Comparison of Mathematical and Statistical Functionality of Machine Learning Tools for Data Analysis Research

Shamitha S. K
Research Scholar, VTU, Belgaum
Bangalore, India

Nikitha Pai
Guest Lecturer,
NMKRV College for Women,
Jayanagar, Bangalore, India

B. Nithya
Asst.Professor,
New Horizon College of Engineering,
Marathahalli

Dr. V. Ilango
Professor,
New Horizon College of Engg
Marathahalli

***Abstract* - Over the last three decades many general-purpose machine learning frameworks and libraries have emerged from both academia and industry. The aim of this paper is to compare mathematical and statistical programming languages on a fair level by showing only facts about the tested programs and attempts have been made to avoid subjective remarks. This could be used as base information to make own decision. The paper takes a closer look on mathematical and statistical programming, data analysis and simulation functionality for huge and very huge data sets. The following machine learning tools have been tested: Mathematica from Wolfram Research Inc., MATHLAB from The Mathworks Inc. This type of functionality is of great interest for econometrics, the financial sector in general, biology, chemistry, physics and have immense usage in other areas as well, where the numerical analysis of data is very important.**

Keywords: Machine Learning frameworks, Mathematica, Mat lab

I. INTRODUCTION

Given the enormous growth of collected and available data in companies, industry and science, techniques for analyzing such data are becoming ever more important. [1] Research in machine learning (ML) combines classical questions of computer science (efficient algorithms, software systems, databases) with elements from artificial intelligence and statistics up to user oriented issues (visualization, interactive mining, user assistance and smart recommendations).[2] Over the last three decades, many general purpose machine learning frameworks, as well as special purpose machine learning libraries, such as for phishing detection or speech processing [1], has emerged from both academia and industry. In this survey, we will only consider the general purpose frameworks. It is good to have a look around to see what languages and platforms are popular in self-selected communities of data analysis and machine learning professionals.[3] The study consists of tables which lists the availability of functions for each program [6][7]. It is divided in functional sections of mathematical, graphical functionality and programming environment, a data import/export interface section, the availability for several operating systems, a speed comparison and finally a summary of the whole information. To rate all these information, a simple scoring system has been used and following machine learning tools have been tested: Mathematica from Wolfram Research Inc, MATHLAB from The Mathworks Inc. A recent poll is titled “What programming/statistics languages you used for an analytics work in 2013“ [6]. The results suggest heavy use of R and Python and SQL for data access.

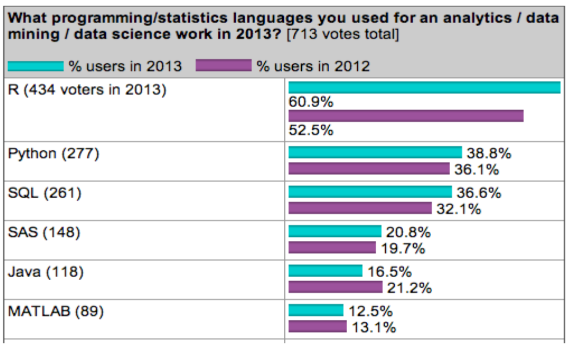

FIG 1: Popular Data Analytic Tools

II. COMPARISON OF THE MATHEMATICAL FUNCTIONALITY

Actually there are a lot of different mathematical and statistical programs in the market which are covering a huge amount of functions. The above figure(FIG 1) discusses the popular Analytical tools. The following tables should give an overview about the functionality for analyzing data in numerical ways and should mark out which functions are supported by which program and whether these functions are already implemented in the base program or whether you need an additional module. The functions are sorted by the categories: Standard mathematics, Linear algebra, Numerical mathematics, Stochastic, Statistics, Other mathematics.

