Predictive Accuracy of Regression / Regression Trees

Data Mining
Prof. Dawn Woodard
School of ORIE
Cornell University

Outline

1. **Measuring Predictive Accuracy**
   - Example: Model Selection in Linear Regression
2. **Regression Trees**
   - Definition
   - Training the Tree
   - Examples

Predictive Accuracy of Regression

So far we have discussed two different types of supervised learning:

1. Classification (categorical outcome)
2. Regression (continuous outcome)

Evaluating Accuracy:

- For classification we evaluated the accuracy of our method by calculating % misclassified in the validation or test data (or looking at false pos. and false neg. rates separately)
- For regression how might we measure predictive accuracy?

Predictive Accuracy of Regression

- Often we measure predictive accuracy using the root mean square error on the validation/test data (smaller is better):

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(Y_i - \hat{Y}_i)^2}$$

  where M is the # of test observations.

- Alternative: Mean absolute error:

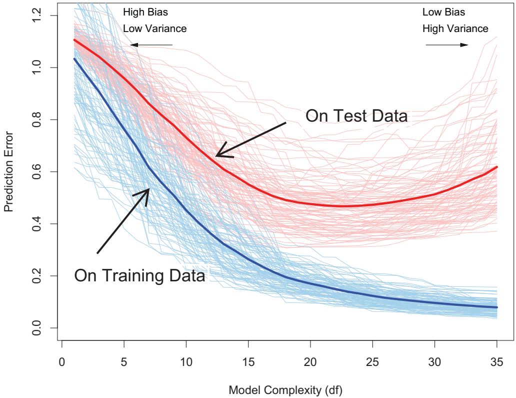
Model Selection for Linear Regression

Which predictor variables to include in the regression model?

- Why is it not necessarily best to include all the available predictors?

Model Selection for Linear Regression

Recall the complexity / accuracy curve. How might "model complexity" be measured for our linear regression model?



Model Selection for Linear Regression

What's a good way to choose a set of predictors that yields high predictive accuracy (accuracy on unseen data such as test data)?

All Subsets Regression

- All Subsets Regression means that we compare all possible subsets of the predictors, and choose a model based on some criterion, like the one on the previous slide.
- What happens as the number of available predictors gets large?

Forward Stepwise Selection

Forward stepwise selection is an alternative:

Backward Stepwise Selection

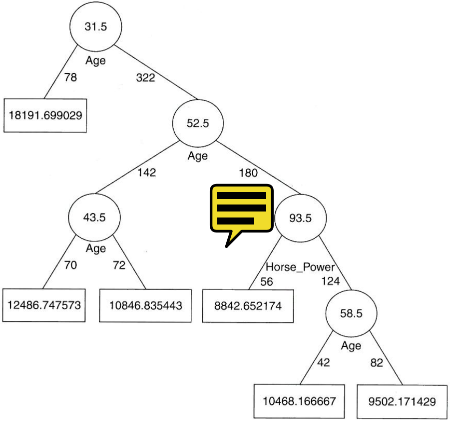
Backward stepwise selection is similar:

Definition of Regression Trees

- Regression trees are a method for prediction of a continuous outcome (as for linear regression)
- they partition the space $\mathbb{R}^p$ of the predictor vector $(X_1, \ldots, X_p)$ and predict a fixed value $\hat{Y}$ on each of the partition sets.
  - In particular:
- These splits correspond to simple logical rules for prediction.
  - Example:

Definition of Regression Trees

A tree for predicting the price of used Toyota Corollas (SPB text):



Definition of Regression Trees

- Key: Regression trees are **flexible enough to accurately approximate almost any relationship between predictors & outcome** (by making the tree big enough)!
- Can the same predictor can appear in more than one split of the tree?

Training the Tree

Training:

- **Outcome:** Y **(continuous)**
  Predictors: $X_1, ..., X_p$ **with** n **observations in training data**
- **For a tree** T**, the predicted value** $\hat{Y}$ **in each leaf is taken to be:**
- **We want to learn the tree structure** T **from the data, i.e.:**

Training the Tree

Define:

- $|T| = \#$ terminal nodes ("leaves") in tree T
- $m = 1, ..., |T|$: terminal node index
- R_m : region in $\mathbb{R}^p$ corresponding to terminal node m
- N_m : $\#$ training observations in R_m

Training the Tree

Want a tree T that has low error on the training data, meaning low sum of squared error:

- The **predicted value** $\hat{Y}$ for leaf m is called:

- The **residual sum of squares** (also called the "sum of squared errors") on the training data is:

Training the Tree

We want a tree that is not too big, while having low RSS. Why?

We use a **criterion for picking a tree** that penalizes the size of the tree:

$$C_\alpha(T) = RSS(T) + \alpha|T|$$

for $\alpha > 0$

Training the Tree

- <u>But</u>: cannot simply evaluate $C_\alpha(T)$ for all possible trees
  - Why?

- Another difficulty: don't know what α to use

Pruning: removing a subtree

Solution:

Let $T \subseteq T_0$ be any tree that can be obtained from T_0 by "**pruning**" (removing a subtree)

- Instead of considering all trees, consider all trees $T \subseteq T_0$ where T_0 is a "full" tree that is good in some sense.
- e.g.: prune at $(\star)$

Training the Tree

Note: Pruning ALWAYS increases the *RSS*. Why?

