Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 2, No.1 : 2024
ISSN : **1906-9685**

LINEAR ALGEBRA APPLIED TO BIG DATA ANALYTICS FOR IMPROVING LEARNING STRATEGIES

Ms. V. KAVYA M.Sc., DCA.,
Assistant Professor, PG & Research Department of Mathematics, Marudhar Kesari Jain College for Women, Vaniyambadi, Tirupattur District, Tamil Nadu :: kavyavijayan38@gmail.com

ABSTRACT:

The sudden increase in the digital universe (Big Data) opened doors for new types of data analytics called big data analytics and new job opportunities. In 2012, only 23% of organizations had an enterprise worldwide Big Data strategy, whereas today 97.2% of organizations are investing in Big Data. A recent Harvard Business Review survey of senior Fortune 500 and federal agency business and technology leaders report that 70% of the respondents plan to hire data scientists. The U.S. Bureau of Labor Statistics (BLS), Occupationaloutlook Handbook 2018 projects that there will be a 34% increase in data analytics jobs from 2016 to 2026. A McKinsey Global Institute research report indicates that the demand for big data analytical talent could be very high and will produce 50 to 60 percent more data analytic jobs. Similarly, a Forbes report indicates that there will be 2.7 million data science and analytic job openings by 2020.

Keywords: Infusing Big Data Analytics, Lecture, Hands on Training, MATLAB, Student Knowledge, Math – Talk, Learning Strategies

INTRODUCTION:

A recent survey from Harvard Business Review indicates that 85% of the organizations that they surveyed revealed that they planned to fill 91% of their data science jobs with new graduates. Though the private sector asks at least a master's degree in mathematics or statistics for data analytics jobs, the government sector requires only a bachelor's degree. Moreover, it is impractical to fill this huge demand for big data analytics through only from graduate degree holders in mathematics related fields.

The Harvard Business Review report also indicates that 70% of the organizations that they surveyed report that finding big data talent as challenging or impossible. The hiring scale for big data jobs is 73%. This high score indicates the amount of difficulty in finding skillful candidates for the job. In order to address this serious problem, Alabama State University with the support of Auburn University employed a unique technique called infusing big data analytics in various undergraduate mathematics and statistics courses.

Our big data course modules walked students through producing working solutions by having them perform a series of hands to big data exercises developed specifically to apply cutting edge industry techniques with each mathematics and statistics course module. We strongly believe that equipping students with such skills greatly improves their employability. Linear algebra concepts such as feature extraction, clustering, and classification involving the manipulation of large matrices are extensively used in big data analytics.

Therefore, this is a natural course to start introducing students to big data analytics. This paper presents our four years of experience in adapting and integrating big data concepts into undergraduate linear algebra courses.

LINEAR ALGEBRA AND BIG DATA

Linear algebra topics such as linear equations, eigenvalue problems, principal component analysis, singular value decomposition, quadratic forms, linear inequalities, linear programming, optimization, linear differential equations, modeling and prediction, and data mining algorithms (frequent pattern analysis, classification, clustering, and outlier detection) are frequently used in practice by Big Data Analytical applications. Particularly, matrix algorithms constitute the core of modern Big Data Analysis. Because, matrices provide a convenient mathematical structure for modeling a wide range of applications' data. For example, information about 'N' objects with 'D' features can be easily described/encoded by an 'NxD' matrix. Manipulations of large matrices are used in feature extraction, clustering, and classification.

Matrix decomposition is used in principal component analysis for dimension reduction. Similarly, application of eigenvectors used in Google's PageRank method.

The versatility of graphs can be seen from their ability to illustrate important aspects of modern computer science, the intricacies geography, linguistic complexities, and the consistency of chemical structures. Incorporating linear algebra allows for representation of these graphs as matrices, this completes the pertinent task of enhancing their computational aspects. Linkeddata is usually represented by a graph in Big Data applications. We define a graph G as an ordered pair

(V(G), E(G)) consisting of a set V(G) of vertices and a set E(G), disjoint from V(G), of edges, together with an incidence function ψG that associates with each edge of G an unordered pair of (not necessarily distinct) vertices of G. Using this definition the vertices of a graph can represent webpages, genes, image pixel, or interacting users, and edges represent relations or links between the vertices. Notions such as centrality, shortest path, and reachability can be derived from the graph using graph analytics. A widely used practical application of large graph analytics is the internet search engine. Some widely used methods in Big Data Analytics that incorporate the utilization of graphs is to visualize big data as graphs (e.g. the World Wide Web), computation for strongly connected large graphs (e.g. PageRank for strongly connected graphs), and finding matchings in bipartite graphs (e.g. internet advertising).

Many practical applications of large graph analytics exist, including internet search engines. One obstacle that presents itself with graphs is the humongous sizes that are involved with the numerous millions of vertices that could exist. Low rank approximation of the adjacent matrices or graph Laplacians in relation to the graphs compared in the analysis of and interpretation of the data.

PRACTICAL BIG DATA APPLICATIONS THAT USE LINEAR ALGEBRA INCLUDE, BUT NOT LIMITED:

1) Google'sPage Rank Algorithm
2) Recommender Systems (e.g., Netflix, Pandora, Spotify)
3) Topic Modeling (e.g., Wikipedia, Genome Sequence Analysis)
4) Social Network Analysis (e.g., Facebook, My Space, LiveJournal, YouTube)
5) Internet Search,
6) Complex System Analysis (e.g., Biological Networks),
7) Image Segmentation

8) Graph Clustering
9) Link Prediction, and
10) Cellular Networks.

In Big Data Analytics, linear equations along with matrices are widely used in large network analysis, Leontief economic models, a model for the economics of a whole country/region in which consumption equals production, and ranking of sports teams. Eigenvalues and eigenvectors are used in Google's PageRank algorithm, networks clustering, and weather system modeling and spectral decomposition, a matrix approximation technique which uses eigenvectors, are used in spectral clustering, link prediction in social networks, recommender systems with side information, densest k-subgraph problem, and graph matchings. Principal component analysis and Singular Value Decomposition techniques are used to compare the structure of folded proteins and in Dimension Reduction techniques such as image compression, face recognition, and El Nino. Optimization, a minimization of a quadratic expression, and linear programming are widely used in the stable marriage problem, production planning, portfolio selection, transportation problem, minimization of production costs, minimization of environmental damage, and maximization of profits. Practical applications like face recognition, finger print recognition, plagiarism finding, and Netflix movie ratings are using similar items and frequent patterns concepts.

INFUSING BIG DATA ANALYSIS IN UG LINEAR ALGEBRA COURSE:

To facilitate the Big Data infusion and active learning in the linear algebra course, we employed a to part module. The first part focused on theoretical and conceptual ideas behind the methodsunder discussion and the second part had hands on experimentation using real world data. The students are advised to use both R and Python general purpose programming languages to complete their projects. The students can also use MATLAB programming to perform their project as well as MS Excel.

The initial set of topics in which we integrated big data analysis methods were chosen using twocriteria: suitability of material for pedagogical integration of big data methods and impact on all computing and Mathematics majors. Instructors may eventually choose to expand the integration of Big Data concepts to other computing and Mathematics courses in the future.

The following big data lectures and lab modules were infused to the existing linear algebra course.

LECTURE:

To begin, the students were provided with a pretest to gauge their understanding of Big Data Analytics and how linear algebra can be applied. This data was later paired with a post test that was given as the last component of the module. The instructor presented the class with a concept of "Big Data" that best suits the linear algebraic view point. In linear algebraic terms we define big data as data that can be represented as an m×n array with large m and large n. The goal of the lecture was to reinforce topics already outlined in the course syllabus while only presenting additional information, if it was absolutely necessary for students to understand aspects of the modules. Some of the topics already incorporated into the course curriculum include linear equations and matrices, eigenvalues, eigenvectors, and singular value decomposition. The lecture focused on methods for gathering data and representing such data in the form of matrices and the utilization of basic applications of linear algebra on said matrices. The primary source for such data was www.data.gov and similar sites. In addition, students were presented with the PageRank algorithm and a scenario utilizing it. Lastly, the lecture introduced the topic of the Leslie Matrix and population change. Examples were kept as simple as possible for students to understand the complexities of certain algorithms or unfamiliar methods.

