Linear Algebra for Machine Learning

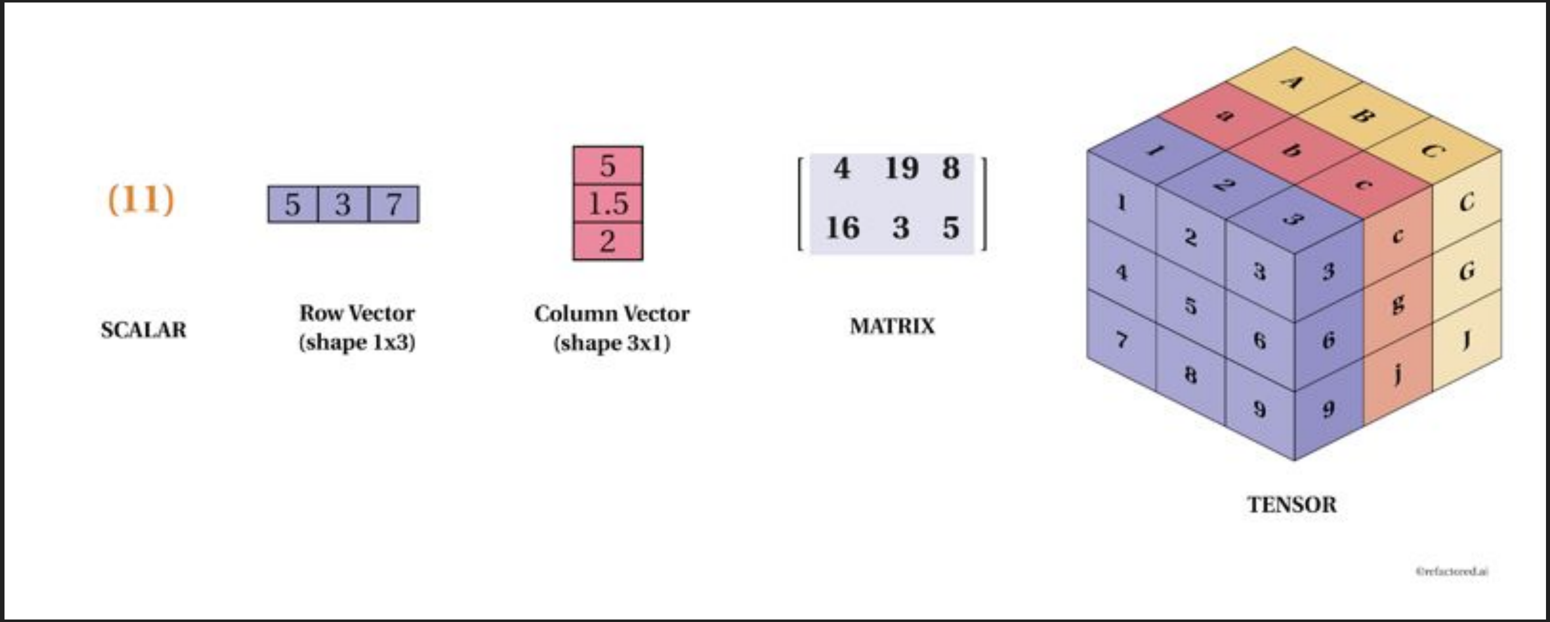
By: Tejumade Afonja

Concept you need to understand

- Scalar, Vector, Matrices and Tensors
- Multiplying Matrices and Vectors
- Identity and Inverse Matrices
- The Determinant
- Linear Dependence and Span
- Norms
- Diagonal and Orthogonal Matrices
- Eigen Decomposition
- Singular Value Decomposition

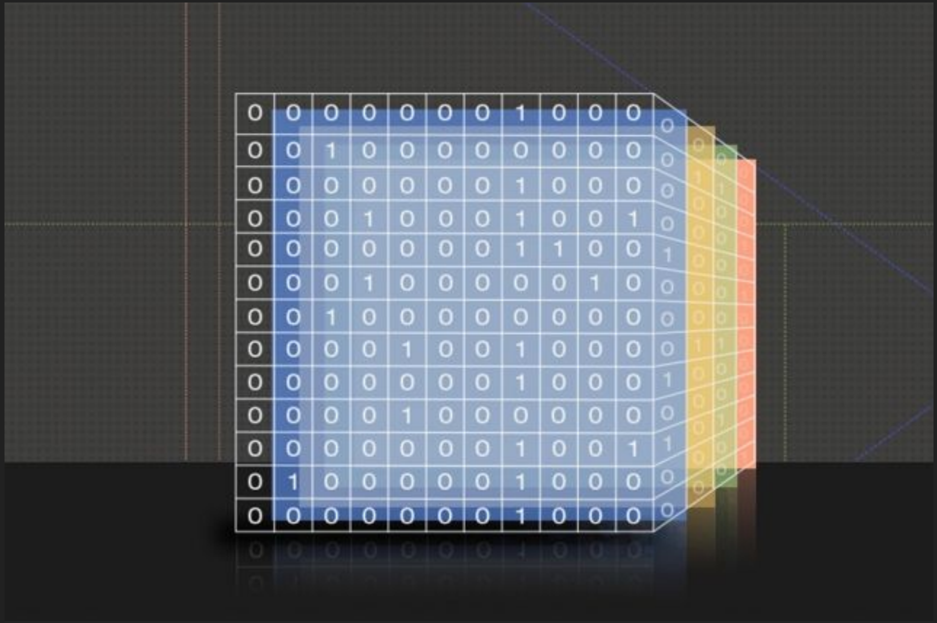
Linear Algebra is a branch of Mathematics that concerns systems of linear equations and their representation through matrices and vector spaces