Conversely, splitting a terminal node always decreases *RSS*.

Training a Tree

- A full tree means that none of its terminal nodes can be split any farther because it either
- We will obtain a "good" full tree as follows.

Growing the Tree

Growing the tree::

1. Start with the tree that has no splits (all observations are in the single terminal node)

2. Split at the variable j and cut point s that yields the largest decrease in $RSS(T)$

3.

4. Repeat until obtaining a full tree T_0.

Growing the Tree

Pruning the Tree

After growing the tree, we **prune it back using “weakest-link” pruning.**

The "**weakest link**" is the internal node that produces the smallest "per-node increase" in $RSS(T)$ when pruned; continue until we get the single-node tree.

$$\text{per-node increase} = \frac{\Delta \text{ in } RSS(T)}{\Delta \text{ in } |T|}$$

$$=$$

Weakest-link pruning:

Weakest-link pruning:

- Start at T_0
- Prune weak link to get $T(1)$
- Prune weak link of $T(1)$ to get $T(2)$
- Continue to get sequence of trees

$$T_0, T(1), T(2), ..., T(L), \text{ for some integer L}$$

- $T(L)$ is the tree with a single terminal node.

Weakest-link pruning

For any α, one can show that

- there is a unique tree $T_\alpha \subseteq T_0$ that minimizes C_α
- the sequence of trees $T_0, T(1), T(2), ..., T(L)$ obtained from T_0 by weakest link pruning **must** contain T_α!

Weakest-link pruning

Now we have to choose **one** of these trees. We can use:

- RMSE on validation data
- RMSE from cross-validation

Summary

Summary: Tree training by "grow & prune"

- Have a criterion $C_\alpha(T)$ that combines
  - how accurate tree is on training data ($RSS(T)$)
  - size $|T|$ of the tree
- Instead of considering all possible trees (too many), consider only trees $T \subseteq T_0$ for some good full tree T_0
- Get T_0 by growing the tree, decreasing $RSS(T)$ as much as possible at each step

Summary

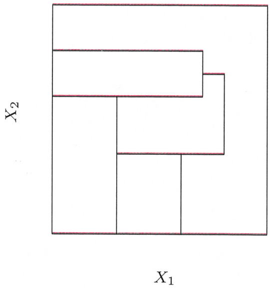
Procedure:

- Grow tree
- Prune, obtaining $T_0, T(1), T(2), ..., T(L)$
- Choose one of these trees by accuracy on validation data

Regression Trees

Consider applying regression trees with $p = 2$ continuous predictors and a continuous outcome variable.

T/F: the following partition of the predictor space could be obtained using regression trees:

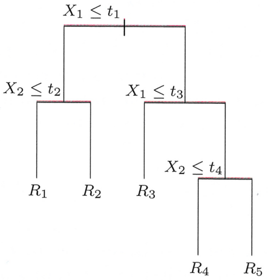


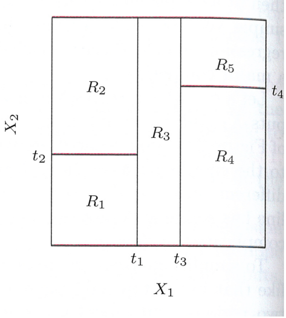
A: True; B: False

Plot from HTF text.

Regression Trees

Here's another example of regression trees for $p = 2$ predictors:

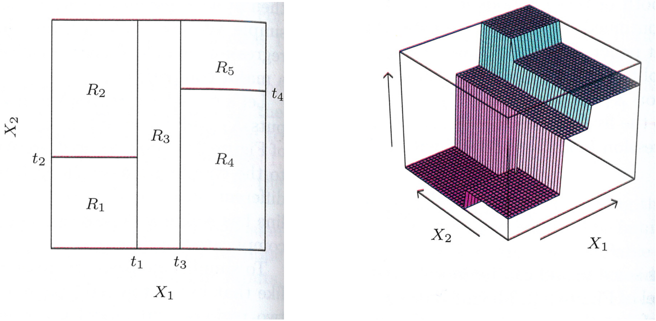




Plots from Hastie, Tibshirani, and Freedman (2009).

Regression Trees

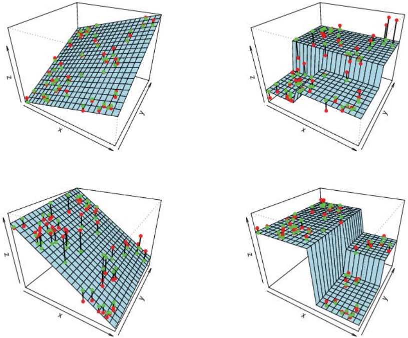
For the same example, here's a perspective plot of $\hat{Y}$ vs. X_1 and X_2:



Plots from Hastie, Tibshirani, and Freedman (2009).

Regression Trees

Now compare the surface of $\hat{Y}$ vs. X_1 and X_2, for a regression tree and for linear regression.



Plots from Rafael Irizarry's course notes.

Regression Trees

Advantage(s) of regression trees over linear regression:

Advantage(s) of linear regression over regression trees: