Regression and Classification Trees

Nate Wells

Math 243: Stat Learning

November 6th, 2020

Outline

In today's class, we will...

- Discuss classifcation trees for classification problems.

Section 1

Classification Trees

Classification Trees

Classification trees are very similar to regression trees, except the terminal nodes predict levels of a categorical variable, rather than values of a quantitative variable

Classification Trees

Classification trees are very similar to regression trees, except the terminal nodes predict levels of a categorical variable, rather than values of a quantitative variable

- To *grow* a classification tree, we need to make cuts based on a metric other than RSS (why?)

Classification Trees

Classification trees are very similar to regression trees, except the terminal nodes predict levels of a categorical variable, rather than values of a quantitative variable

- To *grow* a classification tree, we need to make cuts based on a metric other than RSS (why?)
- For each split candidate, we average the value of the metric on the two proposed subregions, and select the split that minimizes the average value of the metric.

Classification Trees

Classification trees are very similar to regression trees, except the terminal nodes predict levels of a categorical variable, rather than values of a quantitative variable

- To *grow* a classification tree, we need to make cuts based on a metric other than RSS (why?)
- For each split candidate, we average the value of the metric on the two proposed subregions, and select the split that minimizes the average value of the metric.
- The most natural choice is to use *Classification error rate E* (i.e. proportion of obs. in region not in most common class)

$$E = 1 - \max_k(p_{mk}) \quad \text{where } \hat{p}_{mk} = \text{prop. obs. in region m in class k}$$

Classification Trees

Classification trees are very similar to regression trees, except the terminal nodes predict levels of a categorical variable, rather than values of a quantitative variable

- To *grow* a classification tree, we need to make cuts based on a metric other than RSS (why?)
- For each split candidate, we average the value of the metric on the two proposed subregions, and select the split that minimizes the average value of the metric.
- The most natural choice is to use *Classification error rate E* (i.e. proportion of obs. in region not in most common class)

$$E = 1 - \max_k(p_{mk}) \quad \text{where } \hat{p}_{mk} = \text{prop. obs. in region m in class k}$$

  - But because of the greedy algorithm used to split trees, CER tends to overfit to noise

Other Decision Metric

Two common alternatives for decision metric:

Other Decision Metric

Two common alternatives for decision metric:

- The *Gini index G*:

$$G = \sum_{i=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \quad \text{where } \hat{p}_{mk} = \text{prop. obs. in region m in class k}$$

Other Decision Metric

Two common alternatives for decision metric:

- The *Gini index G*:

$$G = \sum_{i=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \quad \text{where } \hat{p}_{mk} = \text{prop. obs. in region m in class k}$$

  - It measures the rate that a random element would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the region

Other Decision Metric

Two common alternatives for decision metric:

- The *Gini index G*:

$$G = \sum_{i=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \quad \text{where } \hat{p}_{mk} = \text{prop. obs. in region m in class k}$$

  - It measures the rate that a random element would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the region
  - The Gini index is small if all $\hat{p}_{mk}$ are close to 0 or 1.

- The *cross-class entropy D*:

$$D = -\sum_{i=1}^{K} \hat{p}_{mk} \ln \hat{p}_{mk} \quad \text{where } \hat{p}_{mk} = \text{prop. obs. in region m in class k}$$

Other Decision Metric

Two common alternatives for decision metric:

- The *Gini index G*:

$$G = \sum_{i=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \quad \text{where } \hat{p}_{mk} = \text{prop. obs. in region m in class k}$$

  - It measures the rate that a random element would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the region
  - The Gini index is small if all $\hat{p}_{mk}$ are close to 0 or 1.

- The *cross-class entropy D*:

$$D = -\sum_{i=1}^{K} \hat{p}_{mk} \ln \hat{p}_{mk} \quad \text{where } \hat{p}_{mk} = \text{prop. obs. in region m in class k}$$

  - It measures the average amount of information conveyed by knowing the region of an observation.

Other Decision Metric

Two common alternatives for decision metric:

- The *Gini index G*:

$$G = \sum_{i=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \quad \text{where } \hat{p}_{mk} = \text{prop. obs. in region m in class k}$$

  - It measures the rate that a random element would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the region
  - The Gini index is small if all $\hat{p}_{mk}$ are close to 0 or 1.