Scalar, Vector, Matrices and Tensors



Tensors

- A tensor is an n-dimensional array
- Any array with more than 2 axes is referred to as a tensor



Scalars

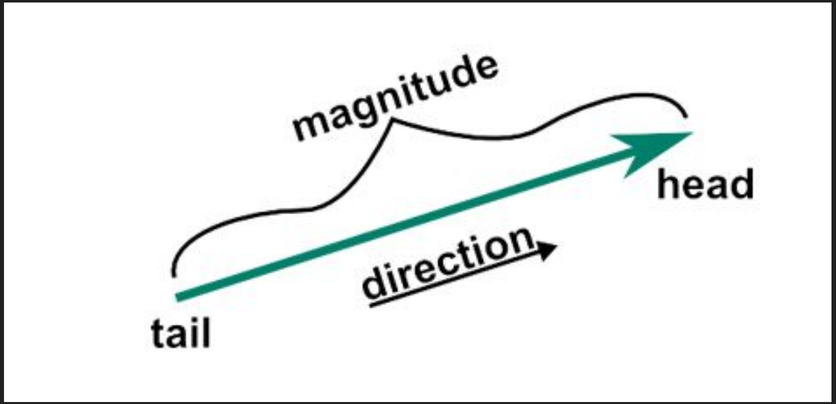
- A Scalar is just a single number usually represented in lowercase italics, x.
- It has only magnitude
- It is a 0-dimensional tensor
- It can be integers, rational number, real numbers, natural numbers etc
- It is a tensor that contain only one number - for example,

$$age\_person1 \; = 25$$

Vectors

- Vectors are arrays of numbers arranged in order usually represented in lowercase boldface
- It has both magnitude and direction
- It is a 1-D tensor
- It can be integers, rational number, real numbers, natural numbers etc

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$



Vector

- In machine learning, a vector is used to represent columns or row in a dataset. Each column(features) can denote particular features .

For example, consider column $\mathbf{age}$ below that denotes the age of 4 users in our dataset and column $\mathbf{user1}$ that denotes the various attributes that characterizes that user like (age, height, gender)

$$\mathbf{age} = \begin{bmatrix} 24 \\ 16 \\ 43 \\ 50 \end{bmatrix} \qquad \mathbf{user1} = \begin{bmatrix} 24 & 5.5 & 0 \end{bmatrix}$$

Matrices

- A matrix is a 2-D array (2-D tensor) of numbers, symbols, expressions, arranged in rows and columns
- It is a real-valued matrix has a height m and width n, $\mathbf{A} \in \mathbf{R}^{\mathbf{mxn}}$
- It can be integers, rational number, real numbers, natural numbers etc

$$\mathbf{A} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}_{\mathbf{mxn}}$$

Matrices

- In machine learning, a matrix can be used to represent an array of features and users. For example, let's say we have a matrix of 4 **AI6Lagos** students whom we represented using 3 features (age, height and gender)

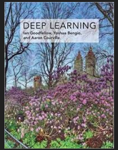
$$\mathbf{AI6Lagos} = \begin{bmatrix} 24 & 5.5 & 0 \\ 16 & 6.9 & 1 \\ 43 & 5.2 & 0 \\ 50 & 6.2 & 1 \end{bmatrix}_{\mathbf{4x3}}$$

Matrices (transpose)

- A transpose of a matrix is the mirror image of the matrix across it's diagonal.

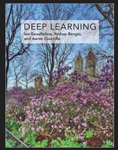
$$\textbf{AI6Lagos} = \begin{bmatrix} 24 & 5.5 & 0 \\ 16 & 6.9 & 1 \\ 43 & 5.2 & 0 \\ 50 & 6.2 & 1 \end{bmatrix}_{\textbf{4x3}} \qquad \textbf{AI6Lagos}^{\textbf{T}} = \begin{bmatrix} 24 & 16 & 43 & 50 \\ 5.5 & 6.9 & 5.2 & 6.2 \\ 0 & 1 & 0 & 1 \end{bmatrix}_{\textbf{3x4}}$$

$$\boldsymbol{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \\ A_{3,1} & A_{3,2} \end{bmatrix} \Rightarrow \boldsymbol{A}^\top = \begin{bmatrix} A_{1,1} & A_{2,1} & A_{3,1} \\ A_{1,2} & A_{2,2} & A_{3,2} \end{bmatrix}$$



More on Matrices

- Matrix $\mathbf{A}$ can be added to Matrix $\mathbf{B}$ to form Matrix $\mathbf{C}$ as long as they have the same shape. This is achieved by adding their corresponding elements $\mathbf{C} = \mathbf{A} + \mathbf{B}$ where $\mathbf{C_{i,j}} = \mathbf{A_{i,j}} + \mathbf{B_{i,j}}$
- We can add a scalar to a matrix or multiply a matrix by a scalar just by performing that operation on each element of a matrix: $\mathbf{D} = \mathbf{a} \cdot \mathbf{B} + \mathbf{c}$ where $\mathbf{D_{i,j}} = \mathbf{a} \cdot \mathbf{B_{i,j}} + \mathbf{c}$
- We can also add a matrix and a vector to yield another matrix: $\mathbf{C} = \mathbf{A} + \mathbf{b}$ where $\mathbf{C} = \mathbf{A_{i,j}} + \mathbf{b_j}$ The vector $\mathbf{b}$ **is added to each row of the matrix.**



