STATISTICS 452/652: Statistical Learning and Prediction

November 28, 2020

PROJECT 2

Due Friday, Dec 4, 11pm, Submission details to be announced

POLICY

1. This project is to be completed *independently*, with no outside help except for a named teammate, if you have been assigned one. You may use whatever class materials you wish in completing this assignment. **BUT DO NOT DISCUSS QUESTIONS OR RESULTS WITH ANYONE ELSE, WITHIN OR OUTSIDE OF THE CLASS**. Failure to follow this directive will result in a failing grade.

2. Late projects will be accepted at a penalty of 2 points/hour (it's a 100 point project), *strictly enforced*.

ASSIGNMENT

The data for this project will be posted to Canvas at 5PM on November 27. This file will be a comma-separated file (csv) that contains n observations of p explanatory variables (labeled `X1`–`Xp`) and one categorical response (`Y`). That's all I'm telling you about the data.

Your job is to develop a method for classifying `Y` based on `X.` You may use any of the techniques covered in class. Methods not covered in class may not be used.

DELIVERABLES

You will produce three required items.

1. I will post a test set of explanatory variables without the response variable attached. You will return a list of predicted values, *in the same order*. The list should be *one column of numbers with no row numbers and no column header*. Use
```
write.table(predictions, file.name, sep = ",", row.names = F, col.names =
F)
```
to submit your code, where `predictions` is the name of the vector containing your predicted values and `file.name` is the location on your computer where you want to

store the results. These will later be uploaded into a Shiny app that we will introduce later.

Look at your file before you submit it to make sure that the format is correct (and also to make sure that the predicted values are sensible!).

2. You will supply a written report answering questions relating to the steps you took to create your model and predictions. *You will also provide a measure of test error and a test confusion matrix.* This gets posted to Crowdmark. Details are given below. **PLEASE ANSWER EACH QUESTION ON A SEPARATE PAGE!** They will get uploaded into separate slots in Crowdmark.

3. You will provide the R code that turns the test X into the predicted Y using *only your final model.* That code should not contain other models that were not directly part of the final predictions. It should not contain additional tuning loops or comparisons of models that do not directly produce the final predictions. This gets posted to Canvas.

REPORT

Your report, which is submitted to Crowdmark, should answer these questions, as numbered below:

1. *What models or machines did you attempt to fit?* For each one, paste the R code from your program for the initial successful model fit. I want to see what you tried. For example, "`fit1 = lm(y~., data=train)`" is what I would list if I used multiple linear regression on all of the variables and my training data were called "train". **The answer should be a list (e.g., with bullets) of nothing but the code for each of these model fits.** Don't list code that did not run. If tuning was involved in the initial fitting process, you can can paste the function with variable names for the tuning parameters (e.g., your function might have "`mtry=mm`" if you looped over a variable called "`mm`").

2. *What process(es) did you use to evaluate and compare models and to select your final model?* I am thinking of Lecture 3, specifically: **Give 1-2 sentences explaining the method, the quantity, how results were turned into decisions.** For example, "I used 50,000 bootstrap resamples, fit all models to each resample, and used largest training error from last resample as my best model." (This example answer is complete, but represents something rather stupid to do...)

3. *Did you tune any methods?* If so, (a) what process(es) did you use to evaluate and compare models and to select your final model (i.e., **I want to see an answer like to the previous question, but relating to how TUNING was done**), and (b) **for each method list all parameter values that were considered** (e.g., "For "Blasting" I use a grid of values with A=(1, 2, 3, ... , 60) and B=(0.00317, sqrt(3.14159)). For "Blooming" I used combinations of (z, γ)=(0.1, 3), (0.5, 6), and (1.1, 12)). I expect maybe 1-2 sentences for each method tuned.

4. *(NEW) Did you do anything else that helped you to complete your analysis?* You can add a **couple of sentences** here if you want to say what you did that is not covered by the questions above. Please don't write a page!!!

5. *What was your chosen prediction machine?* **Paste the code that produced your predicted values, including all values of tuning parameters if any, random number seeds, and explaining any variable names that are not obvious.** I should be able to run your code and produce the same results (or extremely similar if randomization is used). If I try and it doesn't work, there will be a major deduction.

6. **Report your estimate of test error and test confusion matrix.** All that is needed here is the numeric value for your test error (as a proportion and not a percentage), and the confusion matrix **with observed classes along the side and predicted classes along the top.** This is meant to show that you have some idea what to expect from your machine when it is applied to a new test set.

7. *(optional)* **List the variables that you believe are important.** A positive bonus will be given for each correct results. A deduction will be made for each variable listed that was not important.

The main thing here is that I should be able to see what your thought process was and whether you considered (or failed to consider) important ideas.

GRADES

Your grade will be based partly on how well your model performs, and partly on the steps you took to get there. I will compute the misclassification rate between your predicted values and the test set responses. *I will scale these against the best model produced by a member of the class, so this is a competition!* If your prediction machine improves misclassification rate from the naive classifier by only 80% as much as the best machine in the class, your mark for this part will be 80%.

The other part of your grade will depend on your report as described above. This portion will count for 60% of your grade. The remaining 40% will come from your model's performance.

If you supply a list of variables that are in the model, I will give a small bonus for each variable in your list that was part of the true structure that made the data. *I will subtract the same amount for each variable on your list that is not a part of the true structure!*

FINAL COMMENTS

GOOD LUCK, HAVE FUN, and may your machine be much better than rolling a 5-sided die.