Homework 3: COMPAS, CART, Forests

36-462/662, Spring 2020

Due at 10 pm on Thursday, 6 February 2020

AGENDA: Practice with classification trees, and with ensembles of trees.

Our data set this week comes from the analysis, performed by the news organization ProPublica, of the "COMPAS" risk prediction scores performed by ProPublica for Broward County[1], Florida. (We will look specifically at ProPublica's analysis, and the controversy it led to, later in the course; you're welcome to read more about it now, but it's not needed for this assignment.) Specifically, our data file tracks the following information:

- The age of each suspect;
- Their age, binned into categories;
- Their sex;
- Their race;
- Their COMPAS score[2] for risk of violence (1–10);
- Their COMPAS score, binned into categories of "Low" risk (1–4), "Medium" (5–7) or "High" (8–10);
- Their COMPAS score, binned into categories of "Low" (1–4) and "Medium or High" (5–10);
- Whether they were charged with a felony (F) or misdemeanor (M)[3];
- Count of priors[4].
- Whether they had a subsequent conviction for violence within two years.

[1] Mostly: Fort Lauderdale, in the greater Miami metropolitan area.

[2] COMPAS calculates separate scores for risk of "failure to appear" at trial, risk of committing any type of crime, and risk of violence. This is specifically the score for risk of violence.

[3] American law distinguishes between two kinds of crimes. Felonies are more serious crimes, punishable by (in most states) a year or more of imprisonment, or, in some cases, death. Misdemeanors are punishable by shorter terms of imprisonment (typically in city or county jails rather than prisons) and/or fines. Most crimes of violence are felonies, but not all felonies are crimes of violence: fraud, drug dealing, and tax evasion, for instance, are all felonies.

[4] This appears to be the count of prior *convictions* (not just arrests).

Unless explicitly asked otherwise, in this problem set, do not use any of the COMPAS score features.

In an experiment, this week we do not have online reading questions. I do, however, strongly recommend reading the chapter on trees in *ADAfaEPoV*, and the posted lecture notes from 30 January.

1. *Understanding*

   (a) (3) In a few sentences, using your own words, describe the data set in a way which should be comprehensible to a non-statistician. (You may want to actually look at the data file first.)

   (b) (3) In a few sentences, using your own words, explain why one would want to build a statistical model to predict the risk of violence from features like this.

   (c) (3) In a few sentences, using your own words, explain why this assignment is *not* using the COMPAS score.

2. *Data prep* Filter the data so that it only contains black and white arrestees. Pick a random 20% of the remaining data points and set them aside as the testing set; the other 80% of the data will be the training set. In all subsequent problems, fit all models on the training set, and, whenever an evaluation of error is called for, use the testing set.

   (a) (3) Explain, in your own words, why it is important to only evaluate predictions on the testing data.

   (b) (2) Why might we run in to trouble if we randomly divided the data into training and testing sets *without* limiting the data to the two largest racial groups?

3. *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) (4) 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:

      - 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.

4. *Error rates* In this problem, use the baseline tree you grew problem 3.

   (a) (4) 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 classification ($\hat{Y}$) would we make at each leaf? How would we classify Archie, Betty, Chuck and Vernoica (from problem 3c)?

   (b) (3) At a threshold of 0.5, what would our false positive rate be? What would our false negative rate be?

   (c) (5 Pick a range of thresholds between 0.9 and 0.05, and calculate the false negative and false positive rate at each threshold. Display your results as two-dimensional plot of FNR against FPR. 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%?

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

   (b) (4) 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?

6. *Some theory for why ensembles help* Suppose we are trying to estimate a parameter μ, and have b different estimates, $M_1, \ldots M_b$, whose mean would be $\overline{M} = b^{-1} \sum_{k=1}^{b} M_k$. The variance across the estimates would be

$$V \equiv \frac{1}{b} \sum_{k=1}^{b} (M_k - \overline{M})^2$$

(a) (5) Show that the squared error of the average is the average of the squared errors, *minus* the variance of the estimates:

$$(\mu - \overline{M})^2 = \left(\frac{1}{b}\sum_{k=1}^{b}(M_k - \mu)^2\right) - V$$

Hint: In the definition of V, add and subtract μ, expand the square, and simplify.

(b) (5) Explain, in your own words, what this algebra has to do with the advantages of using multiple predictive models, rather than just one.

7. *Bagging* Using `adabag`, or any other suitable package, use bagging to fit an ensemble of 100 classification trees to the COMPAS data. Use the same features as in problem 3. (On my computer, doing this takes about 30 seconds, so you will probably want to cache at least this chunk of your R code.)

   (a) (5) Pick the first five trees from the ensemble and plot them. How do they compare to each other? How do they compare to the tree you grew in problem 3? Does it matter that you looked at the *first* five trees?

   (b) (4) How will your ensemble assess the risk of violence for each of the arrestees from problem 3c?

   (c) (5) Find the prediction of the ensemble for every arrestee in the testing set. (If you use `adabag`, you can get this from the `class` component of `predict.bagging`.) Report the confusion matrix, and use it to calculate (i) the over-all error rate, (ii) the false positive rate, (iii) the false negative rate, and (iv) the positive predictive value.

8. (a) (5) For every arrestee in the testing set, calculate the probability of recivism predicted by the bagging ensemble. (If you use `adabag`, you can get this from the `prob` component of `predict.bagging`.) Apply a series of thresholds to this probability from 0 to 1, and, for each threshold, plot the combination of false negative rate (on the vertical axis) and false positive rate (on the horizontal 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%?

   (b) (5) What would a plot of FNR against FPR look like for a really good classifier? For a really bad one? How good or bad is your baseline tree, by this standard? How about your bagged ensemble of tress?

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. When a problem asks for a plot or table, the plot or table is automatically generated by code included in the source file, rather than being made by hand in an external program, or hard-coded. 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.