Terms related to Matrix

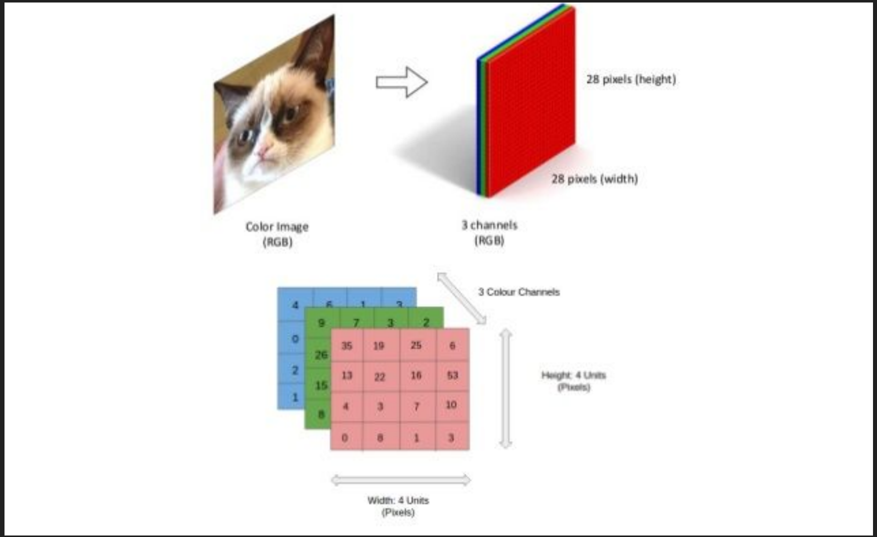
- Order of matrix - If a matrix has 3 rows and 4 columns, order of the matrix is 3*4 i.e. row*column.
- Square matrix - The matrix in which the number of rows is equal to the number of columns.
- Diagonal matrix - A matrix with all the non-diagonal elements equal to 0 is called a diagonal matrix.
- Upper triangular matrix - Square matrix with all the elements below diagonal equal to 0.

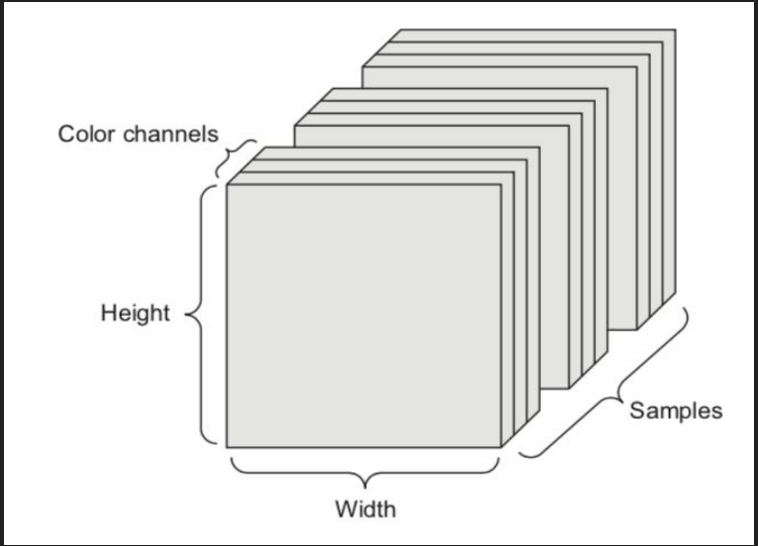
Terms related to Matrix contd

- Lower triangular matrix - Square matrix with all the elements above the diagonal equal to 0.
- Scalar matrix - Square matrix with all the diagonal elements equal to some constant k.
- Identity matrix - Square matrix with all the diagonal elements equal to 1 and all the non-diagonal elements equal to 0.
- Column matrix - The matrix which consists of only 1 column. Sometimes, it is used to represent a vector.
- Row matrix - A matrix consisting only of row.
- Trace - It is the sum of all the diagonal elements of a square matrix

Tensors

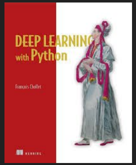
- Example - a color image is a 3D tensor with height, width and channels or a 4D tensor shape samples, height, width and channels



