STAT 452/652 Fall 2023 Project 2

Owen G. Ward

New statistical machine learning tools are developed frequently. As such a key skill is being able to understand and implement techniques which you have not studied previously.

In this project you are required to independently study a classification tool which we did not study in this class and write a brief report illustrating how it is used. You should choose a dataset to use throughout your report which will aid in illustrating the topic and the benefits of the method. This report is worth 15% of your overall grade (15 points).

You will submit a brief (max 6 pages) report. In **Section 1** you will describe how **ROC curves** are computed and how they can be used for assessing a binary classifier (for 7 points). This is discussed in ISLR in Chapter 4.4.2. In **Section 2** you will describe several key components of classification using **Support Vector Machines (SVM)** (for 8 points) You should provide an introduction to SVMs for binary classification. The relevant material in ISLR is Chapters 9.1-9.3 but you are welcome to use other sources for both sections.

You must follow the academic integrity policy and anything you write must be original. The detail required for these methods should be at a similar level to other methods we cover in this class. In particular, aim that your exposition should be readable by someone taking this class. If any text or resource other than ISLR is used than these should be referenced. Any standard referencing format is fine. You should not discuss this with other students in the class.

Your report must address the following key points for Section 1 (7 points)

- Explain why we might want to use a probability different than 0.5 to decide how we classify observations.
- Explain how the choice of the threshold here might change the error rate.
- Explain what the True Positive Rate and False Positive Rate is, for a given threshold.
- Explain what the ROC curve captures.
- What is the AUC of an ROC curve? Use some example data to compute this for a test set.

- You should use two methods from the class and compare their performance on a classification problem with some example data using the metrics available for ROC curves.
- Show some code in your report illustrating how to use this method for a classification problem (similar to what is included in the lecture notes).

Your report must address the following key points for Section 2 (8 points)

- Provide an overview of how the SVM classifier works. What is it trying to optimize?
- How does it differ from other classifiers we have seen in this class?
- What are advantages/disadvantages of this method?
- How do you implement this method in `R`? What parameters need to be tuned or selected? What is the test error of this method for the standard example we have used for classification in class?
  - For this part you **should show sample code which could (hypothetically) be run** in your report. Look at the below link about code chunks for help with formatting this. This code must also be included in the `rmd` file. That is, you do not need the code to actually run in the pdf, but if we copied and pasted the code shown in your pdf it would run (given packages and the data are loaded, etc).

Submission Instructions

- Like Project 1, the pdf you submit and the rmd file **do not need to be identical**. However, the plots and any other output you include in your pdf file (eg, numbers in a table, etc) must be **created exactly** by code in the rmd file you submit. We will check rmd files at random and if we cannot recreate the plots and other output shown in your report you will **receive 0 for the project**.
- If you wish to include equations in your report, you may find that saving the html output to a pdf can make these equations look strange. Similarly, saving the html version as a pdf can often be quite long (longer than the suggested page limit). An alternative is to use the option to `knit` your `rmd` file to a word document (if you have word installed) or to a pdf (which requires a TeX installation). Some further details are given here and here.
- Note if you do not wish to show some code in your report but keep the code in your `rmd` file (so that they match), you can use `chunk options`, such as to hide some of your code. Some details about these options are given here. However, you must include code in your pdf providing an example of how to implement the methods, as described above (similar to what is included in the lecture notes).

- You are free to ask the TA and I questions about formatting but we will not be able to comment if your explanation of either method is "good". Try to compare it to the level of detail providing in the class. If you think someone else in the class could read it and have a similar understanding to what we obtained for LDA (for example), you are on the right track.