A. Standard Mathematics

Standard mathematics functions are an essential part of any kind of mathematical work. Not necessary to mention that these type of functions should be available in all programs. Therefore the following results are not very surprising.[6].Comparison of Mathematica and MATHLAB on standard mathematics are discussed in the table below

Table1.1: Comparison of Mathematica and MATHLAB with respect to standard mathematics

Functions (Version)	**Mathematica** (6.0)	**MATHLAB** **(2008a)**
BesselI	✓	✓
Bessel J	✓	✓
BesselK	✓	✓
Bessel Y	✓	✓
Beta function	✓	✓
Binomial	✓	✓
Factorial	✓	✓
FresnelC	✓	✓
FresnelS	✓	✓
Gamma function	✓	✓
Hyperbolic trig. Function	✓	✓
Incomplete Gammafunc.	✓	✓
Log / Ln / Exp	✓ / ✓ / ✓	✓ / ✓ / ✓
Log-Gammafunc.	✓	✓
Poly gamma	✓	✓
Square root	✓	✓
Sum / Product	✓ / ✓	✓ / ✓
Trig. / arg trig. Functions	✓ / ✓	✓ / ✓

B. Algebra

Algebra and especially linear algebra offers a basic functionality for any kind of matrix oriented work. i.e. Optimization routines are widely used in the financial sector but also very useful for logistic problems (remember the traveling salesman problem).Most simulation and analyzing routines are relying on decomposition equations solving and other routines from algebra.Table 1.2 discusses the comparison between Mathematica and MATHLAB with respect to Algebra

Table 1.2: Comparison of Mathematica and MATHLAB tools with respect to Algebra

Functions (Version)	**Mathematica** **(6.0)**	**MATHLAB** **(2008a)**
Eigenvalues		
Eigenvalues	✓	✓
Eigenvectors	✓	✓
Matrix		
Characteristic polynom	✓	✓
Determinant	✓	✓
Hadamard matrix	M	✓
Hankel matrix	✓	✓
Hilbert matrix	✓	✓
Householder matrix	-	✓
Inverse matrix	✓	✓
Kronecker product	✓	✓
Pascal matrix	-	✓
Toeplitz matrix	✓	✓
Upper Hessenberg form	✓	✓
Decomposition		
Cholesky decomposition	✓	✓
Crout decomposition	-	✓
Dulmage-Mendelsohn decomposition	-	✓
LU decomposition	✓	✓
QR decomposition	✓	✓
Schur form of quadratic matrix	✓	✓
Smith normal form	-	✓

Singular value decomposition	✓	✓
Optimization		
Optimization - linear models (Unconstr. / Constr.)	✓ / ✓	✓ / ✓
Optimization - nonlinear models (Unconstr. / Constr.)	✓ / ✓	✓ / ✓
Optimization - quadratic models (QP) (Unconstr. / Constr.)	✓ / ✓	✓ / ✓
Equation solver		
Linear equation solver	✓	✓
Non-linear equation solver	✓	✓
Ordinary Differential Equation solver	✓	✓
Partial Differential Equation solver	✓	✓
Miscellaneous		
Moore-Penrose pseudo-inverse	✓	✓
Sparse matrices handling	✓	✓

Most simulation and analyzing routines are relying on decompositions, equation solving and other routines from algebra.

C. Numerical Mathematics

Numerical mathematics offers fundamental algorithms for several appliances. It is marked out that especially any kind of interpolation algorithms are commonly used in technical and non-technical businesses. Without really recognizing, interpolation routines are used in nearly any kind of graphical representation. .[6][4]Table 1.3 explains the comparisons of Mathematica and MathLab tools with respect to Numerical functions.

Table 1.3: Comparison of Mathematica and MATHLAB tools with respect to Numerical mathematics

Function (Version)	Mathematica (6.0)	MATHLAB (2008)
Interpolation		
B-Spline interpolation	✓	✓
Classical interpolation (1D/2D/3D/n D)	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
k-Spline interpolation	✓	✓
Pade interpolation	✓	-
Piecewise cubic hermite polynomial interpolation	✓	✓
Piecewise polynomial interpolation	✓	✓
Other functions		
Bisection	✓	✓
Newton method for finding roots	✓	✓
Runge Kutta method for solving ODE	✓	✓

III. COMPARISON OF THE STATISTICAL FUNCTIONALITY

A. Descriptive statistic and Distribution functions

Very important to get familiar with data and to understand samples of data are stochastic and descriptive statistic routines. Distribution functions, their CDF- Cumulative Distribution Function and PDF- Probability Distribution Functions function are commonly used to figure out what are representative samples and what are outliers. A typical “simple” but common usage might be in a productive area to take samples of the manufactured product and to see whether the faulty parts in a sample are within a normal range. More complex usages might be in load balancing simulations of telecommunication hardware. However it might be possible to mention example usages for nearly all kind of business.[6].Comparison of Mathematica and MathLab tools to distribution function is explained in Table 2.1