- The *cross-class entropy D*:

$$D = -\sum_{i=1}^{K} \hat{p}_{mk} \ln \hat{p}_{mk} \quad \text{where } \hat{p}_{mk} = \text{prop. obs. in region m in class k}$$

  - It measures the average amount of information conveyed by knowing the region of an observation.
  - The entropy is small if all $\hat{p}_{mk}$ are close to 0 or 1.

Dealing with Categorical Variables

Both regression and classification trees can easily hand either quantitatitve or binary categorical varaibles.

Dealing with Categorical Variables

Both regression and classification trees can easily hand either quantitatitve or binary categorical varaibles.

- But with some modification, trees can also be used with multi-level categorical variables.

Dealing with Categorical Variables

Both regression and classification trees can easily hand either quantitatitve or binary categorical varaibles.

- But with some modification, trees can also be used with multi-level categorical variables.
- To do so, we recode all multilevel categorical variables as a sequence of dummy binary variables. Then proceed as usual.

Dealing with Categorical Variables

Both regression and classification trees can easily hand either quantitatitve or binary categorical varaibles.

- But with some modification, trees can also be used with multi-level categorical variables.
- To do so, we recode all multilevel categorical variables as a sequence of dummy binary variables. Then proceed as usual.
- But this conversion has a significant downside! The algorithm is biased toward making early splits on categorical variables with many levels.

Dealing with Categorical Variables

Both regression and classification trees can easily hand either quantitatitve or binary categorical varaibles.

- But with some modification, trees can also be used with multi-level categorical variables.
- To do so, we recode all multilevel categorical variables as a sequence of dummy binary variables. Then proceed as usual.
- But this conversion has a significant downside! The algorithm is biased toward making early splits on categorical variables with many levels.
    - Since trees are already prone to high variance, this additional bias can lead to unwanted increases in MSE.

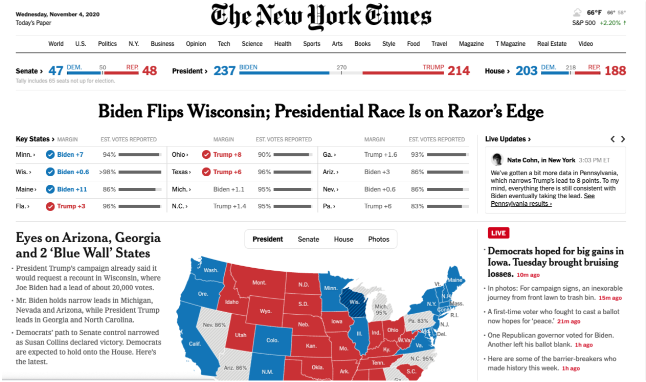
Trees for Classification Problems

Can we predict the winner of a presidential election based on demographics, state polling, economic conditions, and other features?

Trees for Classification Problems

Can we predict the winner of a presidential election based on demographics, state polling, economic conditions, and other features?

- No. Too stressful.



Wednesday, November 4, 2020
Today's Paper

The New York Times

66°F 66° 58°
S&P 500 +2.20% ↑

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Books Style Food Travel Magazine T Magazine Real Estate Video

Senate › 47 DEM. 50 REP. 48
Tally includes 65 seats not up for election.
President › 237 BIDEN 270 TRUMP 214
House › 203 DEM. 218 REP. 188

Biden Flips Wisconsin; Presidential Race Is on Razor's Edge

Key States ›	MARGIN	EST. VOTES REPORTED		MARGIN	EST. VOTES REPORTED		MARGIN	EST. VOTES REPORTED
Minn. ›	Biden +7	94%	Ohio ›	Trump +8	90%	Ga. ›	Trump +1.6	93%
Wis. ›	Biden +0.6	>98%	Texas ›	Trump +6	96%	Ariz. ›	Biden +3	86%
Maine ›	Biden +11	86%	Mich. ›	Biden +1.1	95%	Nev. ›	Biden +0.6	86%
Fla. ›	Trump +3	96%	N.C. ›	Trump +1.4	95%	Pa. ›	Trump +6	83%

Live Updates ›

Nate Cohn, in New York 3:03 PM ET
We've gotten a bit more data in Pennsylvania, which narrows Trump's lead to 8 points. To my mind, everything there is still consistent with Biden eventually taking the lead. See Pennsylvania results ›

Eyes on Arizona, Georgia and 2 'Blue Wall' States

- President Trump's campaign already said it would request a recount in Wisconsin, where Joe Biden had a lead of about 20,000 votes.
- Mr. Biden holds narrow leads in Michigan, Nevada and Arizona, while President Trump leads in Georgia and North Carolina.
- Democrats' path to Senate control narrowed as Susan Collins declared victory. Democrats are expected to hold onto the House. Here's the latest.

President Senate House Photos

LIVE

- Democrats hoped for big gains in Iowa. Tuesday brought bruising losses. 10m ago
- In photos: For campaign signs, an inexorable journey from front lawn to trash bin. 15m ago
- A first-time voter who fought to cast a ballot now hopes for 'peace.' 21m ago
- One Republican governor voted for Biden. Another left his ballot blank. 1h ago
- Here are some of the barrier-breakers who made history this week. 1h ago

Trees for Classification Problems

Can we predict the winner of a presidential election based on demographics, state polling, economic conditions, and other features?

- No. Too stressful.

Wednesday, November 4, 2020
Today's Paper

The New York Times

66°F 66° 58°
S&P 500 +2.20% ↑

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Books Style Food Travel Magazine T Magazine Real Estate Video

Senate › 47 DEM. 50 REP. 48
Tally includes 65 seats not up for election.
President › 237 BIDEN 270 TRUMP 214
House › 203 DEM. 218 REP. 188

Biden Flips Wisconsin; Presidential Race Is on Razor's Edge

Key States ›	MARGIN	EST. VOTES REPORTED		MARGIN	EST. VOTES REPORTED		MARGIN	EST. VOTES REPORTED
Minn. ›	Biden +7	94%	Ohio ›	Trump +8	90%	Ga. ›	Trump +1.6	93%
Wis. ›	Biden +0.6	>98%	Texas ›	Trump +6	96%	Ariz. ›	Biden +3	86%
Maine ›	Biden +11	86%	Mich. ›	Biden +1.1	95%	Nev. ›	Biden +0.6	86%
Fla. ›	Trump +3	96%	N.C. ›	Trump +1.4	95%	Pa. ›	Trump +6	83%

Live Updates ›

Nate Cohn, in New York 3:03 PM ET
We've gotten a bit more data in Pennsylvania, which narrows Trump's lead to 8 points. To my mind, everything there is still consistent with Biden eventually taking the lead. See Pennsylvania results ›

Eyes on Arizona, Georgia and 2 'Blue Wall' States

- President Trump's campaign already said it would request a recount in Wisconsin, where Joe Biden had a lead of about 20,000 votes.
- Mr. Biden holds narrow leads in Michigan, Nevada and Arizona, while President Trump leads in Georgia and North Carolina.
- Democrats' path to Senate control narrowed as Susan Collins declared victory. Democrats are expected to hold onto the House. Here's the latest.

President Senate House Photos

LIVE

- Democrats hoped for big gains in Iowa. Tuesday brought bruising losses. 10m ago
- In photos: For campaign signs, an inexorable journey from front lawn to trash bin. 15m ago
- A first-time voter who fought to cast a ballot now hopes for 'peace.' 21m ago
- One Republican governor voted for Biden. Another left his ballot blank. 1h ago
- Here are some of the barrier-breakers who made history this week. 1h ago

Trees for Classification Problems

Can we predict the species of a Portand tree based on its crown height and overall height?

Trees for Classification Problems

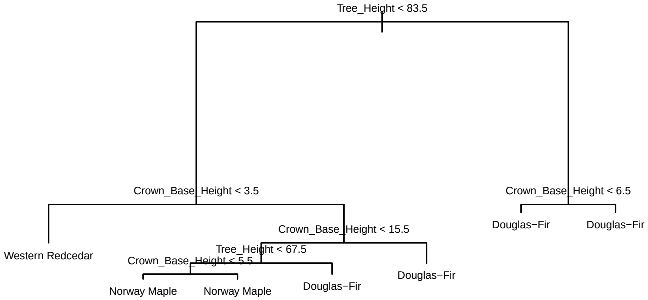
Can we predict the species of a Portand tree based on its crown height and overall height?

