Practical Machine Learning

Exercise 3
Spring 2024

Lisa Stafford

February 20, 2024

Abstract

Model selection is a common problem for machine learning architects where one selects the model that does the best job of generalizing and most appropriate for the given problem. Those selecting the model should choose the best balance between under- and over-fitting. When selecting a model, one must examine the value of different predictive methods to identify the one that best fits observable data. There are two main factors when choosing the most appropriate model in machine learning. One is the reason for choosing it, and the other is the model's performance. Reasons for choosing a model should be based on the available data set, and the problem task.

Introduction

In order to learn and determine which models are most effective, we used a wine data set obtained from the University of Irvine [**?**]. We use a consistent machine learning model for each dataset (in this case using the scikit-learn decision tree library) and then determining what (if any) preprocessing methodologies are most effective in preparing our two datasets for machine learning by evaluating the overall performance of the models after taking these datasets in standard and consistent ways, and modifying them using different preprocessing techniques, then evaluating performance of the model per each preprocessing step/activity.

Dataset Description

Experimental Setup

Results

References

[1] A. Asuncion, D. Newman, UCI Machine Learning Repository, University of California, Irvine (2007). Obtained from https://archive-beta.ics.uci.edu/dataset/186/wine+quality.

[2] C. Harris, K. Millman, S. van der Walt, Array programming with NumPy. Nature 585, 357–362 (2020). DOI: 10.1038/s41586-020-2649-2. https://numpy.org/doc/stable/reference/generated/numpy.isnan.html

[3] M. Waskom, (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021, https://doi.org/10.21105/joss.03021.

[4] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.