Table 2.1 Comparison of Mathematica and MATHLAB tools with respect to distribution function

Functions (Version)	Mathematica (6.0)	MATHLAB (2008a)
General Function		
Contingency tables	-	-
Correlation	✓	✓
Cross tabulation	✓	✓
Deviation	✓	✓
Kurtosis	✓	✓
Markov models	✓	✓
Mean /geometric Mean / Mode	✓ / ✓ / ✓	✓ / ✓ /-

Min / Max	✓ / ✓	✓ / ✓
Quantile / Percentile	✓ / ✓	✓ / ✓
Skewness	✓	✓
Variance	✓	✓
Variance-covariance matrix	✓	✓
Distribution Functions *(PDF / CDF / iCDF/ random number)*		
Bernoulli	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Beta	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Binomial	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Brownian motion	-/-/-/ ✓	-/-/-/-
Cauchy	✓ / ✓ / ✓ / ✓	-/-/-/-
Chi-squared	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Chi-squared (non-central)	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Dirichlet	✓ / ✓ / ✓ / ✓	-/-/-/-
Erlang	✓ / ✓ / ✓ / ✓	-/-/-/-
Exponential	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Extreme value	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
F	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
F (non-central)	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Gamma	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Geometric	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Gumbel	✓ / ✓ / ✓ / ✓	-/-/-/-
Half-normal	✓ / ✓ / ✓ / ✓	-/-/-/-
Hotelling T2	✓ / ✓ / ✓ / ✓	-/-/-/-
Hyper-exponential	✓ / ✓ / ✓ / ✓	-/-/-/-
Hypergeometric	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Kernel	✓ / ✓ / ✓ / ✓	-/-/-/-
Laplace	✓ / ✓ / ✓ / ✓	-/-/-/-
Logarithmic	✓ / ✓ / ✓ / ✓	-/-/-/-
Logistic	✓ / ✓ / ✓ / ✓	-/-/-/-
Log-normal	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Log-normal (multivariate)	✓ / ✓ / ✓ / ✓	-/-/-/-
Negative binomial	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Normal	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Normal (bivariate)	-/-/-/-	-/-/-/-
Normal (multivariate)	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Pareto	✓ / ✓ / ✓ / ✓	-/-/-/-
Poisson	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Rayleigh	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
S	✓ / ✓ / ✓ / ✓	-/-/-/-
Student's t	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Student's t (non-central)	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Student's t (multivariate)	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Uniform	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Von Mises	✓ / ✓ / ✓ / ✓	-/-/-/-
Weibull	✓ / ✓ / ✓ / ✓	✓ / ✓ / ✓ / ✓
Wishart	✓ / ✓ / ✓ / ✓	-/-/-/ ✓

A. *Statistics*

Statistical functions are fundamental for any kind of data analysis. Routines like regression or time series are commonly used to find out trends or to predict future values i.e. for stock market courses. Filter routines are used to smooth or filter effects in data acquisition. Multivariate statistics are used to find patterns or common characteristics in data i.e. for market basket analysis by clustering routines.[9].Comparison of Mathematica and MAthLab tools with respect to statistical functions are explained in the table 2.2.

Table 2.2 Comparison of Mathematica and MATHLAB tools with respect to statistical functions

Functions (Version)	**Mathematica (6.0)**	**MATHLAB (2008a)**
Regression models		
Linear	✓	✓
Loess	✓	✓

