Original Article

Imperative Linear Algebra For Data Science With R-Software

[1]R. Nagarathinam, [2] R Punitha,[3] Mrs.M.Nazreen Banu

[1] Associate Professor, Department of Mathematics, Dr.MGR Educational and Research Institute, Chennai-95
[2] Associate Professor, Department of Mathematics, Thassim Beevi Abdul Kader college for Women, Kilakarai, Tamilnadu, India, varshre@gmail.com
[3] Associate Professor, Department of Mathematics, Thassim Beevi Abdul Kader college for Women, Kilakarai, Tamilnadu, India, nazz_reen@yahoo.com

*Corresponding Author:

R. Nagarathinam, Associate Professor, Department of Mathematics, Dr.MGR Educational and Research Institute, Chennai-95
nagarathinam.hs@drmgrdu.ac.in
DOI: 10.47750/pnr.2022.13.S10.174

Abstract

Data science and machine learning are built on linear algebra. Machine learning and data science make extensive use of linear algebra, a branch of Mathematics. Machine learning relies heavily on linear algebra. Matrix representations are commonly used in machine learning models. Using linear algebra in data science means regularizing, reducing to dimensions, recognizing images, learning algorithms, and analyzing images. Many data science algorithms are based on linear algebra. This article will cover three uses of linear algebra in three different data science domains. We will discuss loss functions from the perspective of machine learning, and image convolution from the perspective of computer vision. Any prospective data scientist must learn R since it is a very strong language designed specifically for data analysis and data visualisation. With linear algebra, R is extremely useful. It has built-in data types like matrices and vectors.

Keywords: Linear Algebra, Data Science, Algorithms, R-Software**.**

INTRODUCTION

The area of mathematics known as "linear algebra" deals with linear equations and how they are represented in vector spaces and through matrices. The study of vectors and linear functions is what linear algebra is, to put it simply. Linear algebra involves the study of matrices, determinants, linear transformations, vector spaces, and subspaces and uses closed vectors that operate under addition and scalar multiplication. It enables us to do mathematical operations and comprehend geometric notions in greater dimensions. Nearly all branches of mathematics, including geometry and functional analysis, are based on linear algebra. Understanding its ideas is essential for comprehending the theory underlying data science.

Data science is a branch of study that combines subject-matter knowledge, programming abilities, and a working understanding of mathematics and statistics to draw out valuable insights from data. Data scientists use machine learning algorithms to analyse data from a variety of sources, including text, images, videos, and audio, to create artificial intelligence systems that can carry out activities that often require human intellect. To study and analyse real-world phenomena using data, "data science" is a concept that combines statistics, data analysis, informatics, and their related methodologies. In the context of mathematics, statistics, computer science, information science, and domain knowledge, it makes use of methods and theories from a variety of domains. A data scientist is a person who writes programming code and uses statistical expertise in conjunction with it to derive insights from data.

Data scientists may be able to avoid using linear algebra for a while, but not for very long. Here are several ways that linear algebra can help with computer vision and machine learning. Few people think of linear algebra when discussing data science in general or specific subfields like machine learning or computer vision. Because the modern tools we use to perform data science algorithms do a great job of disguising the underlying math that makes things function, linear algebra is sometimes overlooked. In data science, linear algebra is used.

Impact of Linear Algebra in Data Science

Area of mathematics called linear algebra is very helpful in machine learning and data science. The following data science fields make use of linear algebra:

- Regularization
- Reduction to dimension
- Image recognization
- Working with data sets
- Machine learning
- Computer vision

REGULARIZATION

Regularization is one of the biggest hurdles in machine learning, especially for beginners in data science. It is when a model is too close a fit for the available data, to the point that it does not perform well with any new or outside data. A concept called "**regularization"** is used to prevent the model from overfitting. This concept makes use of linear algebra as it uses the "**norm".** The norm can be defined simply as the magnitude of a vector. This magnitude can be calculated in various ways one popular way is using the Euclidean Distance. ie, using the distance from the origin. Regularization prevent overfitting as it adds the norm of the weight vector to the cost function. This makes sure that model does not become overly complex as aim is always to reduce the cost function, and therefore have to reduce this norm. This is much better understood by someone who knows the basics of linear algebra.