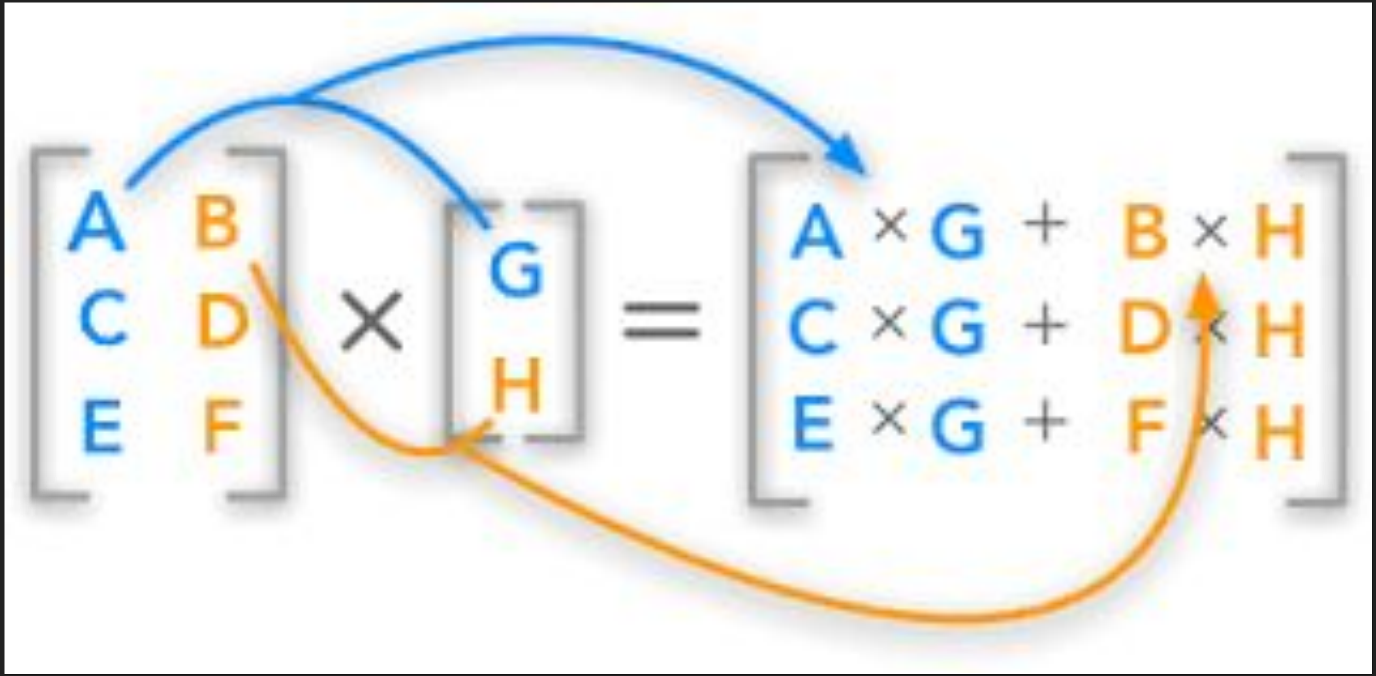


Real World Examples

- Vector data - 2D tensors of shape (samples, features)
- Timeseries data or sequence data - 3D tensor of shape (samples, timesteps, features)
- Images - 4D tensors of shape (samples, height, width, channels) or (samples, channels, height, width)
- Videos - 5D tensors of shape (samples, frames, height, width, channels) or (samples, frames, channels, height, width)



Multiplying Matrices and Vectors



Matrix Product

- Product of a Matrices A and B is a third Matrix C
- Inorder for this product Matrix C to be defined, A must have the same column as B has rows i.e
  If A is of shape $m \times n$ and B is of shape $n \times p$, then C will be of shape $m \times p$
- Matrix Product can be written as $\mathbf{C} = \mathbf{AB}$ and the product operation is defined as below

$$\mathbf{C_{i,j}} = \sum_{\mathbf{k}} \mathbf{A_{i,k}} \mathbf{B_{k,j}}$$

Matrix Product

Example,

$$\begin{bmatrix} 1 & 2 \\ 0 & -3 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 2 & 6 & -3 \\ 1 & 4 & 0 \end{bmatrix} = \begin{bmatrix} 4 & -14 & -3 \\ -3 & -12 & 0 \\ 7 & 22 & -9 \end{bmatrix}$$

Element-wise or Hadamard Product

- It is denoted as $\mathbf{A} \odot \mathbf{B}$
- It is the standard product between two matrices
- The matrices must be of the same shape

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \odot \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & a_{13}b_{13} \\ a_{21}b_{21} & a_{22}b_{22} & a_{23}b_{23} \\ a_{31}b_{31} & a_{32}b_{32} & a_{33}b_{33} \end{bmatrix}$$

Dot Product or Scalar Product

- It sometimes referred to as inner product
- It is the algebraic sum of two equal length sequence or numbers (usually coordinate vectors) and returns a single number.
- The length of a vector is the square root of the dot product by itself

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i-1}^{n} \mathbf{a}_i \mathbf{b}_i = \mathbf{a}_1 \mathbf{b}_1 + \mathbf{a}_2 \mathbf{b}_2 + \cdots + \mathbf{a}_n \mathbf{b}_n$$

Dot Product or Scalar Product

For example:

$$\begin{bmatrix}1 & 3 & -5\end{bmatrix} \cdot \begin{bmatrix}4 & -2 & -1\end{bmatrix} = (1 * 4) + (3 * -2) + (-5 * -1)$$
$$= 4 - 6 + 5$$
$$= 3$$

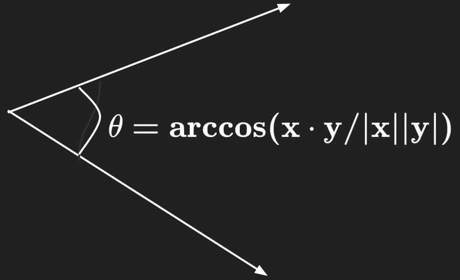
If the vectors are defined as row matrices, then the dot product is defined as a matrix product and can be written as below

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}\mathbf{b}^{\mathrm{T}}$$

Dot Product or Scalar Product

The dot product is used to find the angle between two vectors

$$\mathbf{x} \cdot \mathbf{y} = ||\mathbf{x}|| \; ||\mathbf{y}|| \cos(\theta)$$



Where $||\mathbf{x}||$ is the euclidean length of the vector denoted as $\sqrt{\mathbf{x} \cdot \mathbf{x}}$

If the angle between x and y are 90 degrees, then the dot product $\mathbf{x} \cdot \mathbf{y} = 0$ the vectors are then said to be orthogonal

When will cos(Θ)=1?

Orthogonal and Orthonormal Vectors

- Two vectors are said to be orthogonal if they are perpendicular to each other
- A set of vectors S are said to be orthonormal if every vector in S has magnitude of 1 and the vectors are mutually orthogonal.

