Tree based models - Notes

Ivan Hanigan

January 23, 2014

1 Tree based methods

- We follow the methodology described in [Cutts et al., 2012] who used Decision Tree Analysis (DTA) and Random Forest Models together to infer differing levels of support for variables drawn from the different theories.
- Decision trees are the most widely used data mining method [Williams, 2011] also known as Classification and Regression Trees (CART), Recursive Partitioning trees (RPART) or Conditional Inference trees (CI).
- Tree models partition the dataset using binary splits based on the key variable that reduces the deviance of the "child" nodes compared to the "parent" node. The dataset is successively broken into smaller, more homogenous groups. Splitting continues with a test of all the potential splits at each step to find which variables produce the most homogenous subsets. The procedure continues until a stopping criterion is met such as a minimal reduction in deviance or a minumum number of observations in the resulting subset or others.
- Trees have become, popular multivariate tools for prediction as well as exploratory methods for identifying local structures in data sets (such as interactions), as well as alternatives to statistical descriptive methods like linear or logistic regression, discriminant analysis, and other mathematical modeling approaches [Ritschard, 2006].
- Random Forest models are similar to DTA but involve growing many trees and implementing other data mining techniques
- DTA models produce output that is simpler to interpret
- Cutts et al propse these statistical tools when the aim is testing hypotheses generated from multiple theories
- We did not use the first stage of Principal Components Analysis (PCA) because we felt a) too early to start controlling measurement error and potential collinearity of the variables; and b) the Tree methods adequately deal with the high dimensionality anyway.
- We also did not centre each variable on the grand sample mean so that the relative weight of each variable was even.

2 Statistical Software

- In contrast to [Cutts et al., 2012] we built traditional Regression Tree models [Hastie et al., 2001] using the rpart [Therneau and Atkinson, 2013] and tree (TODO cite Ripley) packages
- These methods give insights into which variables are most important and if there are any interactions
- We then fitted random forest models as additional test of explanatory power (TODO cite Brieman)
- We plan to follow the methodology of [Cutts et al., 2012] more closely in the next phase: implementing conditional inference trees (ctree from the party package - TODO cite Hothorn, Hornik, and Zeileis 2006).
- All analyses were conducted within the R statistical analyses software version 3.0.2

3 Statistical Methodology Notes

- The tree and rpart approach uses out-of-sample validation, whereas ctree does not.
- Out-of-sample validation techniques are usually preferred for complex, highly parameterized, statistical models such as decision trees.
- So results from ctree may be slightly less reliable for predictive purposes.
- However, for producing a descriptive model the approach used by both methodologies is highly suitable.
- The use of Random Forests also offers a very robust method when faced with "Large P, Small N" problems (many potential explanatory variables and not many observations).

4 References

References

[Cutts et al., 2012] Cutts, B. B., Moore, N., Fox-Gowda, A., Knox, A. C., and Kinzig, A. (2012). Testing Neighborhood, Information Seeking, and Attitudes as Explanations of Environmental Knowledge Using Random Forest and Conditional Inference Models. *The Professional Geographer*, (September 2013):121018062314008.

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning. 2nd Edition.*

[Ritschard, 2006] Ritschard, G. (2006). Computing and using the deviance with classification trees. In *Compstat 2006 - Proceedings in Computational Statistics 17th Symposium Held in Rome, Italy, 2006.*

[Therneau and Atkinson, 2013] Therneau, T. M. and Atkinson, E. J. (2013). An Introduction to Recursive Partitioning Using the RPART Routines. Technical report.

[Williams, 2011] Williams, G. (2011). Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery (Use R!).