The Mathematics of Data Science and Machine Learning

Machine Learning theory is a field that intersects statistical, probabilistic, computer science and algorithmic aspects arising from learning iteratively from data and finding hidden insights which can be used to build intelligent applications. Despite the immense possibilities of Machine and Deep Learning, a thorough mathematical understanding of many of these techniques is necessary for a good grasp of the inner workings of the algorithms and getting good results.

Why Worry About the Maths?

There are many reasons why the mathematics of Machine Learning is important and I will highlight some of them below:

1. Selecting the right algorithm which includes giving considerations to accuracy, training time, model complexity, number of parameters and number of features.
2. Choosing parameter settings and validation strategies.
3. Identifying underfitting and overfitting by understanding the Bias-Variance trade-off.

   The bias-variance tradeoff is a fundamental concept in machine learning that relates to the performance of predictive models.

   Bias refers to the error that is introduced by approximating a real-world problem with a simplified model. Models with high bias tend to oversimplify the problem, leading to underfitting and poor performance on both training and test data.

   Variance, on the other hand, refers to the error that is introduced by the model's sensitivity to small fluctuations in the training data. Models with high variance tend to overfit the training data, leading to excellent performance on the training data but poor performance on the test data.

   The bias-variance trade-off occurs because reducing bias often increases variance, and reducing variance often increases bias. Therefore, the goal is to find the right balance between bias and variance that results in the best possible predictive performance on the test data.

   Some techniques for balancing the bias-variance trade-off include cross-validation, regularization, and ensemble methods such as bagging and boosting.

4. Estimating the right confidence interval and uncertainty.

   Confidence in statistics refers to the degree of certainty that a population parameter falls within a certain range of values based on a sample of data. The level of confidence is typically expressed as a percentage, and it represents the probability that the true population parameter falls within the confidence interval.

   For example, a 95% confidence interval means that there is a 95% probability that the true population parameter lies within the calculated range of values. This does not mean that there is a 95% probability that the calculated interval contains the true population parameter, but rather that if the experiment were repeated many times, 95% of the intervals would contain the true population parameter.

The level of confidence chosen for a confidence interval depends on the level of risk one is willing to accept. A higher level of confidence, such as 99%, indicates a lower level of risk, but it may also result in a wider confidence interval and lower precision in the estimate of the population parameter.

Uncertainty refers to the level of confidence we have in our estimate of a population parameter.

Suppose a researcher wants to estimate the mean weight of all adult men in a certain city. The researcher collects a random sample of 100 adult men from the city and calculates their average weight to be 175 pounds with a standard deviation of 10 pounds.

The researcher wants to estimate the population mean weight with 95% confidence. Using a t-distribution with 99 degrees of freedom (based on the sample size minus one), the critical value for a 95% confidence interval is 1.984. The standard error of the sample mean can be calculated as the sample standard deviation divided by the square root of the sample size, which gives 1 pound.

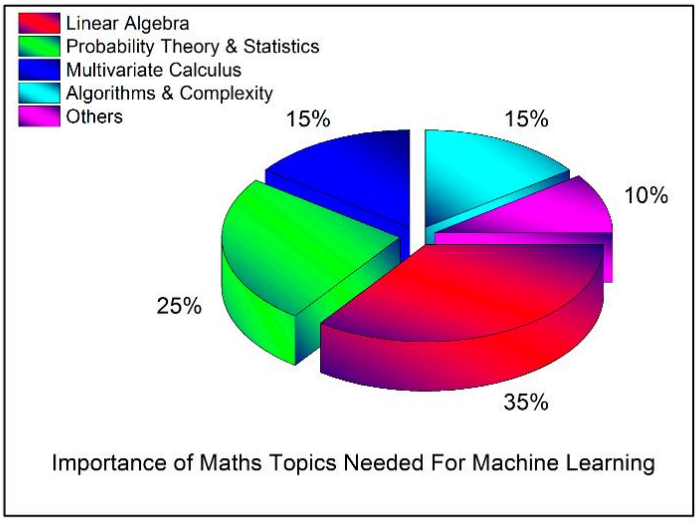
The confidence interval can then be calculated as:

$175 \pm 1.984 * 1 = [172.056, 177.944]$

This means that the researcher is 95% confident that the true mean weight of all adult men in the city lies between 172.056 pounds and 177.944 pounds.

What Level of Maths Do You Need?

The main question when trying to understand an interdisciplinary field such as Machine Learning is the amount of maths necessary and the level of maths needed to understand these techniques. The answer to this question is multidimensional and depends on the level and interest of the individual. Research in mathematical formulations and theoretical advancement of Machine Learning is ongoing and some researchers are working on more advance techniques. I will state what I believe to be the minimum level of mathematics needed to be a Machine Learning Scientist/Engineer and the importance of each mathematical concept.



Linear Algebra: Vectors, matrices, eigenvectors, and eigenvalues are important concepts in machine learning. It is also used in solving systems of linear equations and for dimensionality reduction techniques such as PCA.

Calculus: Differential calculus is useful in optimization problems, which is a central theme in machine learning. Integral calculus is used in probability theory, which is another important area of machine learning.

Probability and Statistics: Probability theory is the foundation of statistical inference, which is used to draw conclusions from data. Statistical concepts such as hypothesis testing, confidence intervals, and regression analysis are important tools in machine learning.

Multivariate Calculus: Multivariate calculus extends calculus to functions of more than one variable. It is used in optimization problems in which the goal is to find the optimal values of multiple variables simultaneously.

Optimization: Optimization is the process of finding the minimum or maximum value of a function. It is used in machine learning to find the best values of model parameters that minimize a cost function.

Information Theory: Information theory is used to quantify the amount of information in data. It is used in machine learning to measure the entropy of a probability distribution or the mutual information between two variables.

Graph Theory: Graph theory is used in machine learning to represent data as graphs or networks. It is also used in developing algorithms for clustering, classification, and anomaly detection.

