Homework 13: COMPAS on CART

36-462/662, Fall 2019

Due at 10 pm on Tuesday, 26 November 2019 (**note date**)

We continue to work with the COMPAS data from last week, but now we are using the CART method to fit classification trees to it. I strongly recommend using the `tree` package; you may find the chapter on trees in *ADAfaEPoV* helpful.

1. (10) *Online questions* are online.
2. *Our first tree* Fit a classification tree to predict recidivism from age, sex, race, number of priors, and the degree of the offense the arrestee was charged with. Use the default settings for the minimum size and deviance-improvement of a split. (Make sure you fit a classification and not a regression tree.) Call this our "baseline tree".
   (a) (5) Plot the resulting tree, showing which features are split on at each node (and at what level), and the probability of recidivism at each leaf.
   (b) (5) What features does the tree actually use? Why does it not use some of the features in its formula?
   (c) (5) Describe how the tree will assess the risk of violence for each of the following arrestees: individuals:
      - Archie, a 19 year old white male with one prior, charged with a felony.
      - Betty, a 22 year old white female with two priors, charged with a misdemeanor.
      - Chuck, a 34 year old black male with no priors, charged with a misdemeanor.
      - Veronica, a 42 year old white female with 12 priors, charged with a felony.
3. *Error rates* In this problem, use the baseline tree you grew in the previous problem.
   (a) (5) Assume we set a probability threshold of 0.5 for classification, which is what we would do to maximize accuracy and/or if we viewed

false negative and false positive errors as equally costly. What prediction ($\hat{Y}$) would we make at each leaf? What prediction would we make for Archie, Betty, Chuck and Vernoica?

(b) (5) At a threshold of 0.5, what would our false positive rate be? What would our false negative rate be?

(c) (5) Suppose we set a threshold of 0.9, so we'd have to be 90% certain that someone would recidivate before we'd predict that they would. (This would be consistent with worrying much more about false positives than about false negatives.) What prediction would we make at each leaf? What predictions would we make for Archie, Betty, Chuck and Vernoica? What would the false negative and false positive rates be?

(d) (5) Suppose that we set a threshold of 0.01, so we'd be willing to predict recidivism if there was even a 1% chance of it. (This would be consistent with thinking false negatives were much worse than false positives.) What predictions would we make at each leaf? What predictions would we make for Archie, Betty, Chuck and Veronica? What would the false negative and false positive rates be?

(e) (5) Consider a range of thresholds between 0.9 and 0.01, and plot the false negative rate and false positive rate for each threshold. (Ideally, you should have a two-dimensional plot, with false positive rate on one axis, and false negative rate on the other axis.) Is there a trade-off between the two error rates? What is the lowest false negative rate we could achieve while keeping the false positive rate under 20%?

4. *Pruning* Fit a classification tree to predict recidivism from age, sex, race, number of priors, and the degree of the offense the arrestee was charged with. Set the minimum size of a split to 1 case, and the minimum deviance improvement to 0. Call this the "maximal tree".

   (a) (5) Plot the tree (but don't label it). How many leaves does the tree have?

   (b) (5) Using `prune.tree()` and `cv.tree()`, plot the number of misclassifications for all the prunings of the maximal tree, down to one leaf. Why is this line flat?

   (c) (5) Using `prune.tree()` and `cv.tree()`, plot the "deviance" for all the prunings of the maximal tree. Explain how "deviance" here relates to the entropy. (*Hint:* Read the chapter on trees in *ADAfaEPov.*) What is the optimal number of leaves?

   (d) (5) Find the pruning of the maximal tree with the optimal number of leaves. Compare it to the baseline tree you grew with the default setting — how (in words) are the two trees similar or different?

5. *Fairness* In this problem, use the baseline tree.

(a) (5) Does the baseline tree use any protected categories to make its predictions? That is, does it meet the basic form of anti-classification?

(b) (5) Plot the distribution of predicted probabilities of recidivism for (i) whites, (ii) blacks, (iii) men and (iv) women. Are the predictions independent of protected categories?

(c) (5) Set a probability threshold of 0.1, i.e., we predict $\hat{Y} = 1$ when we're at least 10% sure that someone will recidivate. What are the false negative and false positive rates for (i) whites, (ii) blacks, (iii) men and (iv) women? Does this meet the requirement of parity in error rates?

(d) (5) Again, set a probability threshold of 0.1. Do the predictions of the tree satisfy demographic parity for blacks and whites, and for men and women?

Rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output.

Extra credit (5 points): Repeat the tree-growing exercise, but add the 1–10 COMPAS score (`v_decile_score`) as a potential feature. What variables does the result tree use? Is the new tree a better predictor than the tree you grew without using the COMPAS score? How can you tell? If the tree uses any features other than the COMPAS score, is that good or bad for the makers of COMPAS?