For example, consider set of vectors $\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ \sqrt{2} \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -\sqrt{2} \\ 1 \end{bmatrix}$. These vectors are orthogonal to one another but not orthonormal. In order for them to be orthonormal, we need to normalize the vectors by dividing the vectors with their length. The magnitude of the new set of vectors will thereby be both orthogonal and orthonormal.

Properties of Matrices

- Commutative - $AB = BA$
- Associative - $A(BC) = (AB)C$ or $A+(B+C) = (A+B)+C$
- Distributive - $A(B+C) = AB + AC$

Are matrix multiplication Commutative?

Systems of Linear Equation

This is denoted as stated below

$$\mathbf{Ax} = \mathbf{b}$$

Where $\mathbf{A} \in \mathbb{R}^{\mathbf{mxn}}$ is a known matrix and $\mathbf{b} \in \mathbb{R}^{\mathbf{m}}$ is a known vector, and $\mathbf{x} \in \mathbb{R}^{\mathbf{n}}$ is a vector of unknown variables we would like to solve for.

Consider:

$$\begin{aligned}
\mathbf{A_{1,1}}x_1 + \mathbf{A_{1,2}}x_2 + \cdots + \mathbf{A_{1,n}}x_n &= b_1 \\
\mathbf{A_{2,1}}x_1 + \mathbf{A_{2,2}}x_2 + \cdots + \mathbf{A_{2,n}}x_n &= b_2 \\
&\cdots \\
\mathbf{A_{m,1}}x_1 + \mathbf{A_{m,2}}x_2 + \cdots + \mathbf{A_{m,n}}x_n &= b_m
\end{aligned}$$

We can represent these systems of linear equation as a matrix form.

Systems of Linear Equation

$$\begin{bmatrix} \mathbf{A_{1,1}} & \mathbf{A_{1,2}} & \cdots & \mathbf{A_{1,n}} \\ \mathbf{A_{2,1}} & \mathbf{A_{2,2}} & \cdots & \mathbf{A_{2,n}} \\ & & \cdots & \\ \mathbf{A_{m,1}} & \mathbf{A_{m,2}} & \cdots & \mathbf{A_{m,n}} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

Solve :

$$3x + 2y - z = 1$$

$$2x - 2y + 4z = -2$$

$$-x + \frac{1}{2}y - z = 0$$

Identity and Inverse Matrices

Given a square m-by-m matrix $\mathbf{A}$, if there is a m-by-m matrix $\mathbf{B}$ such that

$$\mathbf{BA} = \mathbf{AB} = \mathbf{I}$$

Where $\mathbf{B}$ is called the inverse of Matrix $\mathbf{A}$ denoted as $\mathbf{A}^{-1}$ and $\mathbf{I}$ is the identity matrix.

$$\mathbf{I_{3x3}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I_n}$$

Can you solve for x in Ax=b using the inverse of the matrix A?

Hint: multiply both sides by A^{-1}

Identity and Inverse Matrices

Given a square m-by-m matrix **A**, if there is a m-by-m matrix **B** such that

$$\mathbf{BA} = \mathbf{AB} = \mathbf{I}$$

Where **B** is called the inverse of Matrix **A** denoted as $\mathbf{A}^{-1}$ and **I** **is the identity matrix.**

$$\mathbf{I_{3x3}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I_n}$$

Do you think you'll get an identity matrix if you multiply a matrix by its inverse.

Identity and Inverse Matrices

- $\mathbf{A}^{-1}$ doesn't always exists
- In theory, the same inverse can be used to solve the equation many times for different values of **b**.
- $\mathbf{A}^{-1}$ is a useful theoretical tool, however it should not actually be used in practise for most software application because it can be represented with only limited precision on a digital computer and algorithms that make use of the value of **b** can usually obtain more accurate estimates of **x**

Example:

$$\mathbf{A} = \begin{bmatrix} 4 & 7 \\ 2 & 6 \end{bmatrix} \quad \mathbf{A}^{-1} = \begin{bmatrix} 0.6 & -0.7 \\ -0.2 & 0.4 \end{bmatrix}$$