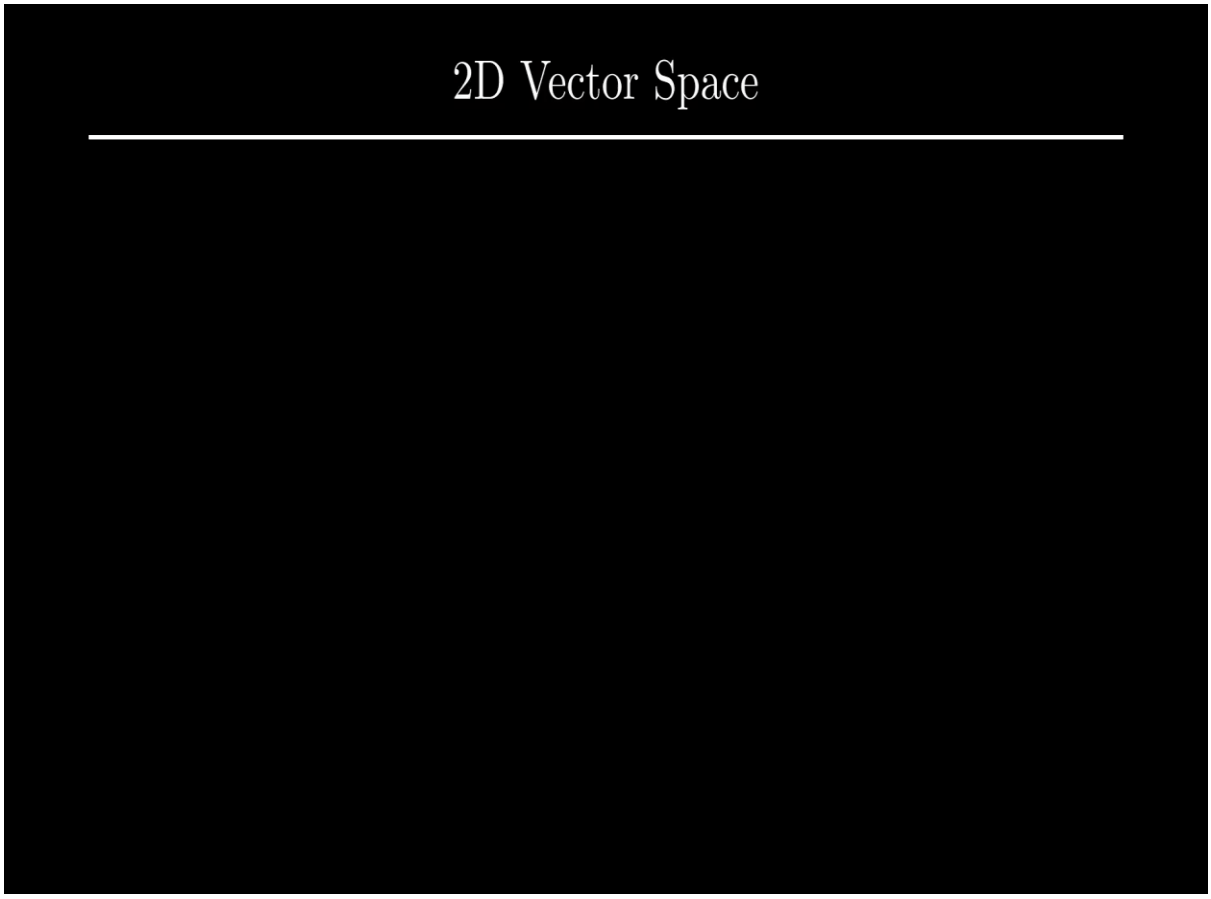
Numerical Methods: Numerical methods are used to solve mathematical problems that cannot be solved analytically. They are used in machine learning to implement optimization algorithms and solve differential equations.

Introduction

Linear algebra is a fascinating branch of mathematics that deals with the study of linear systems of equations. It explores how vectors, lines, and planes can be used to represent real-world phenomena and optimize machine learning algorithms. For instance, large datasets can be represented as matrices, where each feature and sample are expressed as a vector in the matrix. Furthermore, linear transformations can be applied to images, and dimensionality can be reduced by determining the eigenvectors and eigenvalues, a technique commonly used in facial recognition software. With these powerful tools, linear algebra plays a crucial role in ML and is a subject worth exploring in greater detail.

The building blocks of linear algebra are vectors, mathematical objects that describe both the direction and magnitude of an object in an n-dimensional space. This space is known as a vector space, a collection of vectors that can be added together or multiplied by a scalar (a real number). There are also vector spaces that use complex numbers (i.e., non-real numbers) for transformations. For example, quantum computing relies on complex vector spaces to perform quantum algorithms. Having a strong foundation in linear algebra opens up a world of possibilities, whether it's specializing in ML or branching out into new disciplines such as quantum computing. The following gif offers a simple visual representation of a 2-dimensional vector space and two vectors within it:

We can apply most of the commonly known mathematical operations on vectors and matrices. The following animations show how vector addition looks like:



Note that each vector has an arrow pointing to a certain direction and a specific size.

Moreover, vector operations have the following properties:
Let a, b, c be vectors and m, n be scalars.

Many students are unsure of what constitutes a linear function. A common misconception is that it must result in a straight line, but that's not the case. A linear function is defined as one whose

coefficients are either constants or products of a constant. The linearity assumption limits variables, not coefficients, from having an exponent higher than 1.
For example, $f(x) = a + b + c * 4$ and $g(x) = \beta * x1 + \alpha^2 * 2x$, where a, b, c, β, and α are all real numbers, are both linear functions, while $f(x) = x^2$ is not.

The inner product

The inner (aka Dot) product is an operation used to associate pairs of vectors. It allows for a rigorous introduction of intuitive geometrical notions such as the length of a vector or the angle between two vectors. Furthermore, it provides the means to define orthogonality between vectors in a vector space. The dot product is denoted as $\langle a, b \rangle = a^T \cdot b = \sum a_i \cdot b_i$, where a and b are vectors and a^T represents the transpose of a. In other words, it is the sum of the product of each element in the vectors. Thus, it returns a scalar that contains useful information about the vectors. The dot product can be used to find vector length $|a| = \sqrt{\langle a, a \rangle}$ and the cosine of the angle between vectors $\cos(\theta) = \langle a,b \rangle / |a| \cdot |b|$, useful for computing similarity measures in ML and NLP.

Norms

Norms are an essential concept in linear algebra as they provide a way to measure the size or magnitude of a vector. Different norms exist, but the most general form of such a function is the LP norm (also denoted p-norm):

LP norm: $(\sum |x_i|^p)^{(1/p)}$

where p is a positive real number, a_i is the i-th element of the vector a, and the summation is taken over all elements of the vector.

The most commonly used norms are the L1 and L2 norms. The L2 norm, also known as the Euclidean distance, is defined as:

L2 norm: $(\sum|x_i|^2)^{(1/2)}$, case where p=2

The L1 norm, also known as the Manhattan distance, is defined as:

L1 norm: $\sum|x_i|$, case where p=1

The choice of norm depends on the specific problem and context and can have a significant impact on the performance of algorithms such as linear regression or K-means clustering. Norms can also be used to measure the distance between two vectors and provide a way to measure the similarity between two data points in a high-dimensional space.

Outer (tensor) product

This is our tool to create matrices from vectors. The outer product of two vectors of size *n* and m, yields a matrix of size $n \times m$.

In this updated version, I felt I needed to elaborate further on the outer product. It plays an important role across many disciplines.

It is defined as follows: The Outer Product $u \otimes v$ is equivalent to a matrix multiplication $u \cdot v^T$, where u and v are vectors (i.e., 1-dimensional matrices). If you are dealing with complex numbers, then you must apply a conjugate transpose instead of the simple(?) transpose.

Note we have vectors a and b of size *n*, and their outer (or tensor) product yields and $n \times n$ matrix. This operation is closely related to the Kronecker product, which has applications in areas such as quantum mechanics, signal processing, and image compression. The outer product

is a powerful tool for manipulating matrices and is an important concept to understand in linear algebra. Get ready to dive into matrices!