HANDS ON ACTIVITIES FOR STUDENTS:

1) Classify data sets into categories that describe the shape of the data distribution. For this lab activity students were encouraged to use the practical big data techniques explained during the lecture when considering linear equations relating to business problems, tax problems, economic planning models, problems for the input output matrix for an economy producing transportation, and interpret data analytic problems. The data sets for this portion of the activity were prearranged in order to allow for certain controlled outcomes and provide key discussion points.

2) Investigate real data using eigenvalues, and eigenvectors to decipher information from the data set. For this lab, students were encouraged to solve practical problems such as economic development problems, analysis of situations as diverse as land problems, applications in structural engineering, control theory problem, vibration analysis problem, electric circuits problem, and advanced dynamic problem and so on. Of the previous topics stated, instructors were given the freedom to select from this group areas to focus on as part of the second portion of the activity. However, data was once again provided from sources such as www.data.gov.

MATLAB

MATLAB was used as the primary computing tool when calculations would exceed those gained from simple introductory examples. Given that all students were not previously familiar with MATLAB, step by step “cheat sheets” were used when working through examples and as a reference for using MATLAB.

Students were then provided with an out of class assignment to further their research, as well as reinforce their understanding of how linear algebra could be applied to Big Data Analytics.

ASSIGNMENT:

The assignment focused on a complexity analysis of the PageRank Algorithm and was titled “The Mathematics of Google Search.” Given that this was an undergraduate course models were constructed from real world data, but altered as not to produce an unnecessary number of iterations that would distract from the purpose of the assignment.

Upon their return to class students were asked to engage in a classroom discussion based on questions contained in their assignment called “Questions for Class Discussion.”

These questions were meant to understand, from the students’ point of view, the value of the module and their feelings toward applying class material in such a manner.

In order to understand as completely possible the student's competency, certain standards wereconsidered. There were several standards addressed to some degree by this project. The standards are: Students will be able to collect data, display data in a graphical manner, interpret data as a matrix, apply techniques already contained within the curriculum, develop models, determine levels of accuracy needed, organize materials, interpret the data and draw a conclusion from the data explain their thought process.

The criteria which identified indicators of good performance on the task and in class discussions were:

- Accuracy of calculations
- Accuracy of Models and Graphs

- Usage of Algorithms
- Organization of Calculations
- Clear Explanations

As mentioned, the culmination of the module came in the form of a posttest designed to, among other things, show if the students had a better understanding of the topic than when they began.

STUDENT KNOWLEDGE:

Students in each class completed pre and posttests to examine changes over the duration of the module implementation. In each class, there were students that failed to complete the pre, post, or both tests. Overall, scores on the pre-tests averaged just 36.63% while averaging 80.69% on the post-tests. The box plot and paired t-test results are shown respectively. The two tailed P value for the 95% confidence interval less than 0.0001, by conventional criteria, this difference is extremely statistically significant.

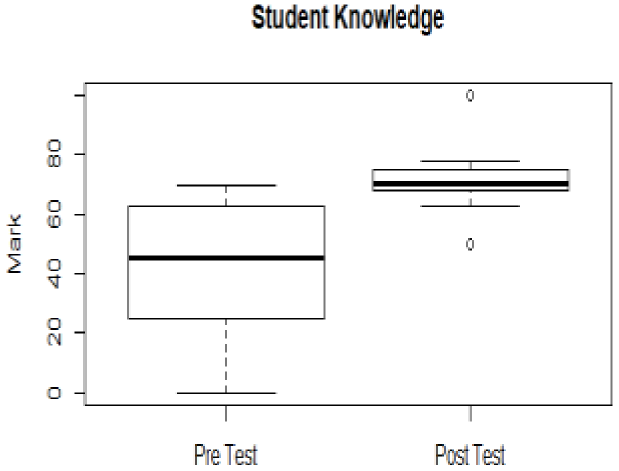

Figure-1: Box Plot (Student Knowledge)

P value and statistical significance:
The two-tailed P value is less than 0.0001 By conventional criteria, this difference is considered to be extremely statistically significant.
Confidence interval:
The mean of PreTest minus PostTest equals -44.0588
95% confidence interval of this difference: From -52.2025 to -35.9152
Intermediate values used in calculations:

t = 10.8667
df = 50
standard error of difference = 4.054

Review of the data:

