Homework 14: Bagging and Boosting

36-462/662, Fall 2019

Due at 10 pm on Wednesday, 4 December 2019

We continue to work with the COMPAS data, but now we investigate using bagging and boosting; you will find it useful to read the notes from the last lecture before Thanksgiving.

There are no online questions this week.

There are a number of R packages available for bagging and boosting; I recommend using the `adabag` package, not least because it's very thoroughly documented (`https://www.jstatsoft.org/article/view/v054i02`).

1. *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) (5) 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?

2. *Bagging* Using `adabag`, or any other suitable package, use bagging to fit an ensemble of 100 classification trees to the COMPAS data. You are, as usual, trying to predict two-year recidivism, and should use, as your predictors: race, sex, age, number of priors, and whether the arrestee was charged with a felony or a misdemeanor. (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 learned in the previous homework? Does it matter that you looked at the *first* five trees?

   (b) (5) How will your ensemble 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.

(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.

(d) (5) Repeat the calculations of error rate, false positive rate and false negative rate separately for blacks and for whites. Which forms of parity (if any) does the bagged ensemble satisfy?

3. *Boosting* Using `adabag`, or any other suitable package, use boosting to fit an ensemble of 100 trees to the COMPAS data. (If it matters, you want the AdaBoost.M1 variant of boosting.) Use the same set of features as in the bagging problem above. Limit the trees to a depth of 1. (If you're using `adabag`, look carefully at the examples of controlling maximum depth in `help(boosting)`.) Again, this can take a while, so you'll want to cache this code chunk.

   (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 learned in the previous homework? Does it matter that you looked at the *first* five trees?

   (b) (5) The trees in the boosted ensemble should be notably different from the trees in the bagged ensemble. Describe the difference, and explain (in your own words) why it exists.

   (c) (5) How will your ensemble assess the risk of violence for each of the arrestees in problem 2b?

   (d) (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.boosting`.) 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.

   (e) (5) Repeat the calculations of error rate, false positive rate and false negative rate separately for blacks and for whites. Which forms of parity (if any) does the boosted ensemble satisfy?

4. *ROC curves*

(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).

(b) (5) Repeat the previous problem, but now for the boosting ensemble.

(c) (5) Plot the false negative rate against the false positive rate for thresholding the numerical, 1–10 COMPAS score. Why are the only sensible thresholds the integers from 0 to 11?

(d) (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 are the bagging and boosting ensembles, by this standard? How good or bad is the COMPAS score itself?

5. *Feature ablation*

(a) (3) Repeat the bagging estimation of problem 2, but using only priors and the degree of the charge as predictive features. (That is, do *not* use the protected categories of race, sex or age as predictors.) What is the new error rate? What proportion of arrestees in the testing set had their classification change?

(b) (3) Repeat the boosting estimation of problem 3, but again, only use prior count and the degree of the charge as predictors. What is the new error rate? What proportion of arrestees in the testing set had their classification change?

(c) (4) Repeat the plots of FNR vs FPR from problem 4 for the new bagged and boosted ensembles.

(d) (5) Over-all, how much worse do the ensembles predict when limited to the un-protected features of priors and charge degree, as opposed to when they had access to the protected features of race, sex and age? Do the limited ensembles predict better or worse than COMPAS?

(e) (3) Explain why, in your own words, this exercise is called feature "ablation".

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.