Note: For us to be able to compute the product between two vectors (or matrices), the vector on the left must have the same number of rows as the number of columns of the vector on the right. In the example provided a has n *rows and b has* n *columns after transposing it (more on this in a little bit).*

Matrix characteristics

Matrices are a collection of numbers arranged in rows and columns, and they are widely used in linear algebra to represent the coefficients of linear equations. Matrices can be of higher dimensions beyond two, and such objects are often referred to as Tensors, with a rank 0 tensor being a scalar, rank 1 being a vector, and rank n being a tensor with n dimensions. The rank of a matrix refers to the maximum number of linearly independent rows/columns in that matrix and is an important concept in linear algebra. Matrices can be used for transformations, data storage, and various other applications in machine learning.

-Okay, more terminology, what do you mean by *linearly independent*? A linearly independent vector x is one that cannot be represented as a linear combination of any other vector in the matrix. Meaning there is no scalar by which we could multiply any of the vectors of the matrix such that the result is the vector x.

Check out the following example, which I hope will help clarify the definition of Rank and Linear independence:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 2 \\ 2 & 4 & 6 & 3 \end{bmatrix}, \quad \text{Rank}(\mathbf{A})=2$$

Here, matrix A is of Rank (2) because it only has two linearly independent vectors. Note that the vectors in black are all linear combinations of each other [1 2] can be multiplied by 2 to get [2 4] and so on, but there is no number by which we could multiply any of the vectors in black that will yield the vector in red [2 3].

Matrix Transpose

To transpose a matrix of size n x n, interchange row i with column i, *for all* $i:1,\ldots,n$. Transposing a matrix has many applications in linear algebra, including changing the orientation of vectors and solving systems of linear equations. The transpose of a matrix is also useful in machine learning as it can simplify certain matrix operations and make it easier to perform mathematical calculations; here is an example:

The transpose of a matrix can also be used to switch the row and column space of a matrix, which can be useful in various mathematical operations. Transposing a matrix can also help simplify certain computations, such as finding eigenvalues, the trace of a square matrix, and derivatives of the determinant. The trace, denoted tr(A), is defined to be the sum of elements on the main diagonal. Furthermore, it is related to the derivative of the determinant, useful for proving statements about matrix algebra.

Determinant

The determinant is a crucial mathematical tool in matrix algebra that has a wide range of applications, including in the field of machine learning. It is a scalar value that can be calculated for a square matrix (a matrix with the same number of rows and columns). The determinant has a number of interesting properties, including the fact that it can be used to determine the invertibility of a matrix and to calculate the area or volume of certain geometric objects.

One important application of the determinant in machine learning is in the calculation of the covariance matrix. The covariance matrix is a crucial component in many machine learning algorithms, such as principal component analysis (PCA) and linear discriminant analysis (LDA), that are used for dimensionality reduction and classification, respectively. The determinant of the covariance matrix provides information about the spread of the data and can be used to evaluate the importance of each principal component in PCA or each discriminant in LDA.

It is computed as follows:

The determinant is only defined for square matrices and, as you can see in the picture, it is calculated by cross multiplying element *a* (2) times *b* (0) minus *c(-1*) times 1*(1).* It is very simple to compute for a 2 × 2 matrix but it grows in complexity as the dimensions of the matrix increase.

Eigen-stuff and the Characteristic Equation

Eigenvectors and eigenvalues are incredibly important concepts in linear algebra and are used extensively in many areas of mathematics and science, including machine learning.

Eigenvectors are used to represent the directions along which a linear transformation acts, while eigenvalues give us information about how much the linear transformation stretches or compresses those directions. You can think of Eigenvectors as the pillars of a matrix; they contain fundamental information about that specific matrix. No surprisingly, they are commonly used in dimensionality reduction algorithms such as Principal Component Analysis (PCA) because, in a way, eigenvectors contain a summary of the information encoded in the matrix.

Cool, but how do we find the eigenvectors of a matrix?

Well, it is done in three steps:
Find the eigenvalue,
Plug it back in, and
Solve the system of equations.

The eigenvalue is a scalar value, denoted by λ, used to stretch or compress eigenvectors that satisfy the equation $A \cdot v = \lambda \cdot v$, where v is the eigenvector, we are trying to find. To find these scalar values we must solve a particular equation called the characteristic equation, denoted as $\det(A - \lambda \cdot I) = 0$, where A is our matrix, λ the eigenvalue and I is the identity matrix. Here is how we solve the characteristic equation:

Eigenvalue Example

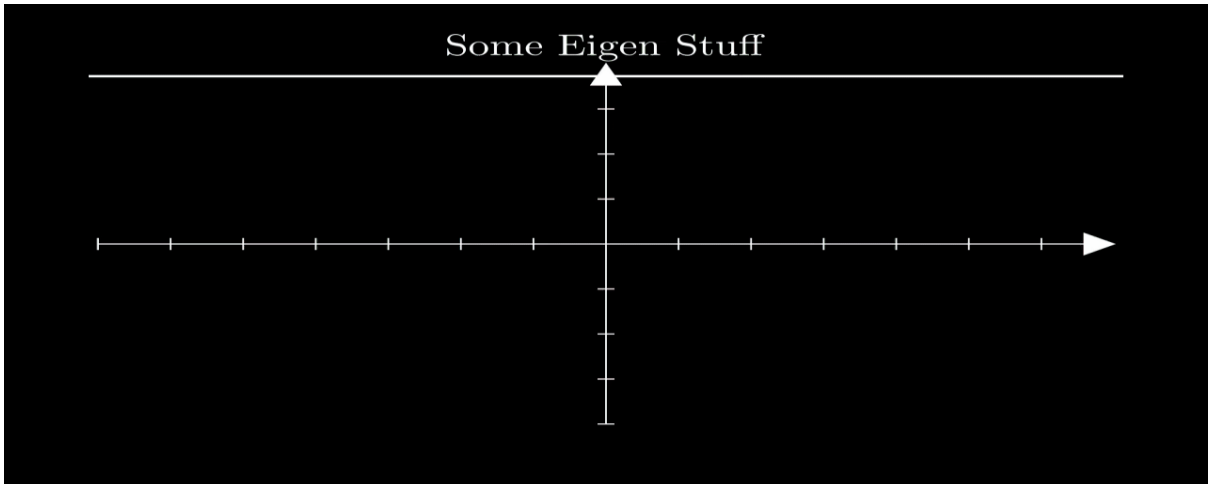
- Characteristic matrix

$$A - \lambda I = \begin{bmatrix} 1 & 2 \\ 3 & -4 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1-\lambda & 2 \\ 3 & -4-\lambda \end{bmatrix}$$

- Characteristic equation

$$\left|A - \lambda I\right| = (1-\lambda)(-4-\lambda) - (2)(3) = \lambda^2 + 3\lambda - 10 = 0$$

- Eigenvalues: $\lambda_1 = -5$, $\lambda_2 = 2$



Note how that the yellow and green vectors only change size as they are transformed, while the red vector changes both size and direction. The quantity by which the yellow and green vectors stretch or compress is determined by the *eigenvalue* and the direction is given by the *eigenvector* itself.

