COMPAS, CART, kNN

36-462/662, Spring 2022

Due at 6 pm on Thursday, 3 March 2022

Agenda: Practice with classification trees, nearest neighbor classifiers, and logistic regression again.

Our data set this week comes from the analysis, performed by the news organization ProPublica, of the "COMPAS" risk prediction scores as used in 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.) COMPAS is a a complicated and propietary algorithm, developed and sold by a company called NorthPointe to local governments across America, which is used to assess the likelihood that people who have been arrested will commit violent crimes if released before their trial[2]. The company does not say exactly how COMPAS works, just that it's a statistical model based on over 100 features.

Specifically, our data file tracks the following information:

- The age of each suspect;
- Their sex;
- Their race;
- Their COMPAS score for risk of violence (1–10);
- 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.

Do *not* use the COMPAS score in this problem set, except in the extra credit questions.

In an experiment, this week we do not have online reading questions. I do, however, strongly recommend reading the chapter on trees *ADAfaEPoV*.

1. **Understanding**
   a. (2) 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. (2) 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. (2) In a few sentences, using your own words, explain why this assignment is *not* using the COMPAS score.
   d. (2) COMPAS, and related models, are mainly used to assess whether it's safe to release someone who has been arrested, before the trial which determines whether they're actually guilty. The outcome variable in this data set is re-arrest and conviction for a violent crime within two years. Does this seem like a good way of measuring pre-trial release safety? Explain.

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

[2] COMPAS actually calculates separate scores for risk of "failure to appear" at trial, risk of committing any type of crime, and risk of violence. This data set only contains 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 state 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).

2. **Data prep** Pick a random 20% of the data points and as a testing set; the other 80% of the data will be the training set. In all subsequent problems, fit all models on the training set. If you need to do any cross-validation, do it *within* the training set. Whenever an evaluation of predictions is called for, use the testing set.
    a. (2) In your own words, why it is important to evaluate predictions on the testing set *alone*?
    b. (2) In your own words, why it is important to *randomly* divide the data?
    c. (2) What is the proportion of people in the testing set who are re-arrested for violence within two years? What is the maximum error rate of any useful classifier here?
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. (4) What features does the tree actually use? Why does it not use some of the available features?
    c. (4) Describe how the tree will assess the risk of violence for each of the following people, arrested after a brawl at the Riverdale Diner:
        - 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 Hispanic female with 12 priors, charged with a felony.
4. **Error rates** In this problem, use the baseline tree you grew Q3.
    a. (3) Set a probability threshold of 0.5 for classification, to maximize accuracy. What classification ($\hat{Y}$) would we make at each leaf? How would we classify Archie, Betty, Chuck and Veronica (from Q3c)?
    b. (3) At a threshold of 0.5, what are the false positive and false negative rates?
    c. (3) Pick a range of thresholds between 0.95 and 0.05, and calculate the mis-classification rate at each threshold. Plot the error rate as a function of the threshold. Is the error rate minimized at a threshold of 1/2?
    d. (3) Using the same range of thresholds as in Q4c, calculate the false negative and false positive rate at each threshold. Plot the FNR as a function of the FPR. Is there a trade-off between the two error rates?
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. (1) Plot the maximal tree (but don't label it). How many leaves does the tree have?
    b. (3) Using `prune.tree()` and `cv.tree()`, plot the number of mis-classifications for all the prunings of the maximal tree, down to one leaf. (Hint: Read the chapter on trees in *ADAfaEPoV*.) Why is this line flat?
    c. (3) Using `prune.tree()` and `cv.tree()`, plot the "deviance" for all the prunings of the maximal tree. How does this "deviance" relates to the entropy? (Hint: Read the chapter on trees in *ADAfaEPov*.) What is the optimal number of leaves?
    d. (3) 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. **Logistic regression plus lasso, again** Using `glmnet()`, as in HW 5, fit a logistic regression with a lasso penalty for predicting recidivism, with the same predictors that were available to the tree in Q3. *Hint*: `glmnet()` needs its `x` argument to be a matrix, where all the columns have to be of the same type, so you will need to convert the categorical variables to numerical codes (e.g., 1 for male and 0 for female, or the other way around). When you do this, you should either code race as binary, white/non-white, or create one fewer indicator variables than there are racial categories in the data. *Note*: If you can't get `glmnet()` to work here, you can get partial credit for just using `glm()` to fit a logistic regression, with no regularization, but be clear in your write-up that this is what you did.
    a. (3) Use `cv.glmnet()` to pick a good value of λ, by cross-validation within the training set. This

reports *two* values of λ, one that minimizes the error, and the largest value of λ where the CV score is within 1 standard error of the score at the minimum. Why might we prefer that larger value of λ? What is that value of λ here?

