Classification Tree, Random Forest, and Bagging

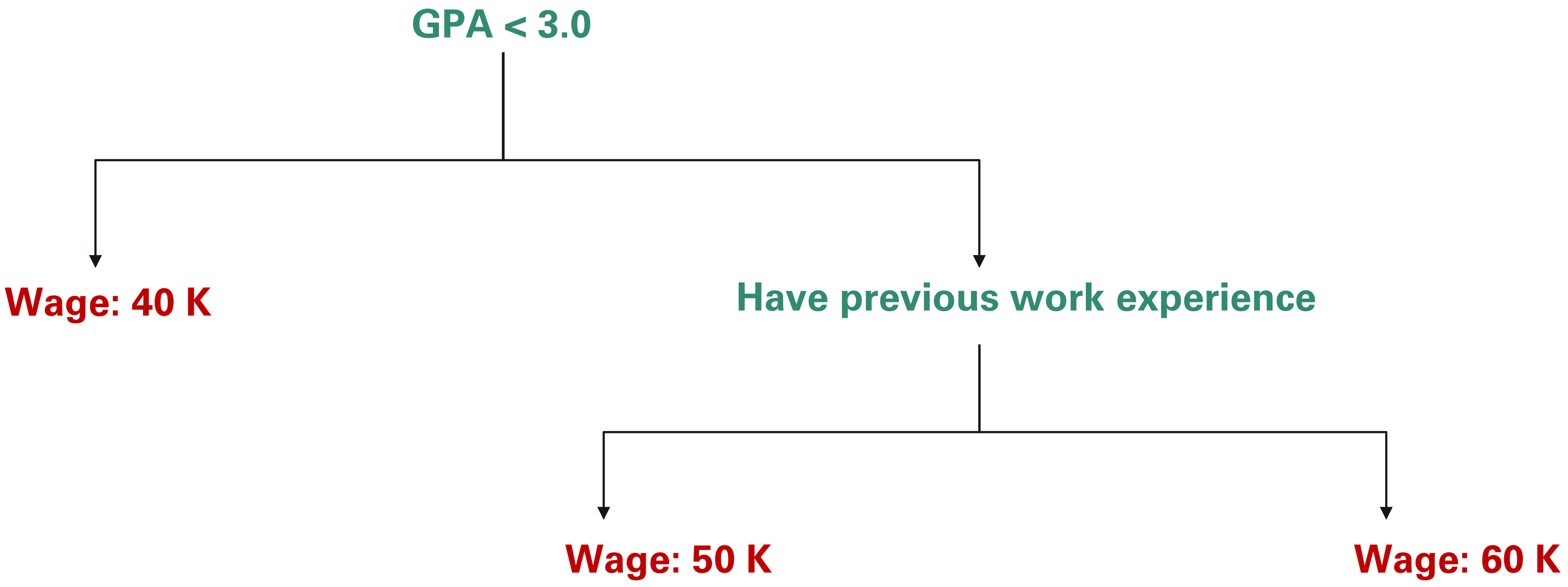
ECON 258

PREPARED BY ARDINA HASANBASRI

Recap from basic tree model!

1) Give the model variables available: GPA, work experience, school ranking, area, etc.
2) Model finds variable cutoff such that it minimizes residual sum of squares

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$



1.1) Classification Tree

I need to put these observations into categories!

Recall classification trees predict categorical responses.

- **We care about what is the most occurring class, but also the proportion of a class that falls into a specific region.**
- **Determining tree splits cannot be done using RSS anymore.**
- **Instead, we will use the Gini Index.**

1.2) Gini Index

The Gini Index is defined by:

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

$\hat{p}_{mk}$ represents proportion of training observation that fall in the mth region and is from a k class.

