HDS-5230-07

Prof. Adam Doyle

Week 8 Assignment

Mar 30th, 2025

Module/framework/package	Name and a brief description of the algorithm	An example of a situation where using the provided GLM implementation provides superior performance compared to that of base R or its equivalent in Python (identify the equivalent in Python)
glm function within Base R's stats package	glm in base R is an extension of tradition linear models, supporting various functions for model's relationship between target and predictors such as binary outcomes and count data, does not necessarily follow normal distribution.	R's glm may outperforms python's statsmodels library for its ease of use and simplicity, built-in tests like Wald and ANOVA, easy implementation on visualization, and model comparison.
CRAN Task View for high-performance and parallel computing with R	Rather than a single algorithm, it is a collection of tools and frameworks for high performance computing with big datasets	There is no exact equivalent to CRAN in Python, meaning it has consistency and centralization of the resources for high performance computing, ensuring great compatibility among packages. So better choice over Python without worrying for version compatibility among HPC tools.
Dask ML, scaling well out of large datasets on a single machine or distributed cluster	GLM in Dask-ML includes models such as Logistic/Poisson/Linear Regression, supporting various algorithms like adm, gradient descent, lbfgs, newton, and proximal grad.	Specifically designed for distributed computing, it can scale across multiple machines seamlessly with great integration with DASK ecosystem. It is also capable of working directly with DASK arrays and data frames efficiently.
Spark R, providing an intergace for using Apache Spark with R	GLMs implemented in SparkR, allowing various regression models to stored data in SparkDataFrame, efficiently handling large datasets.	Since SparkR and PySpark provide quite similar functions regarding ML with Spark, the choice between these may focus on user's preferred programming language.
Apache Spark MLlib, enabling optimization tasks to scale efficiently across a cluster	Various algorithmns including Gradient Descent, Stochastic Gradient Descent, and Limited -memory BFGS.	With RDD-based optimization, Spark's MLlib can distribute the computation of gradients and updates across a cluster, enabling parallelization and distributed computation for large datasets.
Scikit-learn Python, providing multiple regression models and techniques for machine learning	Scikit-learn provides variety of techniques such as OLS, Bayesian Ridge Regression, Lasso, Ridge, ElasticNet, and GLM, making it one of the most widely used tool for machine learning.	Scikit-learn is one of the most widely used tool for machine learning, due to the fact it supports variety of ml techniques from simple regression models to ensemble models, and more. It also provides wide range of modules for tuning and data preprocessing and highly compatible with basic Python libraries like pandas and numpy.