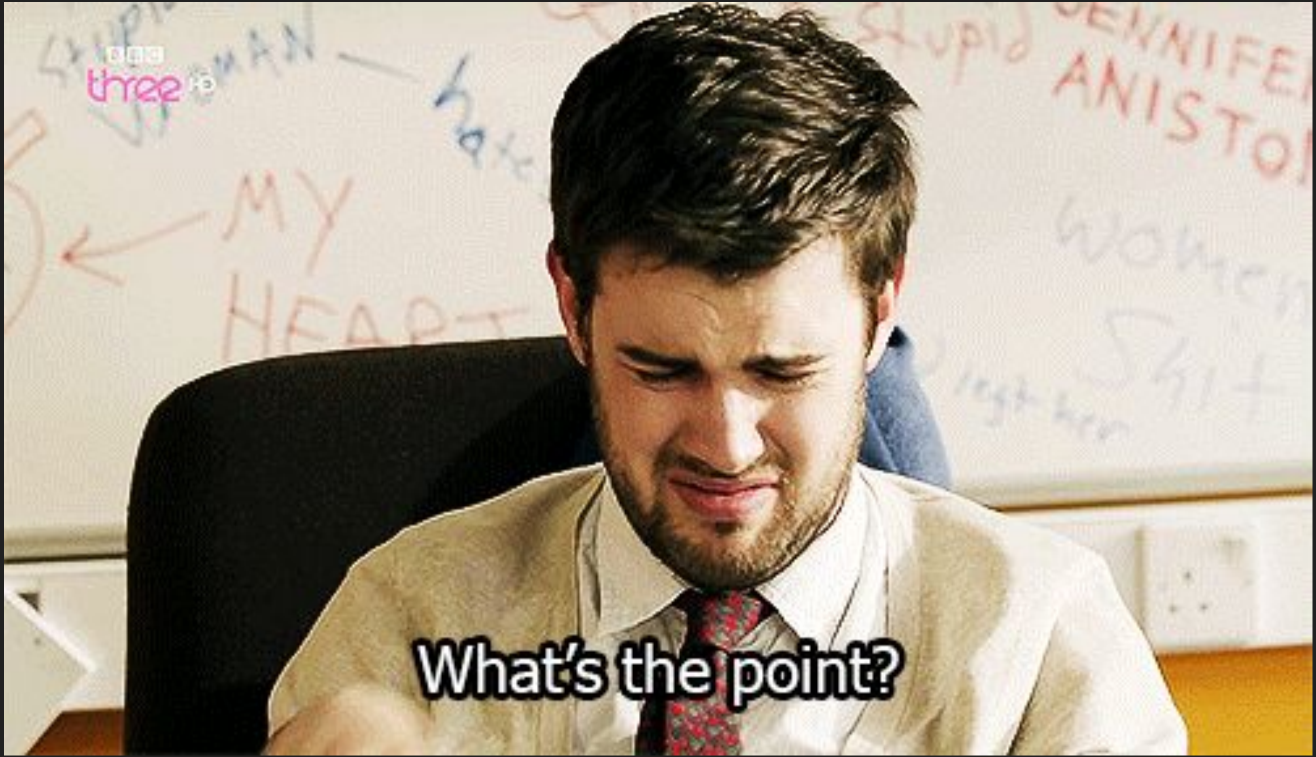
Is the inverse of Matrix A correct?

The Determinant

Can you find the determinant for a non-square matrix?

- It is the area of a square-matrix denoted as **det(A)**
- It is a functions that maps matrices to real scalars
- It is equal to the product of all the eigenvalues of a matrix
- The absolute value of a determinant can be thought of as a measure of how much multiplication by the matrix expands or contracts space
- If the determinant is 0, then space is contracted completely along atleast one dimension, causing it to loose all its volume
- If the determinant is 1, then the transformation preserves volume

Relevance of Linear Algebra to Machine Learning



Dataset and Data Files

- In machine learning, you fit a model on a dataset.
- This is the table-like set of numbers where each row represents an observation and each column represents a feature of the observation

For example: Consider a the first few rows of Iris flowers dataset. You can see that the data is represented below as a matrix where each row represents an observation and each column is the features we want to use to make prediction

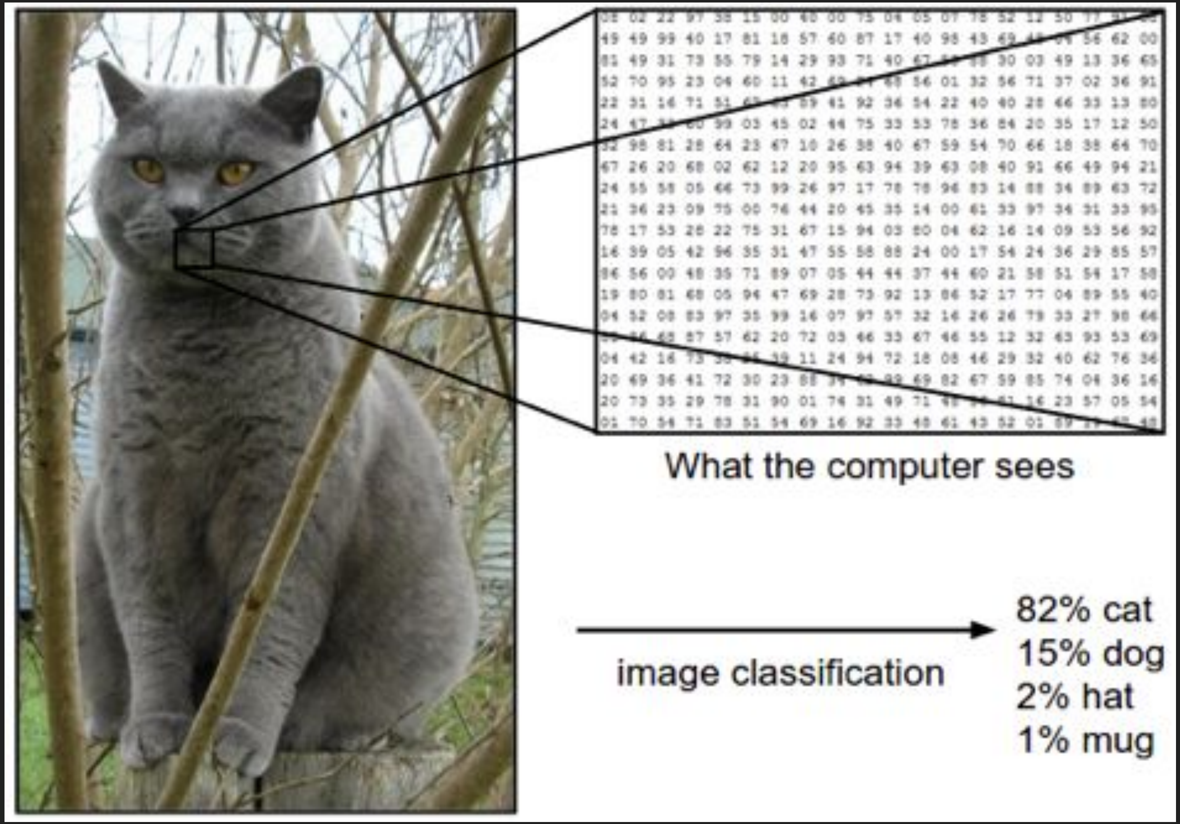
$$\begin{bmatrix} 5.1, 3.5, 1.4, 0.2, \mathrm{Iris-setosa} \\ 4.9, 3.0, 1.4, 0.2, \mathrm{Iris-setosa} \\ 4.7, 3.2, 1.3, 0.2, \mathrm{Iris-setosa} \\ 4.6, 3.1, 1.5, 0.2, \mathrm{Iris-setosa} \\ 5.0, 3.6, 1.4, 0.2, \mathrm{Iris-setosa} \end{bmatrix}$$