Group	**PreTest**	**PostTest**
Mean	36.6275	80.6863
SD	23.3169	15.8600
SEM	3.2650	2.2208
N	51	51

Paired t test results

MATCHED PRE-POST STUDENT KNOWLEDGE:

To better examine gains made by students after using these modules, the analysis was limited tothose students with complete pre- and post-test data. A total of 44 students had completed both the pre- and post-test. Scores for this matched sample increased from pre-test (M=35.14, SD=23.5) to post-test (M=83.61, SD=14.75). Using a paired-samples t-test, changes from pre- test to post-test were statistically significant (t=14.09, p<0.0001). These results are summarized in figure-3 (boxplot) and paired t-test.

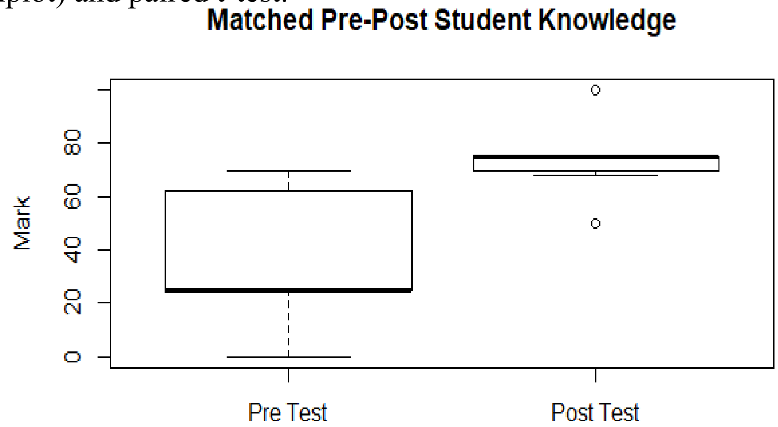

Figure-2: Box Plot (Matched Student Knowledge)

P value and statistical significance:

The two-tailed P value is less than 0.0001

By conventional criteria, this difference is considered to be extremely statistically significant.

Confidence interval:

The mean of PreTest minus PostTest equals -48.4773

95% confidence interval of this difference: From -55.4153 to -41.5393

Intermediate values used in calculations:

t = 14.0910

df = 43

standard error of difference = 3

Review of the data:

Group	**PreTest**	**PostTest**
Mean	35.1364	83.6136
SD	23.4963	14.7463
SEM	3.5422	2.2231
N	44	44

Paired t test results

CONFIDEMCE IN USING BIG DATA MODULES IN CLASS:

In spring, nearly 80% of the overall survey respondents were either juniors or seniors and nearly 30% were enrolled as computer science majors. The sample was balanced in terms of gender (52.9% female), but offered little diversity in terms of race, ethnicity or disability. In fall,nearly 95% of the overall survey respondents were either juniors or seniors and over 38% were enrolled as computer science majors. The sample offered little diversity in terms of race, ethnicity or disability and over 32% were female. In spring, nearly 95% of the overall survey respondents were either juniors or seniors and nearly 28% were enrolled as computer science majors. The sample had a larger number of males (53.1%), with majority of participants identifying as Black (87.5%)

and primarily not identifying with Hispanic or Latino ethnicity (90.6%). Using a 5 point scale (1=little of no confidence…5=A great deal of confidence), students were asked to respond to 31 different potential big data modules/applications. These responses were requested prior to the implementation of modules in math coursework. In spring, only 8 out of 26 modules (30.8%) received an average response of 3 or above, in fall,only 2 out of 26 modules (6.5%) received an average response of 3 or above, and in fall, 30 out of 31 modules (96.8%) received an average response of 3 or above.

STUDENT ACADEMIC EFFICACY, MOTIVATION AND LEARNING STRATEGIES IN MATH COURSES:

Finally, students were asked to respond to survey items pertaining to their level of academic efficacy, motivation and goals in learning math, and strategies that they use and prefer to learn math.