This concept of eigenvectors and eigenvalues is useful in many areas of mathematics, including finding the diagonalization of matrices, solving differential equations, and much more. Additionally, they are a key ingredient in the development of the Principal Component Analysis (PCA) algorithm, which is widely used in machine learning for data compression and visualization.

Vector V can be defined as an element of n dimensional coordinate system i.e V∈R^n.

So for 2D plane [x,y] is general form of vector, for 3D it's [x,y,z] as so on for dimensions.

Transpose of vector(V^T) = vector is represented in column form and it's transpose is the row representation of same.

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} x_1 & x_2 & \dots & x_m \end{bmatrix}.$$

vector transpose

A matrix M can be defined as an element M∈R^(mxn) i.e m collection of n dimensional vectors taken as row of matrix having dimension mxn.

$$M = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1m} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2m} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nm} \end{pmatrix}$$

matrix of dim mxn

Transpose of matrix: -

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{bmatrix}$$

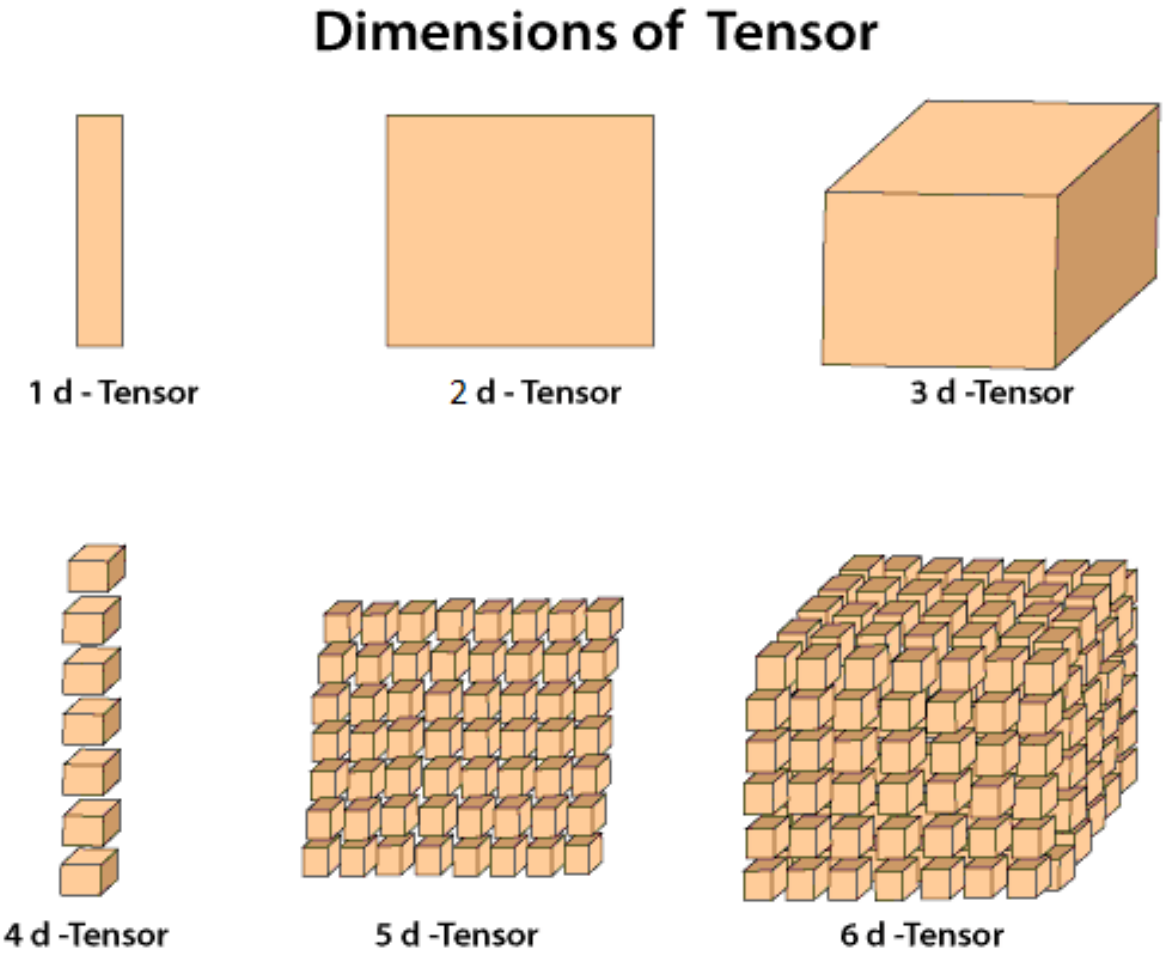
matrix transpose

We can consider vector as a single dimensional array and matrix as an 2 dimensional array.

In machine learning we may need multidimensional array as form of data to perform computation and operation which we refer as tensor.
Tensor is the generalization of vectors and matrices mostly understood as multidimensional array.
Vector is 1st order or single dimensional tensor while matrix is 2D tensor.

elements in linear algebra

We will now study properties and operations of 1D(vector) and 2D(matrix) tensor and that can be generalized to all higher dimensions.



multi-dimensional tensor

Operations on vectors

operations with scalar: - any operation with scalar is performed with every element of vector.

$$(cd)A = c(dA)$$

$$c(A + B) = cA + cB$$

$$(c + d)A = cA + dA$$

Addition and Subtraction: - vectors of same dimension are added and subtracted element-wise.

$$\begin{bmatrix}1\\2\\3\end{bmatrix} + \begin{bmatrix}4\\5\\6\end{bmatrix} = \begin{bmatrix}5\\7\\9\end{bmatrix}$$

Add two vectors:
Element wise addition

Dot product/Inner product :- dot product and inner product are two different things in mathematics but will be same in our context.It is the summation of element-wise product of two vectors of same dimension.

$$\begin{bmatrix}A_x & A_y & A_z\end{bmatrix}\begin{bmatrix}B_x\\B_y\\B_z\end{bmatrix} = A_xB_x + A_yB_y + A_zB_z = \vec{A}\cdot\vec{B}$$

dot product of vectors

Outer product:- element wise scaling of one vector by another that results in matrix is called outer product

$$\mathbf{u} \otimes \mathbf{v} = \mathbf{u}\mathbf{v}^{\mathsf{T}} = \begin{bmatrix}u_1\\u_2\\u_3\\u_4\end{bmatrix}\begin{bmatrix}v_1 & v_2 & v_3\end{bmatrix} = \begin{bmatrix}u_1v_1 & u_1v_2 & u_1v_3\\u_2v_1 & u_2v_2 & u_2v_3\\u_3v_1 & u_3v_2 & u_3v_3\\u_4v_1 & u_4v_2 & u_4v_3\end{bmatrix}.$$

outer product of vectors

Operations on matrices

matrix matrix multiplication:- It can be regarded as inner products of every row on 1st matrix with columns of 2nd matrix.