b. (3) Using the value of λ from Q6a, run `glmnet()` to get the actual regularized logistic regression. Which variables have non-zero coefficients at that value of λ? How does that compare to the variables selected by the baseline tree in Q3?

c. (3) What probability of violence would this model give to each of the four arrestees from Q3c? How would it classify each of them, at a threshold of 1/2?

d. (3) Plot accuracy as a function of the probability threshold, for a range of thresholds from 0 to 1 (as in Q4c). Is the accuracy maximized when the threshold is 1/2? Should it be?

e. (3) For the same range of thresholds, plot the false negative rate versus the false positive rate, for the same range of thresholds (as in Q4d). Is there a trade-off between the two types of error?

7. **Nearest neighbors** We'll build a k-nearest-neighbors classifier, using only age and priors, since the `FNN` package described in the notes for lecture 11 likes its features numerical.

   a. (2) Standardize the features of age and prior count to have mean 0 and variance 1. *Hint*: `scale()`. Report a table of the mean and standard deviation of both variables (before standardizing!). Why does standardizing make sense here?

   b. (2) How, with $k = 1$, would you classify the arrestees from Q3c? Describe in a table the nearest neighbor of each of the four arrestees, giving the (raw, un-standardized) values for all the variables. *Hint*: Apply the *same* transformation to age and priors on this little four-row test set. (That is, make sure you're subtracting the training-set means, then dividing by the training-set standard deviations.)

   c. (3) Plot the mis-classification rate for a range of k from 1 to 50 (as in Q4c). What k is most accurate on the testing set? *Hint*: Same as Q7b.

   d. (3) Plot the false negative rate as a function of the false positive rate, over the same range of k as in Q7c. Is there a trade-off between the two kinds of error?

   e. (3) Use cross-validation within the training set to pick k. Does this give the value of k which is most accurate on the testing set? If not, is it close?

8. **False positive control again** (5) For each of the three methods (trees, logistic regression, kNN), what's the best false negative rate we can get, while keeping the false positive rate $\leq 20\%$? (If you think a method can't get its FPR below 20%, say so.) Why, in the context of pre-trial safety assessment, might we want to minimize the false negative rate, while limiting the false positive rate?

9. **Recommendations** (8): The government of Riverdale County is considering adopting a statistical model to screen arrestees. Which of these models, if any, would you recommend to the county government, and why? (You can assume Riverdale County is very similar to Broward County, where the data were gathered.)

10. **Timing** (1): How long, roughly, did you spend on this assignment?

Presentation 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. 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 plots and tables are generated by code included in the R Markdown file. 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. All parts of all problems are answered with actual coherent sentences, and raw computer code or output are only shown when explicitly asked for. Text from the homework assignment, including this rubric, is included only when relevant, not blindly copied.

Extra credit: COMPAS The COMPAS score is included in this data set; it's a number between 1 and 10. The company which markets COMPAS does not say that it translates directly into a probability for violence, but it does describe those with scores of 1–4 as "low risk", those with scores of 5–7 as "medium risk", and scores of 8 or above as "high risk".

a. (2) Classify each arrestee as "violent" if their score $\geq t$, and plot the resulting error rates, for all t from 1 to 11 inclusive. (Why does this one need to go to eleven?) At what threshold is the error

rate minimized? Is the minimum error rate for COMPAS better than that of the baseline tree, or the best you can attain with the logistic regression or kNN?

b. (2) For the same range of thresholds as in ECa, calculate the false negative and false positive rates, and plot FNR as a function of FPR. Describe the trade-off between the two error rates.

c. (2) Create a single plot showing the FNR-vs-FPR curves for the baseline tree, for the regularized logistic regression, for kNN, and for COMPAS. Does COMPAS have a better curve than any of models you have fit here? What does "better" mean here?

d. (4) Continuing the scenario from Q8, would you recommend that Riverdale County adopt COMPAS? Explain.