Week 08 - GLM Investigation

Overview

The extension of ordinary linear regression known as Generalized Linear Model (GLM) accepts non-normal distributed response variables. The selection of different optimization frameworks and libraries impacts the improvement of GLMs with diverse algorithms that determine both efficiency and scalability and implementation suitability in practical datasets.

This report provides an overview of optimization methods from six different frameworks/modules alongside their performance values and healthcare applicability evaluation.

Generalized Linear Model (GLM) Optimization Across Frameworks

Generalized Linear Models enable linear models to apply non-Gaussian family distribution types (such as Poisson, Binomial, Gaussian and more) to model their response variables. Parameter estimation in GLMs uses numerical methods for efficient optimization tasks that become essential when working with extensive datasets.

Three optimization techniques IRLS and SGD and L-BFGS and Coordinate Descent serve to estimate model parameters in diverse frameworks for GLM implementations. Different healthcare applications utilize the listed optimization techniques from multiple frameworks as shown in this summary table.

Comparison of GLM Implementations

Module/Framework/Package	Algorithm Description	Healthcare Example

Base R (stats library)	**IRLS** method does parameter estimation through successive weight refinements which follows the deviance function. The method suits datasets of small to medium size as well as it comes with default functionality for multiple link functions.	This model predicts readmission rates of hospital patients through its analysis of patient demographics and medical history and past admission records. Hospital resource allocation becomes more effective while reducing patient overcrowding because of this system.
Big Data R (High-Performance Computing libraries such as bigglm from biglm)	The method operates by applying stochastic approximation algorithms to handle big datasets through partial data processing rather than entire memory loading. The approach circumvents memory restrictions yet functions with high efficiency in terms of calculations.	Genetic risk factor analysis of diseases by utilizing big genomic data sets which demand more memory than available systems can handle. By studying genomic data scientists can detect natural genetic risks that contribute to cancer as well as diabetes development.
Dask-ML	The system applies stochastic gradient descent (SGD) alongside Newton's Method which enables distributed out-of-core computation for dealing with big datasets in an efficient manner.	The system facilitates immediate medical imaging data analytics on X-ray and MRI scans that need large memory-independent processing capabilities. The purpose of this approach enables medical institutions to accelerate disease diagnosis processes.
Spark R	The system employs L-BFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) and IRLS to execute efficient distributed GLM coefficient optimization in distributed	The system operates effectively to track continuous patient vital signs by uniting information from several internet-connected medical devices simultaneously. Healthcare

	computing settings. L-BFGS delivers optimal performance in data situations with many dimensions.	providers gain continuous patient health tracking capabilities which enables them to provide rapid responses during critical moments.
Spark Optimization (MLlib)	The system uses Gradient Descent together with L-BFGS methods for optimization which allows expanding its capabilities to distributed computing clusters. MLlib exists as a system that processes extensive datasets with operational effectiveness.	The system tracks disease outbreaks by analyzing time-sensitive information received from various hospitals in multiple regions to forecast upcoming outbreaks. Public health agencies use data-driven decisions that become possible through this system.
Scikit-learn	The implementation of coordinate descent (for Lasso regression) together with SGD ensures high efficiency for large datasets within structured data analysis and feature selection.	The system proves beneficial for timely diabetes diagnosis through patient healthcare information examination. Tests such as glucose level checks and blood pressure tests along with body mass index evaluations enable predictive models to determine at-risk patients before their conditions worsen.

Advantages of Each Framework in Healthcare Applications

A GLM implementation performs best based on the dataset size along with available computational power while using a specific optimization method. The following list summarizes important benefits for each system:

Base R (stats) functions as the perfect tool for healthcare research that requires small-scale data exploration.

Statistical functions available inside the platform make it accessible with user-friendly features. Additionally working with bigglm in the

Big Data R framework represents the best choice for researchers handling medical datasets that need extensive memory along with genomic analysis or hospital billing records.

Dask-ML works best with parallelized machine learning tasks at large scale detected in radiology and pathology imaging. The real-time patient data stream monitoring system Spark R provides high efficiency while processing large volumes of healthcare data.

Spark MLlib: Best for big data analytics in healthcare, particularly useful for epidemiology and population health studies. Scikit-learn gives healthcare practitioners effective tools to select powerful features which boost their predictive modeling capabilities for disease risk evaluation and patient diagnostic practices.

Conclusion

The selection of GLM framework bases on both the targeted healthcare application and the dataset size. Two suitable options for working with small datasets include tools like Base R and Scikit-learn whereas Spark and Dask-ML offer scalable processing solutions for real-time operations using large datasets. Healthcare frameworks serve a critical purpose for analytics progress in the medical field by enabling the ability to forecast disease outbreaks and direct hospital resources more effectively.

Citations & References:

- R Documentation: https://www.rdocumentation.org/packages/stats/
- High-Performance Computing in R: https://cran.r-project.org/web/views/HighPerformanceComputing.html

- Dask ML Documentation: https://ml.dask.org/glm.html
- Spark R API: https://spark.apache.org/docs/3.5.0/api/R/reference/spark.glm.html
- Apache Spark MLlib Optimization: https://github.com/apache/spark/blob/master/docs/mllib-optimization.md
- Scikit-learn Linear Models: https://scikit-learn.org/stable/modules/linear_model.html