# HDS 5230: Week 10 Application Assignment's Instructions

**Name: John Reddy Peasari**

After going through the documentation, please summarize your findings via a table whose structure matches the one shown below:

| **Module/framework/package** | **Name and a brief description of the algorithm** | **Example Description** |
|---|---|---|
| Stats | glm()<br>GLM is in R is a class of regression models that supports non-normal distributions and can be implemented in R using glm() function. This model works well with a variable which depicts a non-constant variance with three important components namely random, systematic, and a link component making the GLM model. | In glm() function family types includes binomial, poisson, gaussian, gamma, quasi. Each distribution performs a different usage and can be used in wither classification and prediction. |
| Big Data Version of R H20 | H2O.GLM Generalized Linear Models<br>It is used to estimate regression models for outcomes following exponential distributions. It includes various regression implementations such as Gaussian, Poisson, Binomial, and Gamma. GLM fits models based on the maximum likelihood estimation via iteratively reweighted least squares.<br>The elastic net penalty can be used for parameter regularization. | H2O can process large datasets because it relies on parallel processes. GLM here can be used for all types of regressions. In GLM data are split by rows but not by columns. Here, the model fitting computation is distributed, extremely fast, and scales extremely well for models with a limited number of predictors with non-zero coefficients (near low thousands). H2O returns the optimal amount of regularization for the given problem. |
| Python Dask | dask_glm()<br>Optimization algorithm for solving minimization problems. Implements distributed generalized linear model family for regularized and unregularized problems. This has convex optimization algorithms for Ibfgs, gradient descent, newton, ADMM, proximal gradient. All the algorithms for regularized problems in dask-glm use the framework of proximal operators. | Generalized linear models built for parallel and distributed machine learning. Dask-glm tries to solve for large scale learning challenges within SciPy ecosystem. Generalized linear model implementations scale well towards larger datasets either using a single CPU or distributed cluster. Sklearn uses single core whereas dask-glm uses full core machine |
| | | |

| | | |
|---|---|---|
| SparkR | spark-glm()<br>Sparks’s generalized linear regression interface allows for specifications of GLMs which can be used for various types of prediction problems including linear regression, poisson regression, logistic regression, and others. A GLM is specified by a distribution of the response and a link function which in turn minimizes the sum of log-likelihoods. Spark.glm is a simple swapper over an ML pipeline that consists of RFormula for preprocessing and encoding and an estimator (GLR) | Spark.glm fits generalized linear model against a spark data frame similar R’s glm() function. Spark only supports up to 4096 features through its generalized linear regression interface |
| L-BFGS | The L-BFGS method approximates the objective function locally as a quadratic without evaluating the second partial derivative of the objective function. L-BFGS is used as a solver for linear regression, logistic regression, multilayer perceptron classifier. Spark MLlib library implements iteratively reweighted least squares (IRLS). It can be used to find the maximum likelihood estimates of a generalized linear model, find M-estimator in robust regression and other optimization problems. | It solves a few optimization problems iteratively by linearizing objective at current solution, solve a weighted least square and repeat above steps until convergence. It also requires the number of features to be no more than 4096. Currently IRLS is used as a default solver of generalized linear regression |
| Scikit-learn | Generalized Linear Regression<br>GLM extend linear models in two important ways. Predicted Y values are linked to a linear combination of the input X variables via an inverse link function h. Then, squared loss function is relaced. Minimization of the problem will be achieved using L2 regularization. | It can be implemented in weather modelling, risk modelling, predictive maintenance. The choice of distribution depends on the type of target values y. |