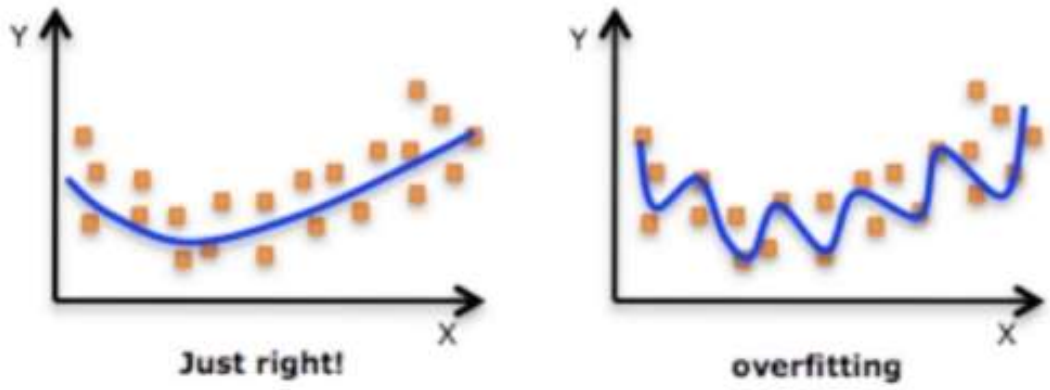

Figure 1

REDUCTION TO DIMENSION

While making Machine Learning Models, often come across data that is made up a hundreds or even thousands of variable. Our model becomes more and more complicated as these variables increase. Dimensionality reduction is the technique that reduces the number of input variables in data set. Since datasets can be easily represented as matrices, certain matrix factorization methods can be used to reduce a matrix and hence the dataset into its constituents parts. Then any operation that used to be performed on the original matrix, could be performed on the smaller matrices. Decomposition method like LU matrix decomposition and QR matrix decomposition can be easily performed using python programming.

LU Decomposition

Consider the system of equation written as a matrix equation:

$X_1+X_2-X_3 = 4$

$X_1 - 2X_2+3X_3= -6$

$2X_1+3X_2+X_3=7$

$$\begin{bmatrix}1 & 1 & 1\\ 1 & -2 & 3\\ 2 & 3 & 1\end{bmatrix}\begin{bmatrix}X1\\ X2\\ X3\end{bmatrix}=\begin{bmatrix}4\\ -6\\ 7\end{bmatrix}$$

$$AX = B$$

We can solve the system using LU Decomposition

Let A= LU and substitute into AX=B

Solve LUX=B FOR X to solve the system

Let UX=Y

LY=B and UX=Y

First solve LY=B and Y and the solve UX=Y for X

LINEAR ALGEBRA IN IMAGE RECOGNIZATION

When implementing data science models, especially in deep learning, it come across data in the form of image, however we cannot just pass an image to a model and expect it to understand it. It need to convert each image into something Mathematical or Statistical to be understood by the model. This is where linear algebra come in.

IMPACT OF LINEAR ALGEBRA IN IMAGE PROCESSING

Image processing is the manipulation of images using mathematical operations. With the introduction of computers, processing is now done using computer graphic algorithms on digital images obtained through a digitization process or directly using any digital device. Digital image processing is the use of a computer to perform image processing on digital images. Linear algebra can be used to perform computer graphics operations such as rotation, skewing, scaling, Bezier curves, reflections, dot and cross products, projections, and vector fields. Other more complex operations, such as filters, necessitate the use of linear algebra in conjunction with other mathematical tools. Linear algebra deals with matrices and all the operation to be Performed on matrix. Any image is made of pixels, which are nothing but coloured square of varying Intensitics (for gray scale image it could be a single number with the intensity, and for coloured images, it could be the RGB value).

Figure 2

WORKING WITH DATA SETS

When building a machine learning model, it will most probably dealing with large data sets having multiple rows and columns. These are nothing but matrices when you spilt your dataset into training and testing data, you are performing operations on these matrices. Matrices are the key data structure in linear algebra and it deals with the various operations performed on matrix, including row and column transformations, transpose of a matrix, addition or scalar multiplication in matrices

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

This is nothing but a **5*4** matrix each record is a row and is indexed with the numbers 0,1,….4 . Each column has its name on top.

MACHINE LEARNING

With the use of machine learning (ML), which is a form of artificial intelligence (AI), software programmes can predict outcomes more accurately without having to be explicitly instructed to do so. Machine learning algorithms predict new results using historical data as input. The way in which a machine learning algorithm learns to improve its prediction accuracy is a common way to classify traditional machine learning. There are four fundamental strategies: reinforcement learning, semi-supervised learning, unsupervised learning, and supervised learning. The kind of data that scientists wish to predict determines the kind of algorithm they use.

