CSC 201 Project 1 – Classification

Intro

Class, this is your long-awaited opportunity to show your skills relating to Data Science. For this project, you will choose a fairly large data set which either has labels or of which labels may be easily derived, clean it, explore it, and fit a few machine learning models to it in order to derive some insight about it. Let's get into the details for how you should approach the project and what you should turn in.

Due Date

The suggested due date for this project is Monday, April 6, 2020. You will have until the middle of dead week to get all your assignments in, but it would be very stressful for you to wait to do all of it until a week before everything is due (speaking from experience).

Choosing a Data Set

I've sent you guys some suggestions for good open-source data libraries. Use that list or just Google until you find a data set that is interesting to you that has at least 10,000 data points ("data points" = rows * columns - empty cells). It will be easier for you if the data is already labeled. In other words, if you found a data set describing different observations of malignant and benign tumors of the brain, it would be best if there was some sort of column with categorical labels that you could use to train a model and validate the model's predictions on the test set, such as a column labeling each row "malignant" or "benign". I would encourage you all to send me your data sets once you've made your decision to make sure that it will work for your project before you go much further with your project.

The Analysis

Once you've chosen the data you want to work with, you should start working with it in R. Try to create some visualizations using the labs you've turned in and examples you find online. If you're data has empty rows, decide how to handle them. If your data has columns that should be numeric but are being processed as character, you should change that.

Your visualizations should be insightful and contribute to the narrative of your analysis. **Data science reports are persuasive essays that are just heavy on the proof.** If your data set was about the 2016 presidential election, you wouldn't include information about astrological signs of voters if it didn't bare any correlation to who they voted for (if it was correlated, that would be very interesting and you should definitely include that). Note: a good place to start is always to show the descriptive statistics of the numerical columns, and the distributions (think bar charts) of the categorical columns.

Once you have sufficiently explored your data to the point that you and your reader (me) can wrap your heads around what basic trends and patterns are in the data, you start modeling. For this project, there is no hard number of models you should use, but the point of the project is to compare the performance of different models on the data set that you choose, so that would be hard to do if you only had one or two models.

Of the models we have covered in class, the following should be considered when performing a classification analysis:

- Logisitic
- Support Vector Machine
- Random Forest
- Naïve Bayes
- K-means

The Report

The deliverable for this project is an R Markdown report. So you should include all the code from your analysis, the visualizations, and your written comments. This is not a recipe, but in general, a data science analysis report is laid out as follows:

I. Introduction
   a. You introduce your project – talk about the data you chose and the environment it describes (if your data set if about life expectancy of smokers versus non-smokers, you would mention some fact about the societal impact of increasing life-expectancy).
   b. Talk about what you discovered in your exploration and modeling of the data.

II. The Data
   a. Give validation to the source of your data
   b. Describe what it is about and how many observations there are

III. The Exploration
   a. Categorically describe the visualizations you create with your data
   b. Consider including a table of summary statistics of your continuous variables
   c. Consider trying to visualize the decision boundary your models will need to fit

IV. The Modeling
   a. Split the data into testing and training
   b. Train your models
   c. Predict with your models on the test set.

V. The Results
   a. Compare each model's results with the actual labeled results from the test set and report on their respective accuracies.
   b. Discuss how the models ranked in terms of predictive accuracy. Attempt to explain why some models performed better than others

VI. Conclusion
   a. This part is crucial to the persuasive aspect of a data science analysis. See the example I provided for inspiration if you would like. Use the results of your modeling to bolster (or direct) the point you are trying to make about the underlying patterns in the data you have chosen.

The only thing you need to turn in to me is the R Markdown (.rmd) file and the output from knitting that file (preferably .html, but you have other options with RStudio).

Please email or call me (318-278-3442) with questions about the projects. Otherwise, we will address FAQ's in the upcoming classes.