Logistic Regression	✓	✓
LOGIT / PROBIT	m- / m	✓ / ✓
Nonlinear / Polynomial	✓ / ✓	✓ / ✓
PSN	-	-
Tobit models	-	-
Test statistics		
Ansari-Bradley test	-	✓
Bartlett multiple-sample test	-	✓
Besley test	-	-
Breusch-Pagan test for homoscedasticity	-	-
Chow Test for stability	-	-
CUSUM test for stability	-	-
Davidson-MacKinnon J-Test	-	-
Dickey Fuller test	-	-
Durbin-Watson test	✓	✓
Engle's LM test	-	✓
Friedman's test	-	✓
F-Test	✓	✓
Goodness of fit test	-	✓
Goldfeld-Quandt test for homoscedasticity	-	-
Granger's causality test	-	-
Hausman's specification test	-	-
Kolmogorov-Smirnov test	-	✓
Kruskal-Wallis test	-	✓
Kuh test	-	-
Lagrange multiplier test	-	-
Lilliefors test	-	✓
Ljung-Box Q-Test	-	✓
Mann-Whitney U test	-	-
Sign test	-	✓
T-Test	✓	✓
Wald test	-	-
Walsh test	-	-
Wilcoxon rank sum / sign test	- / -	✓ / ✓
Z-Test	✓	✓
Filter / smoothing models		
Bandpass / Lowpass / Highpass / Multiband / Bandstop	✓ / ✓ / ✓ /-/ ✓	✓ / ✓ / ✓ / ✓ / ✓
Battle-Lemarie	✓	-
Bessel	✓	✓
Butterworth	✓	✓
Chebyshev	✓	✓
Coiflet	✓	-
Daubechies	✓	✓
Elliptic	✓	✓
Haar	✓	✓
Hodrick-Prescott	-	-
IIR / FIR	✓ / ✓	✓ / ✓
Kernel	-	✓
Linear	✓	✓
Meyer	✓	✓
Pollen	-	-
Riccati	✓	✓
Shannon	✓	-
Savitzky-Golay	-	✓
Time series models		
ARMA / ARIMA / ARFIMA / ARMAX	/ ✓ /-/-	✓ /-/ ✓ ✓
GARCH / ARCH / AGARCH / EGARCH / FIGARCH / IGARCH	/ ✓ /-/-/-/-/- /-	✓ //-/ ✓ / -/-/-/-/-

/ MGARCH / PGARCH / TGARCH models		
Holt's Winter additive / multiplicative	- / -	- / -
Multivariate GARCH models (Diagonal VEC / BEKK / Matrix Diagonal / Vector Diagonal)	-/-/-/-	-/-/-/-
Partial autocorrelation	✓	✓
Spectral analysis	✓	✓
State space models	✓	✓
Time series analysis (Stationary / Non-stat.)	✓ / ✓	✓ / ✓
Wavelets	✓	✓
Multivariate statistics		
ANOVA / MANOVA	✓ / -	✓ / ✓
Cluster analysis (hierarchical/k-means)	✓ / ✓	✓ / ✓
Discriminant analysis	-	✓
Factor analysis	-	✓
Fuzzy clustering	-	✓
Procrustes analysis	-	✓
Principal component analysis	-	✓
Principal coordinate analysis	-	✓
Survival analysis	-	-
Design of Experiments		
Box-Behnken design	-	✓
Central composite design	-	✓
D-Optimal design	-	✓
Full / Fractional factorial design	- / -	✓ / ✓
Hadamard design	-	✓
Response surface design	-	✓
Other statistical functions & models		
Bootstrapping	✓	✓
Duration models	-	-
Entropy models	-	✓
Event count models	-	-
Heckman two step estimation	-	-
Heteroscedasticity	-	-
Jacknife estimation	-	✓
Lagrange multiplier test	-	-
Markowitz efficient frontier	✓	✓
Maximum Likelihood (Unconstr. / Constr.)	✓ / ✓	✓ / -
Monte Carlo simulation	✓	✓

IV. OTHER MACHINE LEARNING TOOLS

Table 3: Machine Learning tools and Libraries

Name	HLD	OS	Language
Aleph	No	Win/Unix	Yap Prolog
C4.5/C5/See5	Yes	Win/Unix	C/C++
Encog	Yes	Win/Unix	Java/.NET
FuzzyML	Yes	Win/Unix	ADA
IBM Cognos	Yes	Web	PowerHouse
IBM SPSS Modeler	Yes	Win/Unix/OSX	Java
JavaML	Yes	Win/Unix	Java
JHepWork	Yes	Win/Unix	Java/Jython/Jruby/BeanShell
Joone	No	Win/Unix	Java
KNIME	Yes	Win/Unix	Java/Python/Perl
LIONsolver	Yes	Win/Unix	C/C++
MLC++	No	Win/Unix	C++