$$c_{11} = a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} + a_{14}b_{41}$$

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \\ b_{41} & b_{42} & b_{43} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{bmatrix}$$

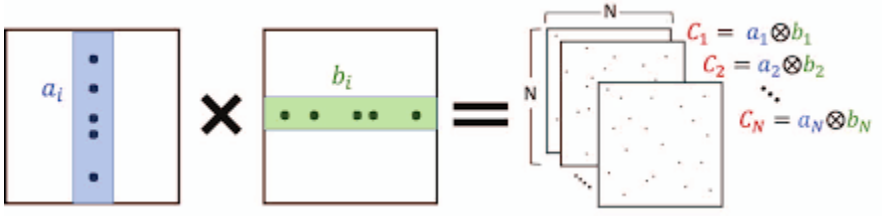
2 x 4 4 x 3 2 x 3

$$c_{22} = a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} + a_{24}b_{42}$$

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \\ b_{41} & b_{42} & b_{43} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{bmatrix}$$

inner product explanation of matrix multiplication

It can also be defined as sum of all outer products of row of matrix 1 with column of matrix 2.



outer product intuition of multiplication

Both intuition of matrix multiplication gives same result respecting the norms of mathematics.

Row Reduction/Gaussian Elimination: - Some certain operations can be produced with rows of the matrices that may change the matrix but doesn't change the interpretation of data or equation represented. These techniques are generally used to solve system of linear equations and inverse of matrix.

Matrix Row Operations

1. Two rows of a matrix may be interchanged.
2. The elements in any row may be multiplied by a nonzero number. ($3R_2$)
3. Multiply a row by a non-zero number and add to another row (Example: $2R_1 + R_3 \rightarrow R_3$)

$$\left[\begin{array}{ccc|c} 3 & 18 & -12 & 21 \\ 1 & 2 & -3 & 5 \\ -2 & -3 & 4 & -6 \end{array}\right]$$

Perform row operations:

Interchange: $R_1 \leftarrow\rightarrow R_2$

$$\left[\begin{array}{ccc|c} 1 & 2 & -3 & 5 \\ 3 & 18 & -12 & 21 \\ -2 & -3 & 4 & -6 \end{array}\right]$$

$3R_1$

$$\left[\begin{array}{ccc|c} 9 & 54 & -36 & 63 \\ 1 & 2 & -3 & 5 \\ -2 & -3 & 4 & -6 \end{array}\right]$$

$2R_2 + R_3 \rightarrow R_3$

$$\left[\begin{array}{ccc|c} 3 & 18 & -12 & 21 \\ 1 & 2 & -3 & 5 \\ 0 & 1 & -2 & 4 \end{array}\right]$$

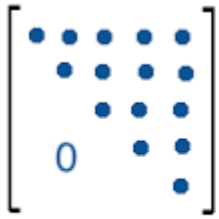
Types and representation of matrix

Diagonal Matrix = A matrix with all elements except diagonal being 0 is called diagonal matrix.

$$D = \begin{bmatrix} x_{11} & 0 & 0 & . & . & 0 \\ 0 & x_{22} & 0 & . & . & 0 \\ 0 & 0 & x_{33} & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & 0 & x_{nn} \end{bmatrix}$$

diagonal matrix

Lower Triangular and upper triangular matrix = A matrix with all elements below diagonal being 0 is called upper triangular and if all elements above diagonal is 0 is called lower triangular matrix.



Upper Triangular Matrix

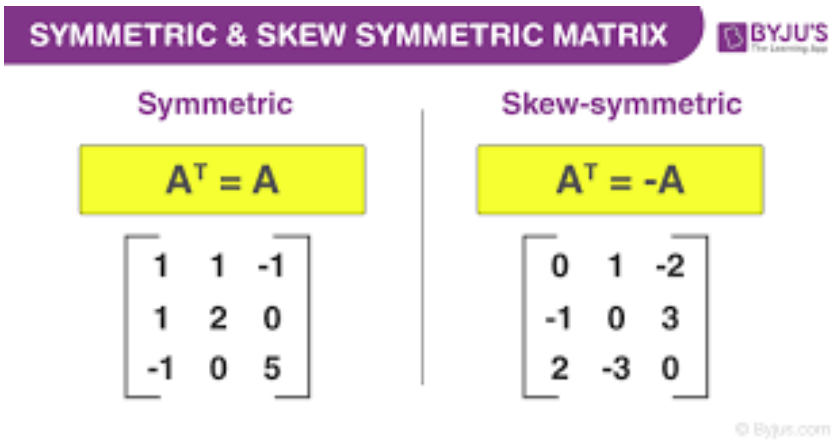
Lower Triangular Matrix

Identity matrix = Diagonal matrix with all element one having property of A.I = A is called identity matrix.

$$I_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Identity matrix

Symmetric matrix = Matrix whose transpose is same to original matrix.
Skew-symmetric matrix = matrix whose transpose is equal to negative of original matrix.



Row Echelon form = A matrix which satisfies following properties is called in row echelon form.

The first non-zero number from the left (the "leading coefficient") is always to the right of the first non-zero number in the row above.
Rows consisting of all zeros are at the bottom of the matrix.

Reduced Row Echelon form = A matrix which satisfies following properties is called in reduced row echelon form.

The first non-zero number in the first row (**the leading entry**) is the number 1.

The second row also starts with the number 1, which is further to the right than the leading entry in the first row. For every subsequent row, the number 1 must be further to the right.
The leading entry in each row must be the only non-zero number in its column.
Any zero rows are placed at the bottom of the matrix.

Echelon form

$$\begin{bmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{bmatrix}$$

Reduced echelon form

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The matrix can be converted to row echelon or reduced row echelon with the help of **row reduction method** discussed above.