Image Processing

- It is very easy for our brains to recognize images.
- But making a computer recognize images is not an easy task, and is an active area of research in Machine Learning and Computer Science in general
- So how can an image with multiple attributes like colour be stored in a computer?
- This is achieved by storing the pixel intensities in a construct called Matrix. Then, this matrix can be processed to identify colours etc.
- So any operation which you want to perform on this image would likely use Linear Algebra and matrices at the back end.

Image Classification

The computer sees a Tensor



One-Hot Encoding

- Sometimes you work with categorical dataset and you will want to represent them with a number because that's what the computer understands
- The common process to do this is by encoding them using some techniques, one of which is one-hot encoding

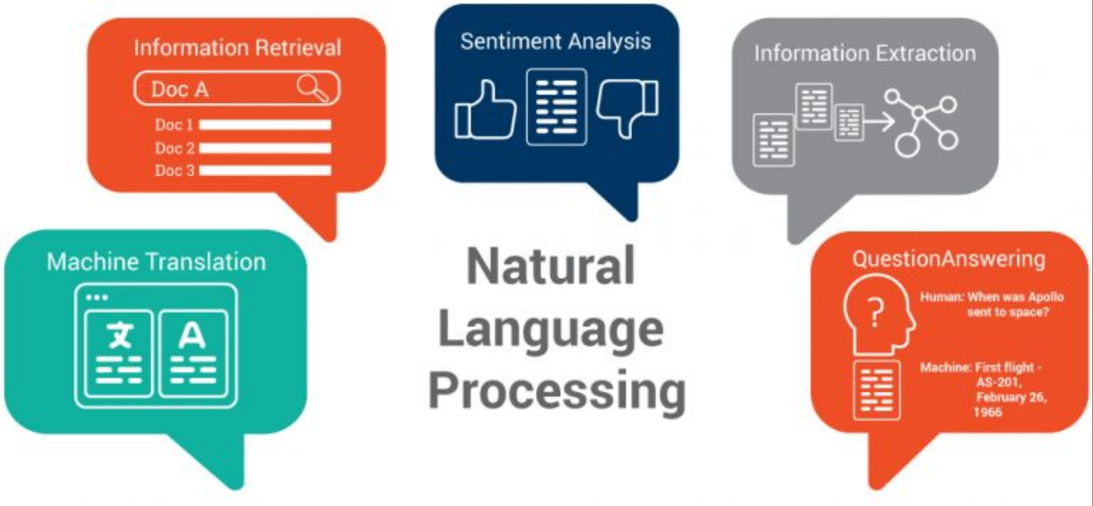
For example: Let's say we have color channel red, green and blue as defined in a vector form below

$$\text{color\_channel} = \begin{bmatrix} \mathbf{red} \\ \mathbf{green} \\ \mathbf{blue} \end{bmatrix}$$

$$\text{one\_hot\_color\_channel} = \begin{array}{c} \begin{array}{ccc} red, & green, & blue \end{array} \\ \begin{bmatrix} \mathbf{1,} & \mathbf{0,} & \mathbf{0} \\ \mathbf{0,} & \mathbf{1,} & \mathbf{0} \\ \mathbf{0,} & \mathbf{0,} & \mathbf{1} \end{bmatrix} \end{array}$$

Natural Language Processing

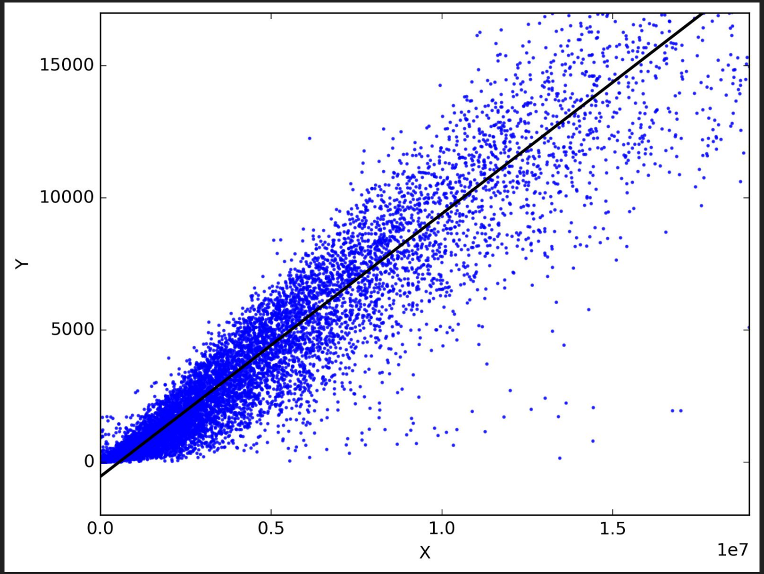
- Another active area of research in Machine Learning is dealing with text and the most common techniques employed are Bag of Words, Term Document Matrix etc.
- All these techniques in a very similar manner store counts(or something similar) of words in documents and store this frequency count in a Matrix form to perform tasks like Semantic analysis, Language translation, Language generation etc.



