WEEK-08

GLM Investigation

Sharath Kasula

High Performance Computing

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)**
A. Base R (stats)	Use the glm.fit function from IRLS for estimating parameters from exponential family distributions.	The statistical program that best works with small to medium-sized datasets would be A which can be used for hospital infection rate analysis. Python equivalent: statsmodels.
B. Big Data Version of R	Incremental or Distributed IRLS methods (biglm and h2o) process distributed and chunked data sets with regularization techniques.	The system works with extremely big datasets spanning throughout multiple states or healthcare records better than the base R system or statsmodels platform.
C. Dask ML	Scalable solvers: ADMM, L-BFGS, Proximal Gradient, Newton's Method. Works with Dask arrays/dataframes.	The system serves as an optimal solution to train logistic regression algorithms when working with large distributed dataset structures. Outperforms scikit-learn on memory-limited systems.

D. Spark R	The IRLS model scales beyond single machines by operating on SparkDataFrames and supports the estimation of GLM families including the Poisson distribution alongside Tweedie distribution as well as optional regularization options.	Excellent for modeling insurance claims or lab data at scale. Base R lacks the capability to process distributed computation operations. Python equivalent: PySpark.
E. Spark optimization	The optimization process relies on Stochastic Gradient Descent (SGD) to apply sub-gradients for loss function optimization. Suited for large-scale streaming data.	The most suitable use of scikit-learn is to build adaptive models for streaming vital measurements and ICU monitoring. The system provides superior performance to static fitting which can be found in base R and scikit-learn.
F. Scikit-learn	The tool supports multiple optimization methods which use both L-BFGS and SAGA along with Newton-CG and Coordinate Descent for regularization and cross-validation processes.	The system demonstrates solid capabilities when used for predicting readmission among high-dimensional logistic regression models. Integrates cleanly into pipelines.