System of equations	Row operations	Augmented matrix
$\begin{aligned} 2x+y-z&=8\\ -3x-y+2z&=-11\\ -2x+y+2z&=-3 \end{aligned}$		$\left[\begin{array}{ccc\|c} 2 & 1 & -1 & 8\\ -3 & -1 & 2 & -11\\ -2 & 1 & 2 & -3 \end{array}\right]$
$\begin{aligned} 2x+y-z&=8\\ \tfrac{1}{2}y+\tfrac{1}{2}z&=1\\ 2y+z&=5 \end{aligned}$	$\begin{aligned} L_2+\tfrac{3}{2}L_1&\to L_2\\ L_3+L_1&\to L_3 \end{aligned}$	$\left[\begin{array}{ccc\|c} 2 & 1 & -1 & 8\\ 0 & \frac{1}{2} & \frac{1}{2} & 1\\ 0 & 2 & 1 & 5 \end{array}\right]$
$\begin{aligned} 2x+y-z&=8\\ \tfrac{1}{2}y+\tfrac{1}{2}z&=1\\ -z&=1 \end{aligned}$	$L_3+-4L_2\to L_3$	$\left[\begin{array}{ccc\|c} 2 & 1 & -1 & 8\\ 0 & \frac{1}{2} & \frac{1}{2} & 1\\ 0 & 0 & -1 & 1 \end{array}\right]$ ← Row Echelon Form
The matrix is now in echelon form (also called triangular form)		
$\begin{aligned} 2x+y&=7\\ \tfrac{1}{2}y&=\tfrac{3}{2}\\ -z&=1 \end{aligned}$	$\begin{aligned} L_2+\tfrac{1}{2}L_3&\to L_2\\ L_1-L_3&\to L_1 \end{aligned}$	$\left[\begin{array}{ccc\|c} 2 & 1 & 0 & 7\\ 0 & \frac{1}{2} & 0 & \frac{3}{2}\\ 0 & 0 & -1 & 1 \end{array}\right]$
$\begin{aligned} 2x+y&=7\\ y&=3\\ z&=-1 \end{aligned}$	$\begin{aligned} 2L_2&\to L_2\\ -L_3&\to L_3 \end{aligned}$	$\left[\begin{array}{ccc\|c} 2 & 1 & 0 & 7\\ 0 & 1 & 0 & 3\\ 0 & 0 & 1 & -1 \end{array}\right]$
$\begin{aligned} x&=2\\ y&=3\\ z&=-1 \end{aligned}$	$\begin{aligned} L_1-L_2&\to L_1\\ \tfrac{1}{2}L_1&\to L_1 \end{aligned}$	$\left[\begin{array}{ccc\|c} 1 & 0 & 0 & 2\\ 0 & 1 & 0 & 3\\ 0 & 0 & 1 & -1 \end{array}\right]$ ← Reduced Row Echelon Form

$$\left[\begin{array}{ccc} 1 & 3 & -1\\ 0 & 1 & 7 \end{array}\right] \xrightarrow{\text{add row 2 to row 1}} \left[\begin{array}{ccc} 1 & 4 & 6\\ 0 & 1 & 7 \end{array}\right].$$ ← Row Echelon Form

$$\left[\begin{array}{ccc} 1 & 3 & -1\\ 0 & 1 & 7 \end{array}\right] \xrightarrow{\text{subtract 3 } \times \text{ (row 2) from row 1}} \left[\begin{array}{ccc} 1 & 0 & -22\\ 0 & 1 & 7 \end{array}\right].$$ ← Reduced Row Echelon Form

System of linear equations

we all know about linear equations

a0 + a1x1 + … + anxn = c

This is the system of m linear equation of n variables shown below

$$
\begin{aligned}
a_{11}x_1 &+ a_{12}x_2 &+ \cdots + a_{1n}x_n &= b_1 \\
a_{21}x_1 &+ a_{22}x_2 &+ \cdots + a_{2n}x_n &= b_2 \\
\vdots &\quad \vdots &\quad \vdots &\quad \vdots \\
a_{m1}x_1 &+ a_{m2}x_2 &+ \cdots + a_{mn}x_n &= b_m
\end{aligned}
$$

m linear equations of n variables

If all the bi's are 0, it is called homogeneous system of linear equations

$$
\begin{aligned}
a_{11}x_1 &+ a_{12}x_2 &+ \cdots + a_{1n}x_n &= 0 \\
a_{21}x_1 &+ a_{22}x_2 &+ \cdots + a_{2n}x_n &= 0 \\
\vdots &\quad \vdots &\quad \vdots &\quad \vdots \\
a_{m1}x_1 &+ a_{m2}x_2 &+ \cdots + a_{mn}x_n &= 0.
\end{aligned}
$$

homogeneous system of linear equation

Solving the system of linear equation
The system of linear equation can be solved by converting the augmented matrix formed into row echelon or row reduced echelon form and substitution.

Types of Solutions

There are three types of solutions which are possible when solving a system of linear equations

1. Independent
2. Consistent
3. Unique Solution

A row-reduced matrix has the same number of non-zero rows as variables
The left-hand side is usually the identity matrix, but not necessarily
There must be at least as many equations as variables to get an independent solution.

$$
\begin{array}{cccc}
\textbf{x} & \textbf{y} & \textbf{z} & \textbf{rhs} \\
\left[\begin{array}{ccc|c}
1 & 0 & 0 & 3 \\
0 & 1 & 0 & 1 \\
0 & 0 & 1 & 2
\end{array}\right]
\end{array}
$$

When you convert the augmented matrix back into equation form, you get x=3, y=1, and z=2.
Dependent
Consistent
infinitely many solutions

Write answer in parametric form

A row-reduced matrix has more variables than non-zero rows
There doesn't have to be a row of zeros, but there usually is.
This could also happen when there are less equations than variables.

x	y	z	rhs
1	0	3	4
0	1	-2	3
0	0	0	0

The first equation will be x + 3z = 4. Solving for x gives x = 4–3z.
The second equation will be y — 2z = 3. Solving for y gives y = 3 + 2z.

The z column is not cleared out (all zeros except for one number) so the other variables will be defined in terms of z. Therefore, z will be the parameter t and the solution is … x = 4–3t, y = 3 + 2t, z = t. since t can be any parameter, so infinite solutions or simply there are less equations than variables so infinite values possible for any one variable.

Inconsistent
No Solution

A row-reduced matrix has a row of zeros on the left side, but the right-hand side isn't zero.

x	y	z	rhs
1	0	3	4
0	1	-2	3
0	0	0	2

There is no solution here. You can write that as the null set Ø, the empty set {}, or no solution.

It indicates that one the equation is in participant to give solution, so, inconsistent.

System of equations	Row operations	Augmented matrix
$2x + y = 7$ $\frac{1}{2}y = 3/2$ $-z = 1$	$R_2 + \frac{1}{2}R_3 \to R_2$ $R_1 - R_3 \to R_1$	$\left[\begin{array}{ccc\|c} 2 & 1 & 0 & 7 \\ 0 & 1/2 & 0 & 3/2 \\ 0 & 0 & -1 & 1 \end{array}\right]$
$2x + y = 7$ $y = 3$ $z = -1$	$2R_2 \to R_2$ $-R_3 \to R_3$	$\left[\begin{array}{ccc\|c} 2 & 1 & 0 & 7 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & -1 \end{array}\right]$
$x = 2$ $y = 3$ $z = -1$	$R_1 - R_2 \to R_1$ $\frac{1}{2}R_1 \to R_1$	$\left[\begin{array}{ccc\|c} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & -1 \end{array}\right]$

The matrix is now in reduced row-echelon form. Reading this matrix tells us that the solutions for this system of equations occur when x = 2, y = 3, and z = -1.

Inverse of matrix

The inverse of a matrix A is a matrix that, when multiplied by A results in the identity. The notation for this inverse matrix is A^–1.

The matrix which have their inverse are called invertible matrices and others are called singular,i.e matrix where AA^-1 = I exist are invertible matrices.