ACADEMIC EFFICIENCY:

Students were asked to respond to five items related to their academic efficacy as it pertains to the math class in which they were enrolled. Overall, students reported agreat deal of confidence in their academic abilities with the average for each term above 4 (on a 5 point scale). Students believed that they would learn if they tried, worked hard, and did not give up. They also believed that they could master the skills and figure out the most difficult class work.

GOALS IN MATH:

While all goals were important to them, students believed that getting a good grade was most important. They also wanted to meet requirements for their degree, improve their ability to communicate math ideas to others, learn new ways of thinking and specific procedures for solving math problems.

PREFERRED LEARNING ENVIRONMENTS:

When asked to indicate their perceptions of statements describing different learning environments, students reported the greatest agreement with “the instructor explains the solutions to problems” and “the assignments are similar to the examples considered in class.” Students also indicated situations in which they compared their math knowledge to other students, studied their notes, explained ideas to others, worked in small groups, and got frequent feedback on their mathematical thinking. They were less supportive of having the class critique their solutions, exams that prove their skills and group presentations.

GENERAL LEARNING STRATEGIES USED BY STUDENTS:

In general, students reported using a variety of strategies in their math classes and not giving up when they get stuck. They most frequently reported finding their own ways of thinking and understanding and reviewing their work for mistakes or misconceptions. They also reported checking their understanding of what a problem asking, studying on their own and using their intuition about what an answer should be.

MOTIVATION TO LEARN MATH-TALK VALUE:

Students reported high levels of task value, indicating their belief in the importance and utility of course content in their math classes. Their understanding of math is extremely important to them and their motivation to learn math is strong.

LEARNING STRATEGY:

CRITICAL THINKING:

In terms of learning math, students reported many strategies that require critical thinking. They reported developing their own ideas based on course content and evaluating the evidence before accepting a theory or conclusion. They also reported questioning what they read or heard in class and thinking of possible alternatives.

SELF REGULATION:

Students reported using many effective self regulation strategies in their math classes. In particular, they pay careful attention to concepts that they find confusing and focus on studying and reviewing these, so they learn them.

TIME AND STUDY ENVIRONMENT MANAGEMENT:

Another positive strategy reported by students related to the management of their time and study environment. They reported attending class regularly, finding a place to study and keeping up with the weekly readings and assignments.

The reliability of these scales was generally supportive, with internal consistency estimates ranging from 0.491 to 0.926, with a median of 0.867. Perceptions were also very positive as overall scale means exceeded the scale midpoints.

CONCLUSION:

There are many linear algebra big data modules and infused them into existing core undergraduate mathematics courses over a period of four years. The modules were taught using examples that were worked through interactively during class. The students then worked on assignments that incorporated the new big data instructional concepts. We have evaluated thebig data modules effectiveness through pre- and post-tests, and surveys. The paired-samples t- test results show that matched pre-post student knowledge is statistically significant. Regarding confidence in using big data modules in class, we had mixed results. Students' perception was very positive as overall scale means exceeded the scale midpoints. We feel the courses were a success but indicated there was room for improvement.

REFERENCES:

1. Anton, Howard (1987), Elementary Linear Algebra, John Wiley & Sons.
2. Arrow, J. (1963), Social Choice and Individual Values, Wiley.
3. Ball, W.W. (1962), Mathematical Recreations and Essays, MacMillan (revised by H.S.M. Coxter).
4. Bennett, William (March 15, 1993), "Quantifying America's Decline", Wall Street Journal
5. Casey, John (1890), The Elements of Euclid, Books I to VI and XI (9th ed.), Hodges, Figgis, and Co.,
6. Clark, David H.; Coupe, John D. (Mar. 1967), "The Bangor Area Economy Its Present and Future", Report to the City of Bangor, ME.
7. Dalal, Siddhartha; Folkes, Edward; Hoadley, Bruce (Fall 1989), "Lessons Learned from Challenger: A Statistical Perspective", Stats: the Magazine for Students of Statistics: 14-18
8. Davies, Thomas D. (Jan. 1990), "New Evidence Places Peary at the Pole", National Geographic Magazine 177 (1): 44.
9. Ebbing, Darrell D. (1993), General Chemistry (Fourth ed.), Houghton Mifflin.
10. Ebbinghaus, H. D. (1990), Numbers, Springer-Verlag.