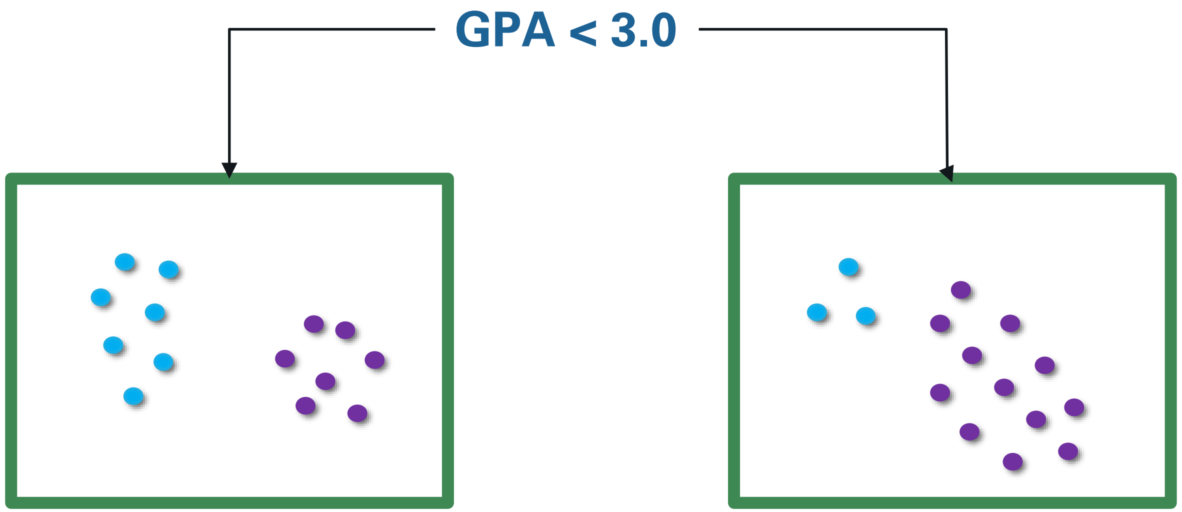
The Gini index measures node purity. When $\hat{p}_{mk}$ is closer to zero or one, gini index is small.

The smaller, the better, meaning our prediction is more precise.

Suppose my tree predicts whether a student in this class becomes an economist (blue) or a data scientist (purple).

The tree checks if cutting region by GPA is good.

Suppose GPA is cut at 3.0.