The inverse of the matrix can be found using following steps:-

write the augmented matrix consisting of original matrix and Identity matrix with original on left and identity on left.
perform gaussian elimination such that original matrix on left is converted to identity.
During the operation, converted matrix obtained on right side from identity matrix is the reverse of original matrix.

$$[A \quad I] = \begin{bmatrix} 0 & 1 & 2 & 1 & 0 & 0 \\ 1 & 0 & 3 & 0 & 1 & 0 \\ 4 & -3 & 8 & 0 & 0 & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 3 & 0 & 1 & 0 \\ 0 & 1 & 2 & 1 & 0 & 0 \\ 0 & -3 & -4 & 0 & -4 & 1 \end{bmatrix}$$

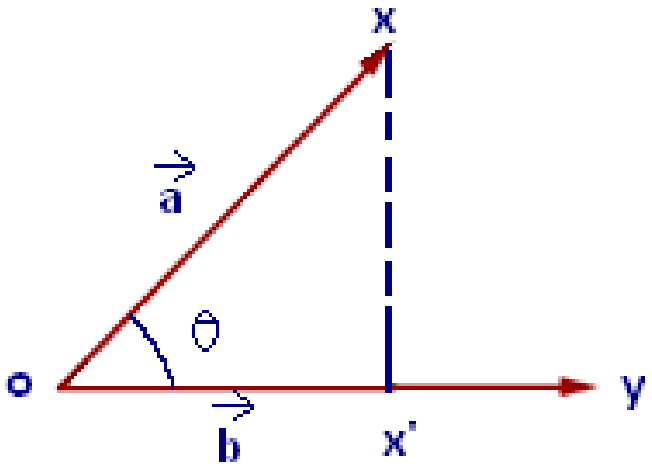
$$\sim \begin{bmatrix} 1 & 0 & 3 & 0 & 1 & 0 \\ 0 & 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 2 & 3 & -4 & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 & -9/2 & 7 & -3/2 \\ 0 & 1 & 0 & -2 & 4 & -1 \\ 0 & 0 & 1 & 3/2 & -2 & 1/2 \end{bmatrix}$$

$$A^{-1} = \begin{bmatrix} -9/2 & 7 & -3/2 \\ -2 & 4 & -1 \\ 3/2 & -2 & 1/2 \end{bmatrix}$$

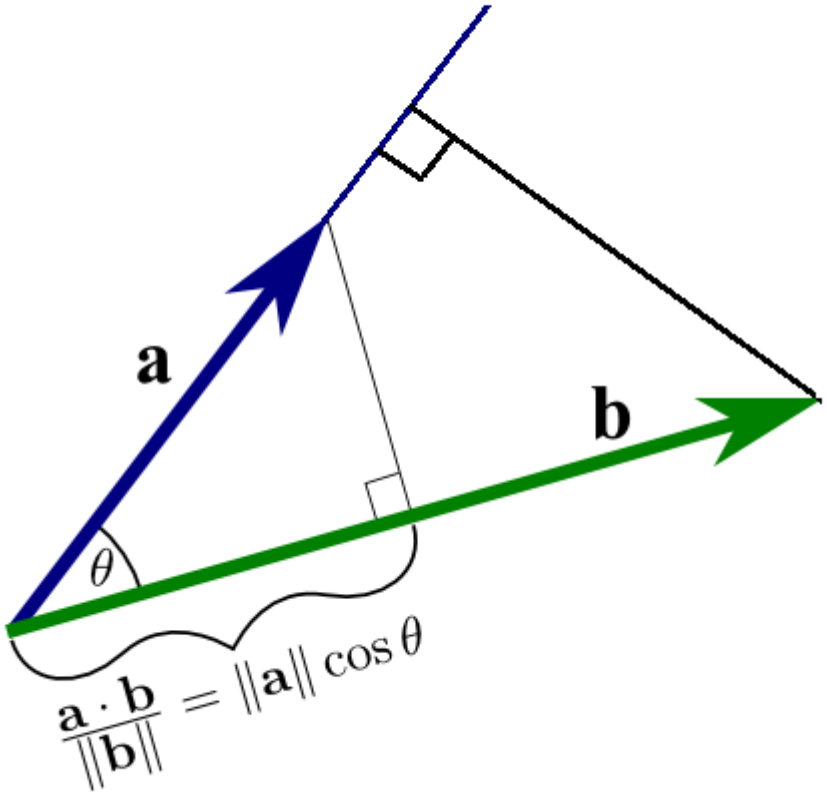
finding inverse of invertible matrix

Projection of vector

Mathematically, projection of vector a on vector b means the part of vector a projected in direction of vector b. This is very intuitive and easy to visualize.



Projection of a on b or proj(a,b) can be calculated as:- **a.b/(|a||b|)**



Projection Matrix

A projection matrix is a matrix which transforms vector from one dimension to other.

Well, this is very overlook definition, to understand it in depth, we need to know to about basis, orthogonal projection and stuffs which are easy and required very much in linear algebra but not understanding it will not make much difference in course of Machine Learning here or anywhere.

Taking an example :-

$$P = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix},$$

This matrix P transforms any vector into y=x.[multiply matrix P with vector [x y]' ,
Geometrically to the projection of any matrix A in matrix B can be given as A.Proj(B) where Proj(B) gives projection matrix for B.
So Not going much in depth in this topic, projection matrix for any matrix A can be written as :- $P=A(A'A)^{-1}A'$.

Eigen Vectors and Eigen Values

Eigen vector of a matrix A is a vector represented by a vector X such that when X is multiplied with matrix A, then the direction of the resultant matrix remains same as vector X.

It means that matrix obtained by product of matrix A and vector X,i.e matrix AX is just a scaled form of vector X.So,AX can be represented as some λX.

AX = λX, and this λ is called as eigen value for that eigen vector. i.e matrix AX is in same direction as X with it's value/magnitude scaled by λ or it's eigen values.

Lets understand it more simply,The matrix A is multiplied by a vector X to produce a new transform vector AX.(dim(A) = mxn,dim(X) = nx1,so dim(AX=mx1,hence AX a vector)

When a matrix is multiplied by a vector,there are two possibilities:-

The new transformed vector(product of matrix and vector) is just a scaled form of the original vector.i.e AX = λX.

the transformed vector has no direct scalar relationship with the original vector which we used to multiply to the matrix.

If the new transformed vector is just a scaled form of the original vector then the original vector is known to be an eigenvector of the original matrix. Vectors that have this characteristic are special vectors and they are known as eigenvectors. Eigenvectors can be used to represent a large dimensional matrix.

The value by which newly transformed vector is scaled from original vector is called eigen value and large multi-dimensional matrix form of data can be represented by eigen values as features with the importance of feature being eigen value.

Finding eigen values and eigen vectors:-

We use the general definition (AX=λX) to find eigen values and eigen vectors.
A.v = λ.v => (A-λI).v = 0, to calculate eigen values,we do |A-λI| = 0.
so we solve determinant of |A-λI| = 0 to calculate all possible eigen values for that matrix.
The no. of unique λ's obtained represent the no. of eigen vectors vi's for that matrix with them being scaled by λi's.