Mlpy	No	Win/Unix	Python
MS SQL Server	Yes	Win	.NET
Neuroph	No	Win/Unix	Java
Oracle Data Miner	Yes	Win/Unix	Java
Orange	No	Win/Unix	C++/Python
PCP	Yes	Win/Unix	C/C++/Fortran
Pyml	No	Win/Unix	Python
R	Yes	Win/Unix	C/Fortran/R
RapidMiner	Yes	Win/Unix/OSX	Java/Groovy
Salford Systems	Yes	Win	C/C++/.NET?
SAS Enterprise Miner	Yes	Win/Unix	C
scikit-learn	Yes	Win/Unix/OSX	C/C++/Python/ Cython
Shogun	Yes	Win/Unix	C/C++/Python/ R/MATHLAB
Statistica	Yes	Win	.NET/R

All the tools and libraries referred in table 3 are commercial, close-source products, while the others are licensed under various open-source licenses (GNU (L) GPL, Apache or MIT) with a strong preference towards GPL and LGPL. In terms of operating system (column OS), as most of them rely on virtual machines (Java, Python), they are running cross-platform (Windows, Unix, Mac OS X).[9][10]The few exceptions are large commercial applications developed for Windows operating system. The ability to handle large data sets (column HLD) is largely impacted by two factors: the programming language and environment used to develop the tool and the supported machine learning methods. [7] One can observe that most of the products originating in Python world, such as Mlpy, Pyml and YAPLF, have problems in handling large data sets, may be due to the lack of mature Python libraries for large data processing at the time tool development was started. [5] Machine learning methods also impact this criteria, some of them, such as neural networks, being not well suited candidates for large data sets handling. Programming language support and interfacing (column Language) is an important criterion when it comes to integrate a library in your own application.[8] All of the surveyed products are supporting at least one external interface, which usually are its native language / platform. Many of them offer support for additional programming languages as well. The most popular languages are Java [10], C/C++ [9] and Python [6], followed by .NET, FORTRAN, R etc.

CONCLUSION

Shortly after we started this survey, we have been overwhelmed by the large number of libraries, tools, projects addressing machine learning, showing huge interest in this topic among research teams in academia and industry, equally. Applying popular machine learning algorithms to large amounts of data raised new challenges for ML practitioners. Traditional ML libraries does not support well processing of huge data sets, so that new approaches are needed based on parallelization of time-consuming tasks using modern parallel computing frameworks, such as MPI, Map Reduce. A sequel survey will investigate machine learning solutions designed for distributed computing environments, such as grids or cloud computing.

Our future plan aims at building a smart platform for problem solving applied in the field of Machine Learning, which will be able to smartly support end- users in their activities by selecting the most appropriate method for a given data set, or tweaking algorithms parameters.

REFERENCES:

[1] Daniel Pop, Gabriel Iuhasz , Overview of Machine Learning Tools and Libraries.

[2] Saeed Abu-Nimeh1 , Dario Nappa2 , Xinlei Wang2 , and Suku Nair , A Comparison of Machine Learning Techniques for Phishing Detection.

[3] http://www.scientificweb.com/ (website).

[4] http://www.ml.cmu.edu/research/dap- papers/dap-guang-xiang.pdf (website).

[5] http://www.cs.nyu.edu/˜mohri/pub/hbkb.pdf (website).

[6] http://www.kdnuggets.com/polls/2013/languages-analytics-data-mining-data-science.html(website).

[7] Harshad Mehandale , Machine Learning On cloud Harnessing The Big Data Juggernaut (article) ,Business world.

[8] http://machinelearningmastery.com/best-programming-language-for-machine-learning/ (website).

[9] Threaling K, Some Thoughts on the Current State of Data Mining Software Applications http://www.thearling.com/text/dsstar/top10.htm, 1998.

[10] Zheng Zhu, Data Mining Survey, ver 1.1009, 2010, http://www.dcs.bbk.ac.uk/˜zheng/doc/datamining.pdf.