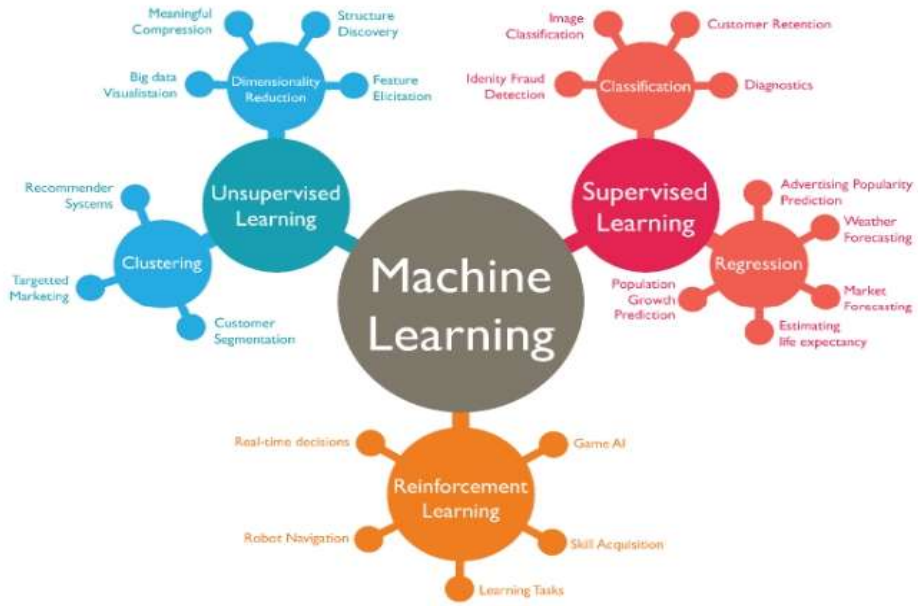

Figure.3

LINEAR ALGEBRA IN MACHINE LEARNING

Vectors, matrices, and linear transforms are the focus of the mathematics subject known as linear algebra. From the notation used to describe the operation of algorithms through the implementation of algorithms in code, it serves as a crucial basis for the field of machine learning. Despite the fact that linear algebra is essential to the subject of machine learning, the close connection is frequently ignored or described using impersonal ideas like vector spaces or certain matrix operations.

The following are significant fields of application made possible by linear algebra:

Data representation

Representation using learnt models

Eigenvectors for word embeddings

REPRESENTATION OF DATA

Data, which serves as the fuel for ML models, must first be transformed into arrays before being fed to your models. Matrix

multiplication is one of the operations carried out on these arrays (dot product). Additionally, the output is returned, which can also be seen as a changed matrix or tensor of integers.

EMBEDDING WORDS

Just above it is the idea of using a lesser dimensional vector to express large-dimensional data (consider a large number of variables in your data).

EIGENVECTORS (SVD) (SVD)

Principal component analysis allows us to decrease the number of characteristics or dimensions in the data without sacrificing the quality of any of them.

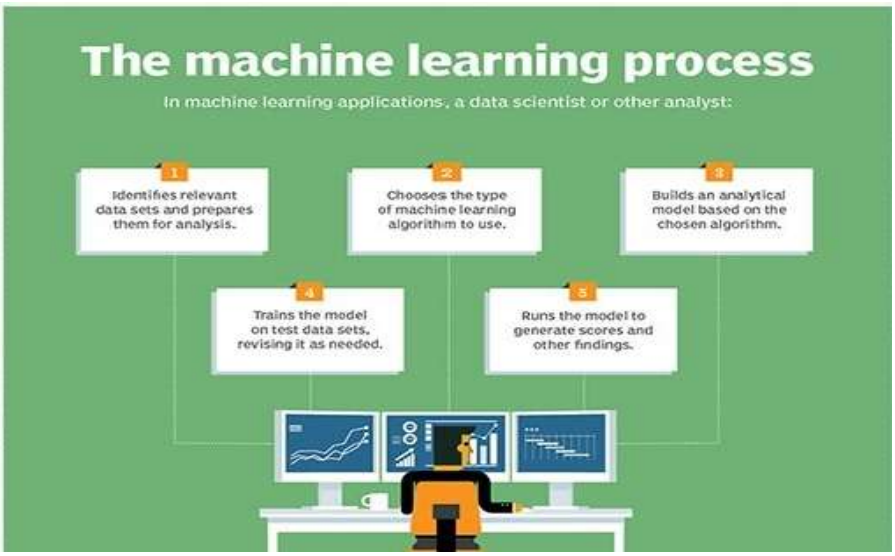

Figure 4

LINEAR ALGEBRA IN COMPUTER VISION

For machine learning, data science, and other related fields like computer vision and natural language processing, linear algebra is a potent tool. With its own implementations in the form of matrices, vectors, and tensors, linear algebra is used in computer vision. This includes basic operations, such as linear transformation, matrix operations, linear combination, and dependency of variables.

