ACTL3142

Week 8: Moving Beyond Linearity

Tetian Madfouni

t.madfouni@unsw.edu.au

Announcements

Assignment late penalties have been removed for the first 72 hours after due date. Check Ed for the specifics (it is a bit weirder than you might think, so please check)

Consultations sessions were also posted on Ed, go to any if you have questions

Recap: Decision Trees

What is a (simple) decision tree?

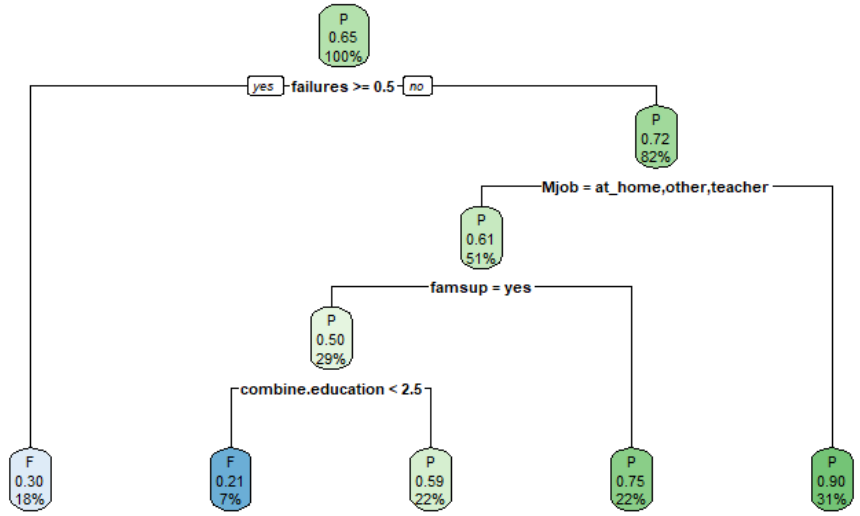
- Has a bunch of nodes with logical statements, if the logical statement is true, you follow the right (generally) path, if its false, you follow the left path. At the bottom of the tree (leaf nodes), you have a prediction for what the observation should be.

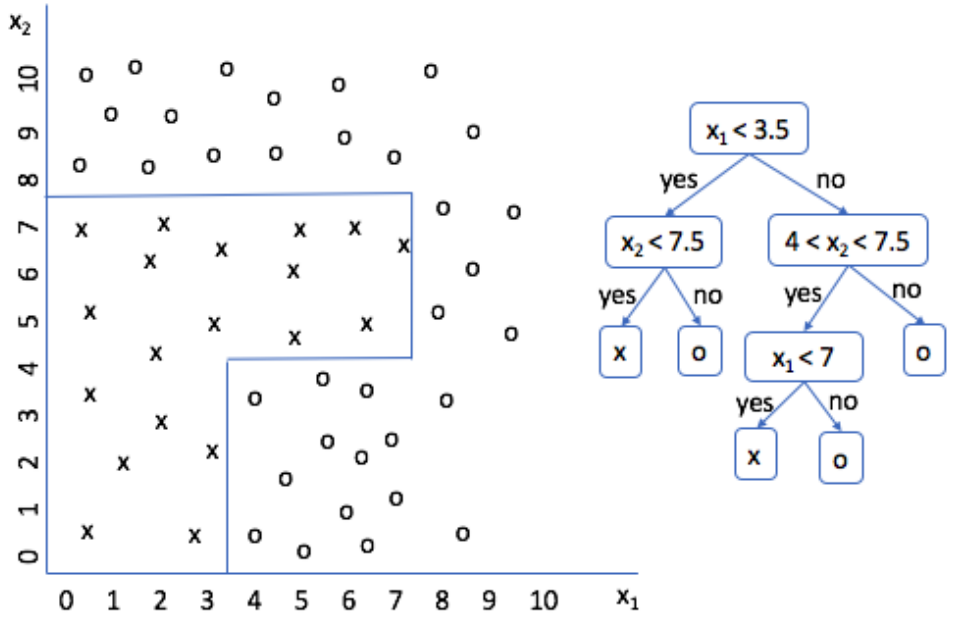
How are they made?

- We recursively find splits that maximise the decrease in some error measure (or just minimises the error at that step) until we meet some stopping criteria. Anyone remember the 3 main stopping rules?

What is pruning and why do we do it?

- Trees are very prone to high bias and high variance. Pruning is basically cutting down a tree to decrease variance while keeping a somewhat low bias





Recap: Ensemble Methods

What is an ensemble method and why are they useful for trees in particular?

- Basically methods that make many models (generally with high bias and low variance, often called weak learners) and aggregate their predictions. This generally reduces bias while increasing variance much less.

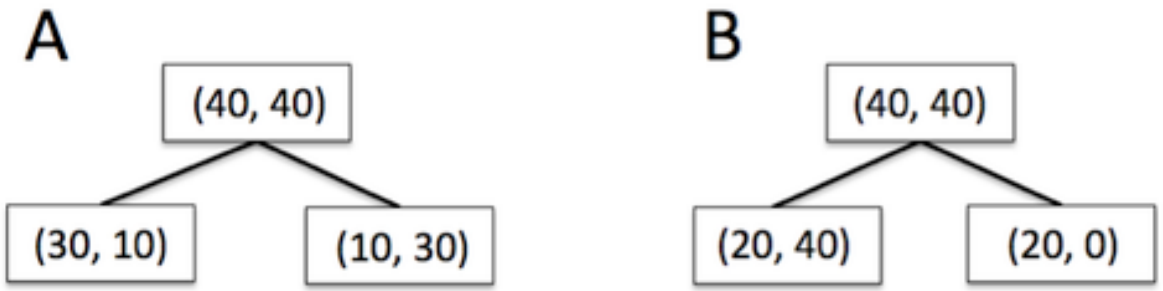
Three main ways presented in the lecture, what are they and how do they work?

- Bagging – bootstrap (what is this?) the data set B times to make B data sets. Train a model on each. To predict future observations, we take an aggregation of the B model's predictions. Important side note is out-of-bag error, what is this and why is it useful?
- Random Forest – same as above, except at each split we can only select a predictor to split on out of some $m < p$ predictors. This decorrelates each tree and reduces the variance further
- Boosting – uses sequential (rather than parallel) learning of each weak learner. At each step we fit a (very) weak learner to the *residuals* (initially these are just the values of the response), we then update the full learner by adding a shrunken version of the new tree. And we then update the residuals (residuals being difference between true response values, and the predictions of the full learner).

Short Recap: Error Measures

I mentioned measuring error at each split earlier, for regression trees we just use MSE, what about for classification?

- Can use mis-classification, however this tends to be “insensitive” and not great for growing trees. Refer to picture below and listen to me ramble
- We instead tend to use the Gini index or cross-entropy. There’s more on them in the lectures but they basically measure “node purity”. If you want to google more on this go for it, but practically there’s almost 0 difference between the two, Gini is computationally cheaper so most people opt for that



https://sebastianraschka.com/faq/docs/decision-tree-binary.html