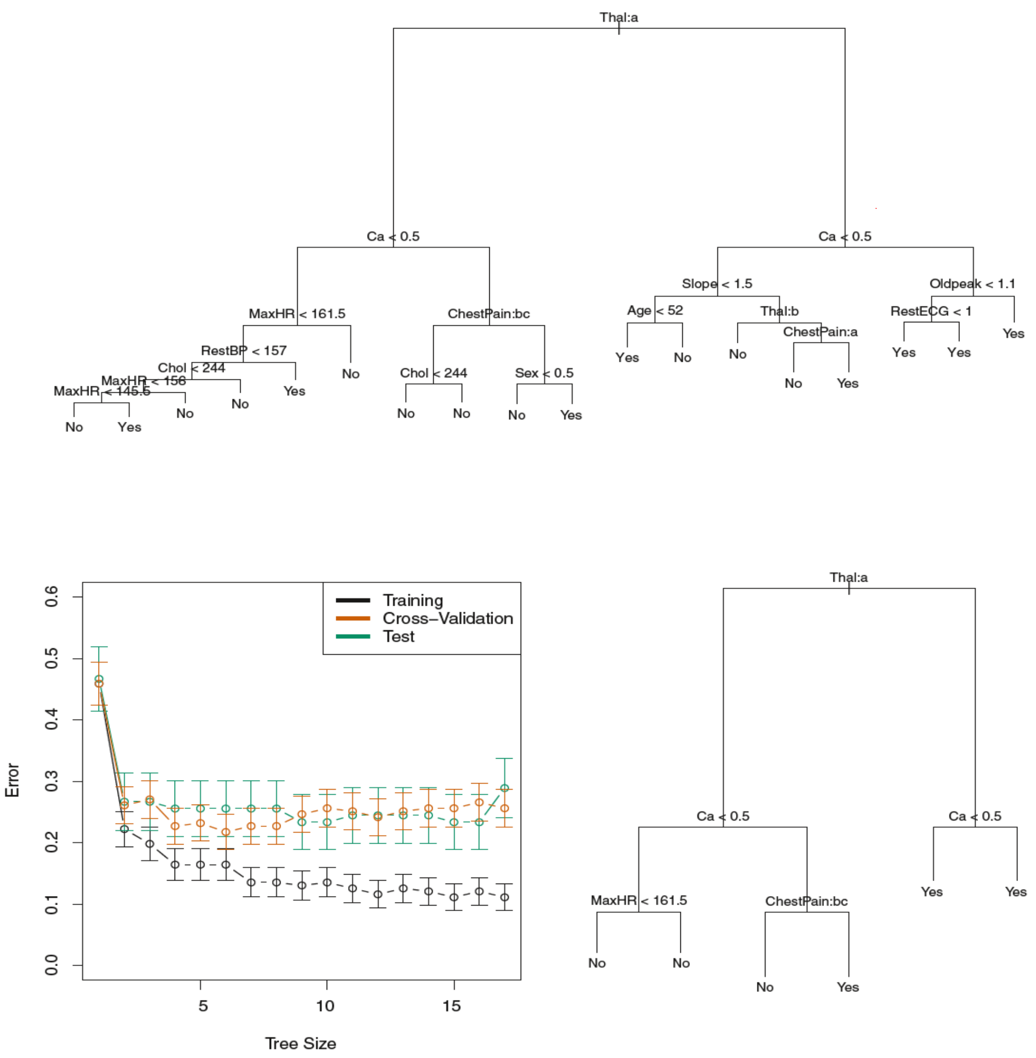


FIGURE 8.6. Heart *data.* Top: *The unpruned tree.* Bottom Left: *Cross -validation error, training, and test error, for different sizes of the pruned tree.* Bottom Right: *The pruned tree corresponding to the minimal cross-validation error.*

1.3) Example of Classification Tree

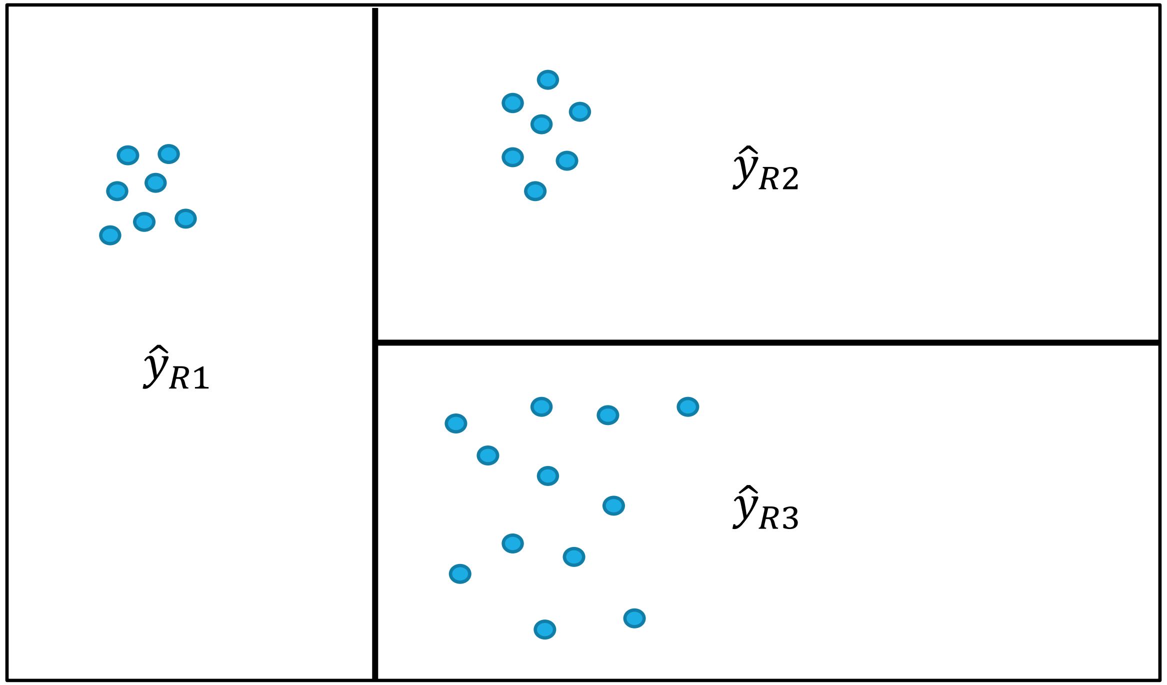
- 303 patients who have chest pain
- Outcome variable: Yes, No for presence of heart disease

Exercise: “Lesson 16a Classification Tree Exercise”

2.1) Problem of High Variance in Regression Trees

The y prediction for an observation depends on the average y in region i.

Regression trees can suffer from high variance.



2.2) Method 1: Bagging to Reduce Variance

One way to reduce variance is using a bagging method.

Intuition for procedure:

1) Suppose you have a set of n independent observations $Z_1, Z_2, Z_3, \ldots Z_n$. Each has a variance of σ^2.
2) If we take the average of all the Z, then the variance of the average is $\frac{\sigma^2}{n}$.
3) Therefore, for a number of training sets, build a decision tree for each, and average the resulting prediction for an observation.