Linear Regression

- Linear regression is a statistical method used to describe the relationship between variables by finding the line of best fit between them.
- It is often used in machine learning for predicting numerical values in simple regression problems

$$\mathbf{y} = \mathbf{m} \cdot \mathbf{x} + \mathbf{b}$$



Recommenders Systems

- Predictive modeling problems that involve the recommendation of products are called recommender systems, a sub-field of machine learning
- Examples include the recommendation of books based on previous purchases and purchases by customers like you on Konga, and the recommendation of movies and TV shows to watch based on your viewing history and viewing history of subscribers like you on Netflix.

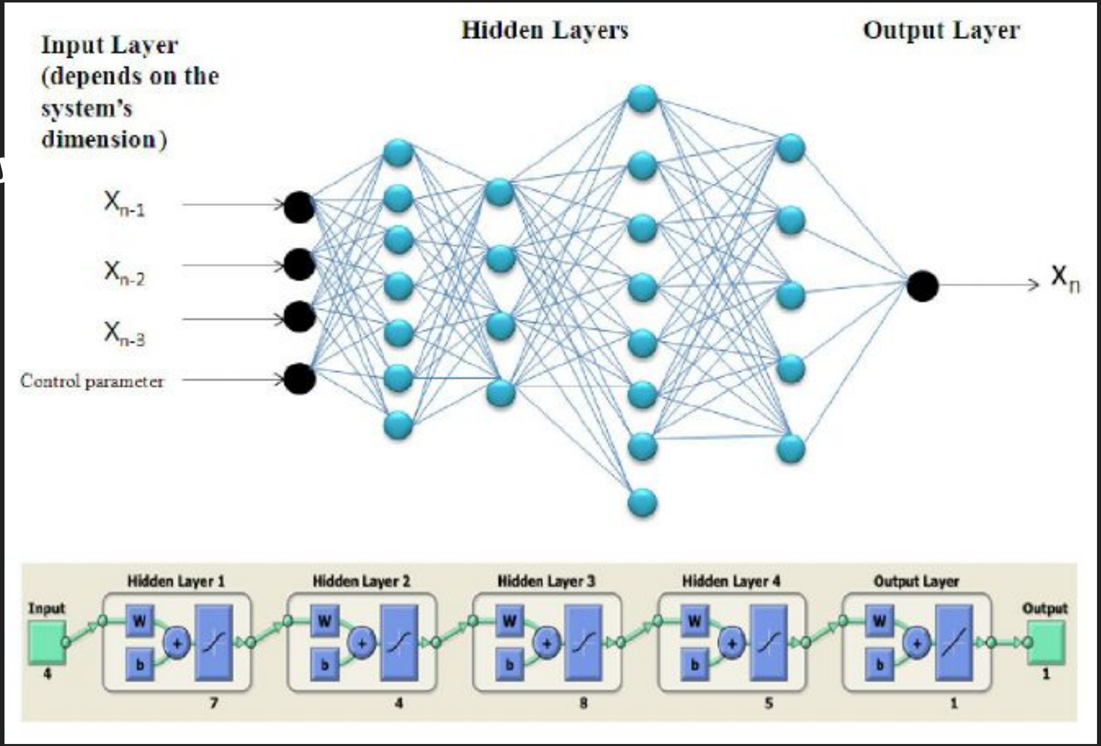
Recommenders Systems

- The development of recommender systems is primarily concerned with linear algebra methods. A simple example is in the calculation of the similarity between sparse customer behavior vectors using distance measures such as Euclidean distance or dot product

- Matrix factorization methods like the singular-value decomposition are used widely in recommender systems to distill item and user data to their essence for querying and searching and comparison

Deep Learning

- Deep learning is the recent resurgence in the use of artificial neural networks with newer methods and faster hardware that allow for the development and training of larger and deeper (more layers) networks on very large datasets



Deep Learning

- Deep learning methods are routinely achieving state-of-the-art results on a range of challenging problems such as machine translation, photo captioning, speech recognition, and much more
- At their core, the execution of neural networks involves linear algebra data structures multiplied and added together. Scaled up to multiple dimensions, deep learning methods work with vectors, matrices, and even tensors of inputs and coefficients, where a tensor is a matrix with more than two dimensions.
- Linear algebra is central to the description of deep learning methods via matrix notation to the implementation of deep learning methods such as Google's TensorFlow Python library that has the word "tensor" in its name

https://machinelearningmastery.com/examples-of-linear-algebra-in-machine-learning/


BUT WAIT!
THERE'S MORE!
ICANHASCHEEZBURGER.COM

What we didn't cover

- Linear Dependence and Span
- Norms
- Eigendecomposition
- Singular Value Decomposition
- The Moore-Penrose Pseudoinverse
- The Trace Operator
- Principal Component Analysis

References

1. Deep Learning Book by Ian Goodfellow, Yoshua Benjio and Aaron Courville
2. Deep Learning with Python by François Chollet
3. 3blue1brown - Essence of Linear Algebra
4. Machine Learning Mastery - https://machinelearningmastery.com/examples-of-linear-algebra-in-machine-learning/
5. Intro to Linear Algebra - https://speakerdeck.com/orevaoghene_ahia/introduction-to-linear-algebra
6. Wikipedia

All Images are from Google search image


THANK YOU
FOR YOUR ATTENTION