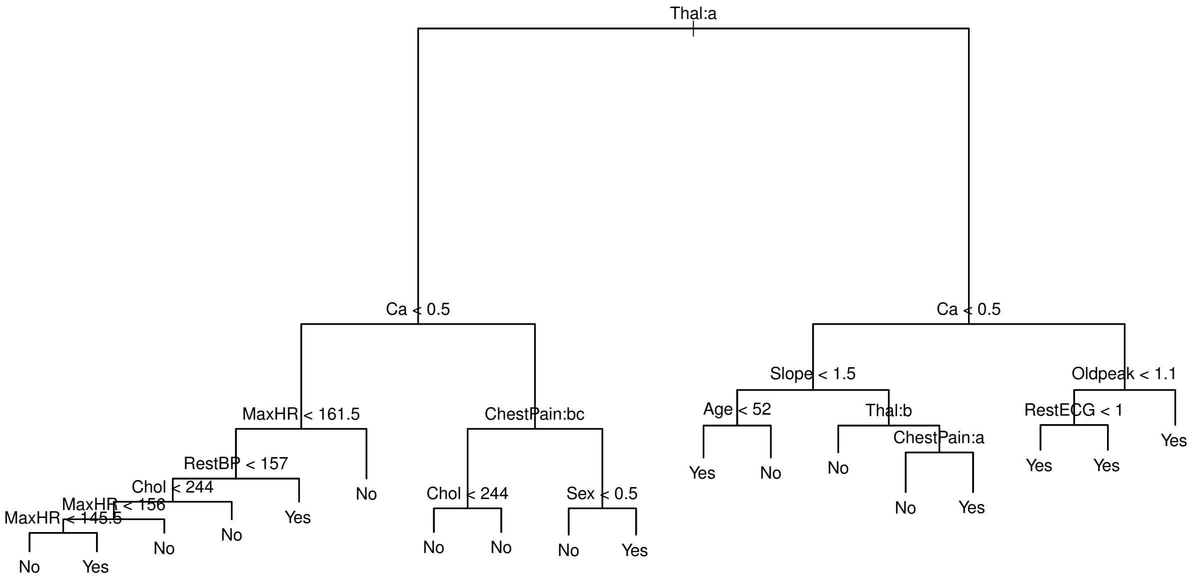
Example: Predicting heart disease

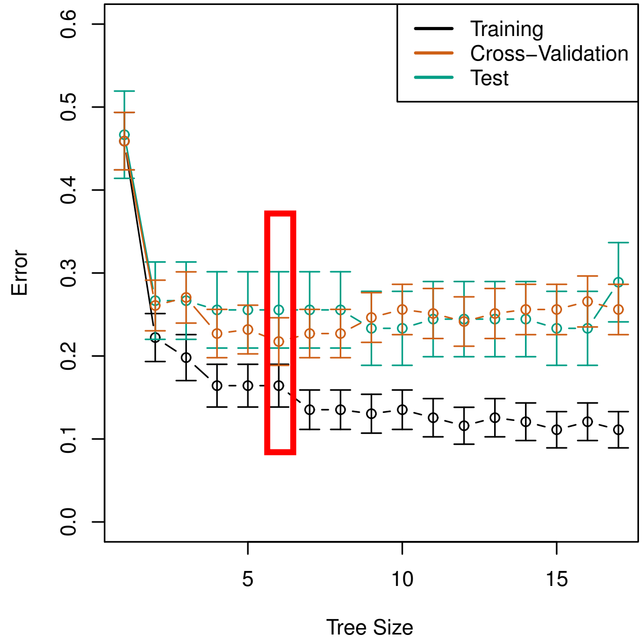
- Predict whether a patient with chest pain has heart disease based on Age, Sex, Chol (a cholesterol measure), and other heart and lung function measures
- Some terminal nodes have the same predicted value
  - Reason: Increased node purity
- Example
  - RestECG ≥ 1: $\frac{9}{9}$ with Yes
  - RestECG < 1: $\frac{7}{11}$ with Yes



Example: Predicting heart disease

- Predict whether a patient with chest pain has heart disease based on Age, Sex, Chol (a cholesterol measure), and other heart and lung function measures
- Cross validation to prune tree





Example: Predicting heart disease

- Predict whether a patient with chest pain has heart disease based on Age, Sex, Chol (a cholesterol measure), and other heart and lung function measures
- Pruned tree after cross-validation: