Classifiers

Rebecca Nugent
Department of Statistics, Carnegie Mellon University
http://www.stat.cmu.edu/∼rnugent/PCMI2016

PCMI Undergraduate Summer School 2016

July 11, 2016

What did we think about last time?

- Extending our Linear Model to include More Variables
- Updating Assumptions
- Regression Trees

Now we'll try

- flipping our regression trees into classification trees
- look at other ways to carve up feature space to label observations

Reminder of our Regression Tree

Use hyper-rectangles to partition feature space into subgroups of similar observations

- predictor variables (X) define the partitions
- response variable (Y) defines the closeness/similarity of the subgroups

Want to answer the questions:

- What variables are useful for prediction and separation?
- What values are useful "cutoffs"?

Hierarchical Binary (Decision) Tree Structure

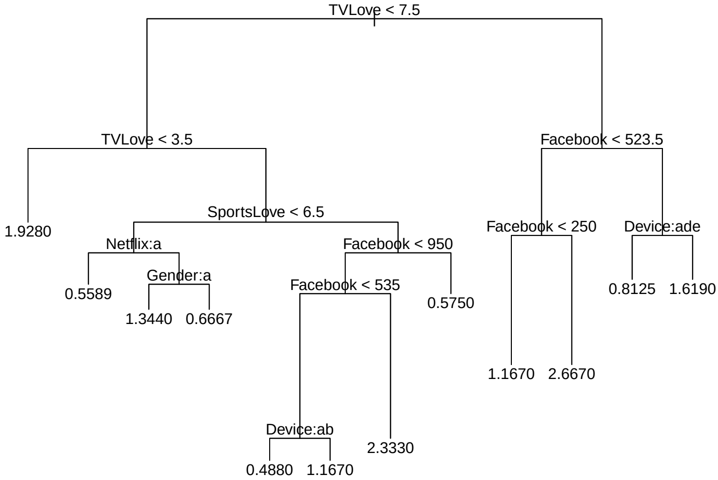
- Begin with root node/all observations
- Search for best split (based on reduction in $D = \sum(Y_i - \mu_i)^2$)
- Partition; recursively search for next best split on each "side"
- Stop splitting when size threshold or splitting criteria met
- In final leaves, each group assigned their average Y value

Reminder of our CMU Undergraduate Data

Surveyed about 100 students in Regression course;
interested in how much time spent watching TV/movies

- time spent watching TV/movies
- time spent doing HW
- how many classes
- gender
- major
- class year
- Facebook friends
- Netflix account
- how much do you love TV/movies (scale 1-10)
- average sleep per night
- device used to watch TV/movies
- how much you like sports (scale 1-10)
- favorite TV show

Back to our CMU Undergrads



Classification Tree

Can similarly partition our feature space using hyper-rectangles with goal of finding similar subgroups with respect to a categorical variable. In this context, often called a set of labels/classes.

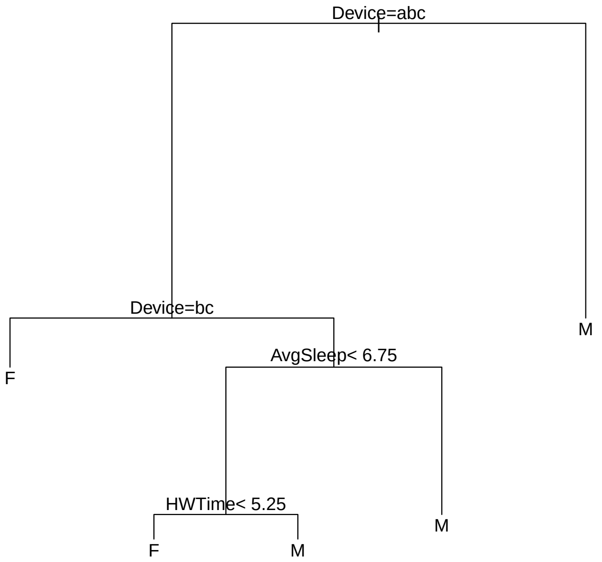
- Root node = all observations
- Search over feature space for "best split"
- Partition; recursively search again, etc
- Each final leaf/node m is assigned a set of class probabilities: $\hat{p}_{m1}, \hat{p}_{m2}, ..., \hat{p}_{mK}$, the proportion of observations in the node from each class
- Best split defined as optimizing the Gini Index

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

where m are our leaf nodes, $k = 1, ..., K$ the set of classes

Common to predict binary variable (two classes):
Gini $= 2p(1 - p)$ where $p =$ prob of one class

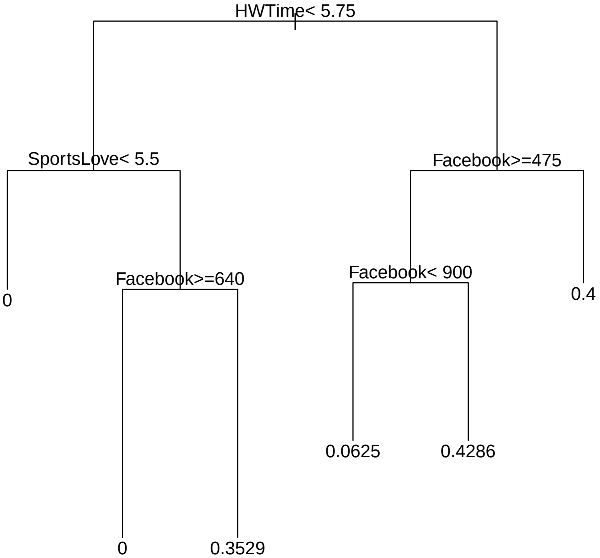
CMU Undergrads: Predicting Gender



More Detailed Tree Results

```
node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 98 46 F (0.5306122 0.4693878)
   2) Device=Computer,Laptop,Phone 82 32 F (0.6097561 0.3902439)
     4) Device=Laptop,Phone 46 13 F (0.7173913 0.2826087) *
     5) Device=Computer 36 17 M (0.4722222 0.5277778)
      10) AvgSleep< 6.75 20  7 F (0.6500000 0.3500000)
        20) HWTime< 5.25 9  1 F (0.8888889 0.1111111) *
        21) HWTime>=5.25 11  5 M (0.4545455 0.5454545) *
      11) AvgSleep>=6.75 16  4 M (0.2500000 0.7500000) *
   3) Device=Tablet,TV 16  2 M (0.1250000 0.8750000) *
```

CMU Undergrads: Predicting Math Major



More Detailed Tree Results

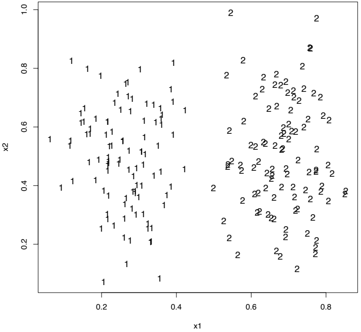
```
node), split, n, deviance, yval * denotes terminal node

 1) root 98 13.387760 0.1632653
   2) HWTime< 5.75 60  5.400000 0.1000000
     4) SportsLove< 5.5 27  0.000000 0.0000000 *
     5) SportsLove>=5.5 33  4.909091 0.1818182
      10) Facebook>=640 16  0.000000 0.0000000 *
      11) Facebook< 640 17  3.882353 0.3529412 *
   3) HWTime>=5.75 38  7.368421 0.2631579
     6) Facebook>=475 23  3.304348 0.1739130
      12) Facebook< 900 16  0.937500 0.0625000 *
      13) Facebook>=900 7  1.714286 0.4285714 *
     7) Facebook< 475 15  3.600000 0.4000000 *
```

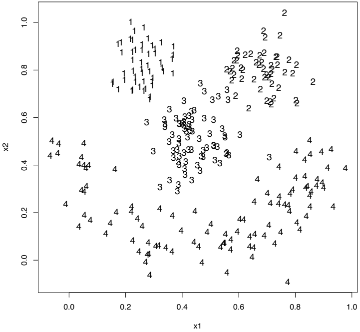
How else could we discriminate between classes?

It's really a question of how we could carve up our feature space to best separate the classes. So far, we've used hyper-rectangles.

What's the downside to that? What else might we use?

- Linear combinations of variables? *Linear Discriminant Analysis*
- Quadratic curves? *Quadratic Discriminant Analysis*





Class Posterior Probability

When estimating our decision rules, essentially estimating the *posterior probability* of class membership.

If the observation is in this location, what is the chance of it belonging to this class? $P(\text{Class } j|x)$?

Can use Bayes' Rule to figure out most likely class:

$$P(Class\ j|x) = \frac{\pi_j \cdot p_j(x)}{\sum_{l=1}^{L} \pi_l p_l(x)}$$

Choose the more likely class by ratio of class probs: $\frac{P(Class\ l|x)}{P(Class\ k|x)}$

Discriminant Analysis uses class labels to help separate classes by within-class and between-class variance; want to maximize ratio

Looking at our Bayes Rule, what do we need to estimate/assume?