Significant contributions from linear algebra have been made to the difficult calculations used in computer vision. Large and complex matrices are necessary for the computations and operations involved in the complex matrix multiplications used in computer vision techniques. The computer vision system uses a variety of techniques to extract data from images, including compression, rotation, flip-flopping, convolution, noise reduction, object detection, facial recognition, etc**.**

The study of visual knowledge extraction is known as computer vision. and the fundamental idea of linear algebra is applied to the extraction. Vectors, matrices, and tensors are LA concepts.

VECTOR

A vector is a type of 1D array that is typically described with magnitude and direction.

MATRIX

It is a 2D array of numbers called a matrix. Think about an image's matrix-based pixel representation as an example. Projections, translations, rotations, scaling, and affine transformations are some of the operations of matrices.

TENSOR

It is a generalisation of matrices and vectors.

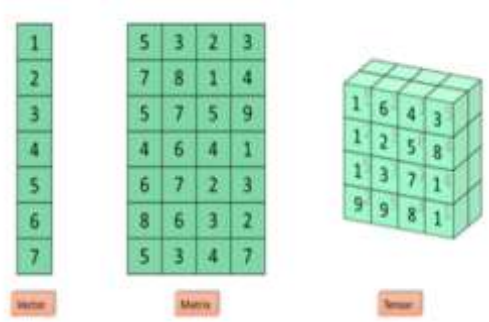

Figure. 5

VECTOR PRODUCT: MATRIX

The identity matrix is created by combining the unit vectors $e_1,..., e_m$. A simpler approach to refer to random vectors $a_1,...a_n$ from the same vector space in a matrix is to use the letter I.

VECTORS AND MATRICE IN DATA SCIENCE

The some essential of linear algebra in the context of data science applications.

Vectors organize information that cannot be expressed as a single number and for which there exist a concept of scaling and addition. Matrices group together multiple vectors. The matrix –vector product expresses a linear combination of the column vectors of the matrix. solving a linear system

$A_x=b=Ib$, to find $x€R^m$ for given $b€R^m$, re –expresses the linear algebra $b= b_1e_1+…b_me_m$,

$I=[e_1e_2….e_m]$, as another linear combination $b=x_1a_1 + x_2a_2+…..x_na_n$, $A=[a_1a_2…a_n]$ for certain problems the linear combination A_x might be more insightful, but the above transformation is information preserving, with b_1x both having the same number of components

Finding the best approximation of some given $b€R^m$ by a linear combination A_x of the column vector of $A€R^{m*n}$ is known as a least squares problem and transforms the information from the m components of b into n components of x, and knowledge of the form of the column vectors. If m>n and the form of the column of A can be successfully stated, the transformation compresses information.

USAGE OF R SOFTWARE

COMMANDS TO FIND MULTIPLICATION OF MATRIX USING R

```
>
> #Multiplication of matrix
>
> A<-matrix(c(11,12,13,14,15,16,17,18,19),nrow = 3,byrow = T)
> B<-matrix(c(20,21,22,23,24,25,26,27,28),nrow = 3,byrow = T)
>
> A*B
     [,1] [,2] [,3]
[1,]  220  252  286
[2,]  322  360  400
[3,]  442  486  532
```

COMMANDS TO FIND INVERSE OF MATRIX USING R

```
> #Inverse of matrix
>
> B<-matrix(c(30,31,40,41,50,51,60,61,70),nrow = 3,byrow = T)
>
> A<-solve(B)
> A
             [,1]    [,2]        [,3]
[1,] -0.16208333 -0.1125  0.17458333
[2,] -0.07916667  0.1250 -0.04583333
[3,]  0.20791667 -0.0125 -0.09541667
>
```

COMMANDS TO FIND DETERMINANT OF A USING R

```
>
> #Determinant of A
> det(A)
[1] -0.0004166667
> #Calculating eigenvalues and eigenvectors
> A<-matrix(c(30,31,40,41,50,51,60,61,70),nrow = 3,byrow = T)
> e <- eigen(A)
> e$values
[1] 147.737576   5.317459  -3.055035
```

COMMANDS TO FIND EIGENVECTOR USING R

```
> e$vectors
           [,1]       [,2]        [,3]
[1,] -0.3948374  0.4437557 -0.74478185
[2,] -0.5497457 -0.8199420 -0.06303763
[3,] -0.7361271  0.3616296  0.66432391
> |
```

EIGENVECTORS APPLICATIONS IN DATA SCIENCE

The principal component analysis of a machine learning method makes use of the idea of an eigenvector.

