Developing Documentation in SciKit-Learn and Applying Statistical Models

Phil Cork, Julia Piscionere, Dylan Waks, and Allison Hansen

Department of Data Science, Computer Science, College of Charleston

SciKit-Learn is an open-source software that enables predictive data analysis. Built in Python on the foundations of NumPy, SciPy, and matplotlib, SciKit-Learn makes complex statistical models easily applicable to datasets in a range of contexts. SciKit-Learn also handles preprocessing, data reduction, and model selection such as classification, regression, and clustering. Documentation is vital to users' ability to capitalize on the capabilities of the software. Due to the nature of open-source software having different algorithms contributed by various members of the community, there are some inconsistent practices for documentation throughout. Here, we undertook pieces of a larger issue to standardize documentation usually relating to missing parameters within a method of a statistical model. The statistical models we studied include Kernel Density, Mini Batch KMeans, CCA, PLS SVD, Incremental PCA, and Lasso. This process required both theoretical knowledge of the statistical model and a practical understanding of the SciKit-Learn codebase, so we have included examples of some of the statistical models invoked using sample data sets while describing our experience contributing to the documentation.