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x).$$

2.3) Bootstrapping

1. **Since we typically do not have multiple training sets, we can bootstrap samples from the training set.**

2. **Bootstrap is a resampling method (cross validation is also a resampling method).**

3. **Bootstrapping means you are repeatedly obtaining distinct data sets from the original sample.**

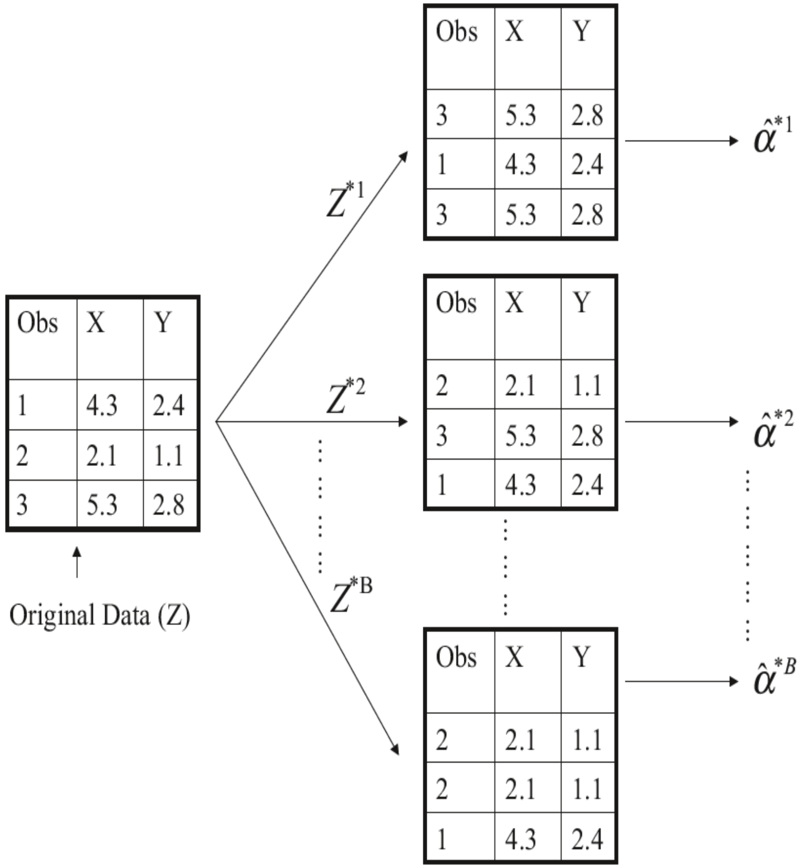


FIGURE 5.11. *A graphical illustration of the bootstrap approach on a small sample containing $n = 3$ observations. Each bootstrap data set contains n observations, sampled with replacement from the original data set. Each bootstrap data set is used to obtain an estimate of α.*

2.4) Possible Issue with Bagging

Bagging can average trees that are highly correlated with each other.

If there is a variable that is a high predictor, then every tree will be similar.

- **This high predictor variables will be chosen at the top each time.**

- **Since you are averaging similar trees, the variance do not decrease as much from a regular regression tree.**

2.5) Method 2: Random Forest

To reduce the issue of too much correlation across trees, we can do random forest.

Procedure for random forest:

1) When building a tree and considering a split, only consider a random sample of variables (m out of p).
2) This creates trees that are not necessarily correlated since algorithm is forced to choose different variables.

Problem with random forest and bagging, variable interpretation is harder than a normal decision tree.

3.1) Measuring Variable Importance

It is easy to interpret a single decision tree.

- The variable at the root is the most important.
- Based on the algorithm, each split minimizes RSS. Thus, top variables are more important based on when they are split.

Can we still measure variable important in bagging or random forest?

- YES!

1) Can calculate how much a variable matters by taking it out of the model and see how much accuracy fell (mean square error for example).
2) Can calculate how much RSS or node impurity changes when variable is taken out.

3.2) Partial Dependence Plots

We now know the most important variables for prediction, how do we measure how output changes as a variable changes?

Drawing partial dependence plots.

-> Draw the marginal effect of variable x1 conditional on the observed values of all other variables.

$$\bar{f}_s(z_s) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(z_s, z_{i,c})$$

3.3) Partial Dependence Algorithm

Let's say we care about the variable x_1.

Copy the training data and replace original value of x_1 with a constant x_{1i} from observation i.

Now predict outcome based on the model.

Calculate the average prediction.

Repeat for different x_1 for a different observation.

Plot x_{1i} and the average prediction.

Coding Exercise: “Lesson 16b Bagging and Random Forest”