- YES!

Implementing classfication trees in R

As with regression trees, we use the `tree` package. We restrict our attention to the 3 most common tree species.

```
library(tree)
tree_model<-tree(Common_Name ~ ., data = common_trees)
plot(tree_model)
text(tree_model, pretty = 0, cex = .5)
```



Summary

We can also gather information on the model using the `summary()` function:

```
##
## Classification tree:
## tree(formula = Common_Name ~ ., data = common_trees)
## Number of terminal nodes:  7
## Residual mean deviance:  0.6324 = 5844 / 9242
## Misclassification error rate: 0.1153 = 1066 / 9249
```

Summary

We can also gather information on the model using the `summary()` function:

```
##
## Classification tree:
## tree(formula = Common_Name ~ ., data = common_trees)
## Number of terminal nodes:  7
## Residual mean deviance:  0.6324 = 5844 / 9242
## Misclassification error rate: 0.1153 = 1066 / 9249
```

- Here, the **deviance** reported is given by

$$-2\sum_m\sum_k n_{mk}\ln\hat{p}_{mk} \qquad \text{where } n_{mk} \text{ is number of obs. in region m in class k}$$

- Residual mean deviance is deviance divided by $n - |T_0|$.
- A small deviance indicates a good fit to *training* data

Pruning Classification Trees

We use `cv.tree` to prune (just like for regression trees).

Pruning Classification Trees

We use `cv.tree` to prune (just like for regression trees).

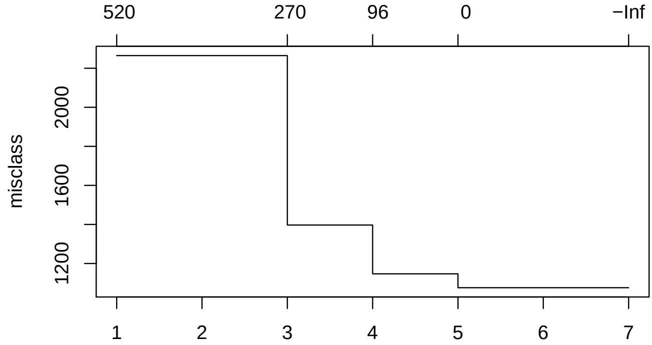
- While we use the Gini index or the entropy to grow the tree, it is still desirable to use misclassification rate to prune the tree

Pruning Classification Trees

We use `cv.tree` to prune (just like for regression trees).

- While we use the Gini index or the entropy to grow the tree, it is still desirable to use misclassification rate to prune the tree

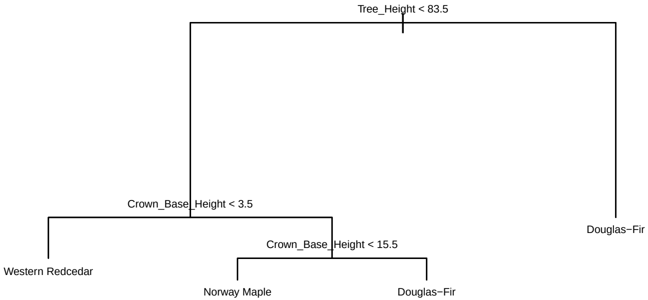
```
set.seed(1)
cv_tree_model<-cv.tree(tree_model, FUN = prune.misclass)
plot(cv_tree_model)
```



Pruning Trees, cont'd

We use the `prune.misclass` function to prune the trees to the desired number of nodes:

```
pruned_tree_model<-prune.misclass(tree_model, best = 4)
plot(pruned_tree_model)
text(pruned_tree_model, pretty = 0, cex = .5)
```



Misclassification

How well does the tree do on test data?

```
tree_preds<-predict(tree_model, common_trees_tst, type = "class" )
conf_mat<-table(tree_preds, common_trees_tst$Common_Name)
conf_mat
```

```
##
## tree_preds          Douglas-Fir Norway Maple Western Redcedar
##   Douglas-Fir              4709          124              137
##   Norway Maple              174          936              146
##   Western Redcedar          190           56              465
```

```
(sum(conf_mat) - sum(diag(conf_mat)))/sum(conf_mat)
```

```
## [1] 0.1192158
```