Assume you have data that is very high in dimensionality and has a lot of features. It's possible that the data contains redundant features. A lot of features will also decrease efficiency and take up more disc space. In this case, PCA eliminates some of the less significant features; eigenvectors save the day. Let's go over the PCA algorithm. Let's say we wish to condense a "n"-dimensional dataset into a "k"-dimensional dataset. We'll proceed in stages.

Step 1: The data are subjected to mean normalisation and feature scaling.

Step 2: We find out the covariance matrix of our data set.

Step 3: Finding the covariance matrix's Eigenvectors is step three. It is a sophisticated statistical notion. We will discover "n" Eigenvectors matching "n" Eigenvalues because our data is in "n" dimensions.

Step 4: The fourth step entails choosing "k" Eigenvectors that correspond to the "k" biggest Eigenvalues and creating a matrix in which each Eigenvector is a column. It's time to locate the less data points now. Let's say you want to shrink a data point from the data set, "a," to dimensions of "k." Multiply the dimensions of the matrix U and transpose it.

After discussing Eigenvectors, let's move on to a more complex and valuable idea in linear algebra known as singular value decomposition, or SVD for short. To fully comprehend it, a thorough investigation of linear

SINGULAR VALUE DECOMPOSITION

Consider receiving a feature matrix A. We divide our matrix A into three constituent matrices for a specific purpose, as the name would imply. Additionally, it has been asserted on occasion that SVD is a generalisation of Eigen value decomposition. A data set's redundant characteristics are removed using SVD. Let's say you have a data set with 1000 features. There will undoubtedly be redundant features in any real data set with this many attributes.

Figure. 6

This tiger can be rendered in monochrome and seen as a matrix whose elements stand in for the pixel intensity and pertinent position. In other words, the matrix comprises data in the form of rows and columns concerning the intensity of pixels in the image. This image displays many images with various resolutions that correlate to various levels. Just assume for the time being that a higher rank indicates that there is more information about pixel intensity.

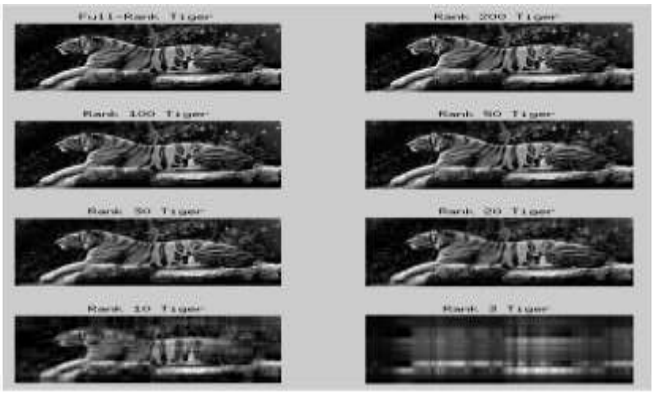

Figure. 7

It is obvious that we can achieve a reasonably good image with 20 or 30 ranks rather than 100 or 200 ranks, and in a scenario of highly redundant data, that is what we want to do. What I'm trying to say Is it true that we don't need to save every piece of data from the original dataset in order to generate a valid hypothesis? However, some of the features make it difficult to determine the optimum algorithm. For instance, multicolinearity in linear regression is caused by the presence of redundant characteristics. Additionally, some qualities don't matter for our model. The method fits better, is more time efficient, and uses less disc space when these features are removed. The application of singular value decomposition

CONCLUSION

In the actual world, linear algebra is quite useful. In data science, linear algebra techniques are utilised to increase algorithm performance and produce more accurate findings. In this paper, it has gathered thatarethe uses of linear algebra in data science and provided an overview of each technique. To analyse the data sets, the data scientists can utilise linear algebra as a tool. Given the continually growing search outputs and the accessibility of the available evidence, which is a specific issue for the study sector in terms of quality improvement, machine learning algorithms are of special relevance. Regularization, dimension reduction, image identification, machine learning, and computer vision were all topics I covered. R is very useful for linear algebra.

REFERENCES

[1] Sorin Mitran - Linear algebra for data science – University of north Carolina at chapel hill
[2] Herstein .L.N (2016) topic in algebra, second edition, wiley student edition
[3] Santiago M.L(2001) Modern algebra, Tata Megraw- hill publishing co ltd
[4] victor A. Bloomfield university of Minnesota Minneopolis, USA –using R for Numerical Analysis in Science and Engineering.
[5] Mike X Cohen, Practical Linear Algebra for Data Science, Released September 2022, Publisher(s): O'Reilly Media, Inc.,ISBN: 9781098120610.
[6] Kenneth Hoffman, Ray Kunze (1996) Linear Algebra,