Student name: Alistair Pattison

Term student will be working on it: Winter 2024

Topic: Support Vector Machine and Neural Network classification methods

Description:

Your independent comps project is to learn about Support Vector Machine (SVM) and Neural Network classification methods. We would also like you to learn how Shapley values are used to enhance the explainability of these models. Depending on your previous work done in this area, you may also need to learn about validation methods in ML.

We would like you to use An Introduction to Statistical Learning with Applications in R (ISLR), Second Edition, as your primary reference for learning about these methods. Chapter 9 of ISLR covers Support Vector Machines (you can focus on models for two classes) and Chapter 10 covers Neural Networks. How deep you go into chapter 10 is up to you, and may be a function of the data you chose (see below), but we hope that you cover through CNN (section 10.3) and read about fitting neural networks (section 10.7). To understand Shapley values, we suggest An introduction to explainable AI with Shapley values and sections 9.5 and 9.6 in Interpretable machine learning. Validation method resources include section 5.1 in ISLR and A machine learning framework for sport result prediction. You are free to use any other textbook or online sources to better understand these ideas.

After gaining a basic understanding of these predictive methods, we'd like you to apply them to a dataset of your choosing. We expect your analysis of this dataset to be novel, meaning it hasn't been analyzed in any of your textbook or online references. With your interest in sports analytics, areas of interest could be score/win prediction or injury prediction in a professional sport of your choice. Here are just a few examples of these types of applications and data sources, though feel free to use any reputable data source for this project.

- Pappalardo, L., Rossi, A., Natilli, M., & Cintia, P. (2021). Explaining the difference between men's and women's football.
- Sara Hedar (2020). Applying Machine Learning Methods to Predict the Outcome of Shots in Football
- Pappalardo, L., Cintia, P., Rossi, A., et al. (2020). Explainable Injury Forecasting in Soccer via Multivariate Time Series and Convolutional Neural Networks.
- NFL Big Bata Bowl 2024, Link to data
- Major League Soccer Dataset, Link to data

In your analysis, we expect you to

- Use basic EDA and visualizations to gain an understanding of the basic features of your data.
- Develop standardized ML models using an appropriate Neural Network and SVM for any two class label prediction based on your choice of data/models.
- Conduct a comparative analysis of your chosen models in terms of accuracy and precision for the chosen outcomes.
- Enhance the explainability of these models, particularly using Shapley values, to provide insights into the feature importance and decision-making process.
- Establish a standardized pipeline for data preprocessing, model training, validation, and testing to ensure reproducibility and comparability across studies.

We prefer that you use R for the majority of your statistical analysis since this is an important computational tool for many statisticians, but if you feel there is a compelling reason to use another software program please first consult with your committee members.

The products of your comps project are a 25-minute talk and a paper. Your talk should give an overview of your predictive models and should address the key results of your analysis with appropriate visualizations. The audience is your fellow statistics majors, and you can assume that they have taken Statistical Inference and Regression. You will not be able to fit everything that you learned into your talk, so you should focus on conveying the main ideas. Your paper should then present the talk's material along with additional details described above. Your work (including images, tables, etc.) should be fully reproducible and you should submit a supplemental file containing your R code along with the paper (this can be your raw .Rmd file if you write the paper in R Markdown).

Your paper should be written in the style of a journal article and contain the following sections:

- Abstract: a short (<250 word) summary of your project. Be sure that your abstract is completely self contained.
- Introduction: introduce the topic, both the methods and the data set, that you will explore. Summarize the relevant background information and tell the reader why the work is important/interesting.
- Data: give an overview of the data set you will explore in your application.
- Methods: outline the details of the methods here. Don't focus on R implementation, but rather the underlying details. A randomly selected statistics major should be able to understand the method after reading this section.
- Results: apply the method to the data set here. Be sure to give the necessary implementation details and clearly interpret the results.

- Conclusion: briefly summarize your work. In this section you can also outline areas for future research and/or discuss any limitations of your current application.
- R Appendix: please attach a knitted Markdown doc to document your complete analysis. This appendix doesn't count towards your page count. It should be readable by a randomly selected statistics major who wants to replicate your results.

Your paper should not contain R code or raw output. You will be provided with examples of journal articles on the Independent Comps Moodle page.