Determinant

Determinant is a very important concept of core linear algebra but we can understand determinant as a function which maps every square matrix with a unique no. used to solve many mathematical equations and matrix systems

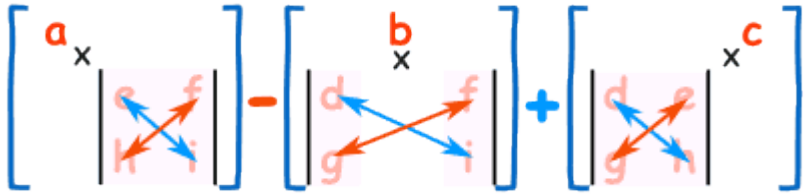
For a 1×1 Matrix
Let A = [a] be the matrix of order 1, then determinant of A is defined to be equal to a.
For a 2×2 Matrix
For a 2×2 matrix (2 rows and 2 columns):

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

determinant of A = ab - cd
For a 3×3 Matrix
For a 3×3 matrix (3 rows and 3 columns):

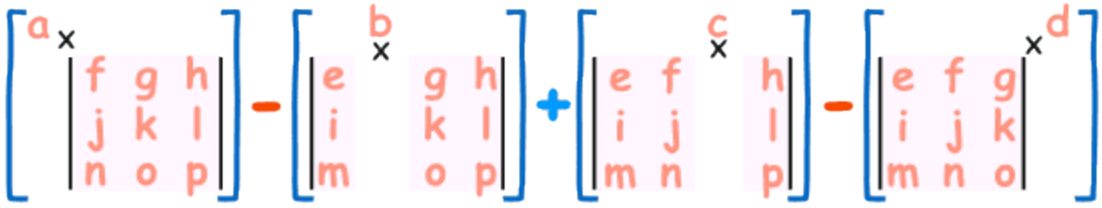
$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

The determinant is: **$|A| = a(ei - fh) - b(di - fg) + c(dh - eg)$.**



For higher dimension matrices

The pattern continues for higher order matrices with for **4x4** being:-



As a formula:

$$|A| = a \cdot \begin{vmatrix} f & g & h \\ j & k & l \\ n & o & p \end{vmatrix} - b \cdot \begin{vmatrix} e & g & h \\ i & k & l \\ m & o & p \end{vmatrix} + c \cdot \begin{vmatrix} e & f & h \\ i & j & l \\ m & n & p \end{vmatrix} - d \cdot \begin{vmatrix} e & f & g \\ i & j & k \\ m & n & o \end{vmatrix}$$

Notice **the +−+− pattern (+a… −b… +c… −d…).**

Finding Eigen vectors

The eigen values of matrix calculation was discussed as $\det|A-\lambda I| = 0$ giving all possible unique values of λ's.

After getting eigen value λ,the vector X can be calculated by solving:-

$(A-\lambda I)X = 0$

Computing λ and v

- To find the eigenvalues λ of a matrix *A*, find the roots of the *characteristic polynomial* :

$$det(A - \lambda I) = 0$$

Example: $A = \begin{bmatrix} 5 & -2 \\ 6 & -2 \end{bmatrix}$

Ax=λx

$$det\left(\begin{bmatrix} 5-\lambda & -2 \\ 6 & -2-\lambda \end{bmatrix}\right) = 0 \text{ or } \lambda^2 - 3\lambda + 2 = 0 \text{ or } \lambda_1 = 1,\ \lambda_2 = 2$$

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, x_2 = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$$

Principal Component Analysis

By definition: -

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Well this definition may not be for our purpose currently,so for followers of my lecture:-
PCA can be understood as a method of finding most important principal component vectors of matrix or feature vector of large data represented as matrix(both are same things).

Lets break it,finding most important feature of matrix…….,how to do that?

Well it's simple,find all the eigen vectors and eigen values of **square matrix obtained**,and the eigen vectors are the feature or **principal components** with their importance value reflected by respective eigen values.so,most important vector is eigen vector with highest eigen value and so on…
But

how the hell will I get **square matrix** every time?

Obviously,you will not get square matrix each time,so as standard method,All matrix are first multiplied with their transpose to form square matrix and then all methods are applied to get principal component vectors.

So let's analyze steps for PCA :-

Let's have a data of m products having n feature(n dimension) in form of matrix A of dim mxn,so A'(transpose of A) is matrix of dim nxm.
A = m x n
A' = n x m
A'.A = S (Covariance Matrix)
Calculate eigen values and vectors

we multiply A and A' to get a matrix S = A'A,dim of S is nxn(nxm x mxn = nxn)
This S is called covariance matrix,we do **eigendecomposition** of S.

Eigen decomposistion is a very simple step,for this nxn matrix,there exist n eigen values and n eigen vectors(each having length n i.e equal to length of column vector).

we sort all eigen vectors as per there eigen values in decreasing order and make a set of them i.e matrix of dim nxn(n eigen vector each having length n).

Now,suppose if we want to reduce the dim of data from n to k,so we take only k eigen vectors forming nxk matrix.

To reduce the feature dim,we multiply A(dim mxn) with this newly formed nxk matrix,giving a new matrix of dim mxk(mxn x nxk).

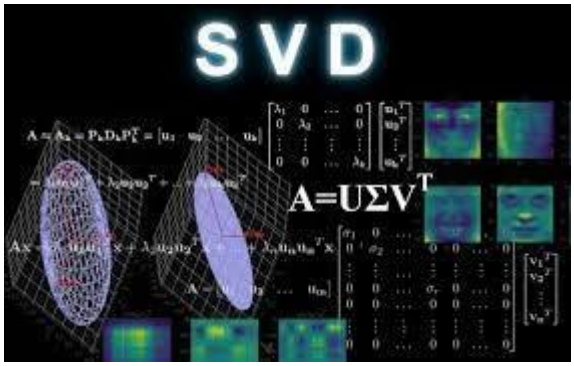
Now,we have matrix having m products with there top k features which is < n.

Remember we sorted eigen vectors as per there eigen values,that is greater the eigen value,more important the feature,so top k eigen vector(having top k eigen values) -> top k feature vector to form.

Step 3 to 5 is called **eigendecomposition** and is important concept in PCA,we will also see it's use in SVD next.

This is how we have PCA for dimensional reduction in real life machine learning/data science.

Singular value Decomposition



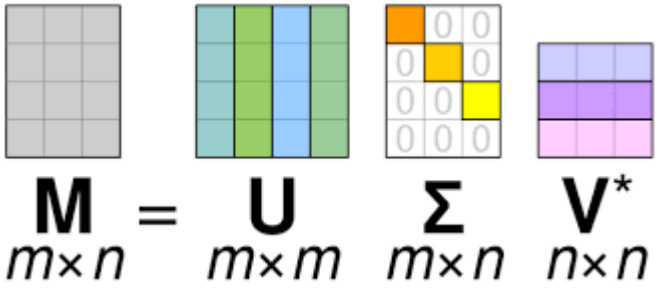
We decomposed a square matrix in terms of it's eigen values in PCA.

This decomposition of square matrix in form of it's eigen vectors is called eigendecomposition.

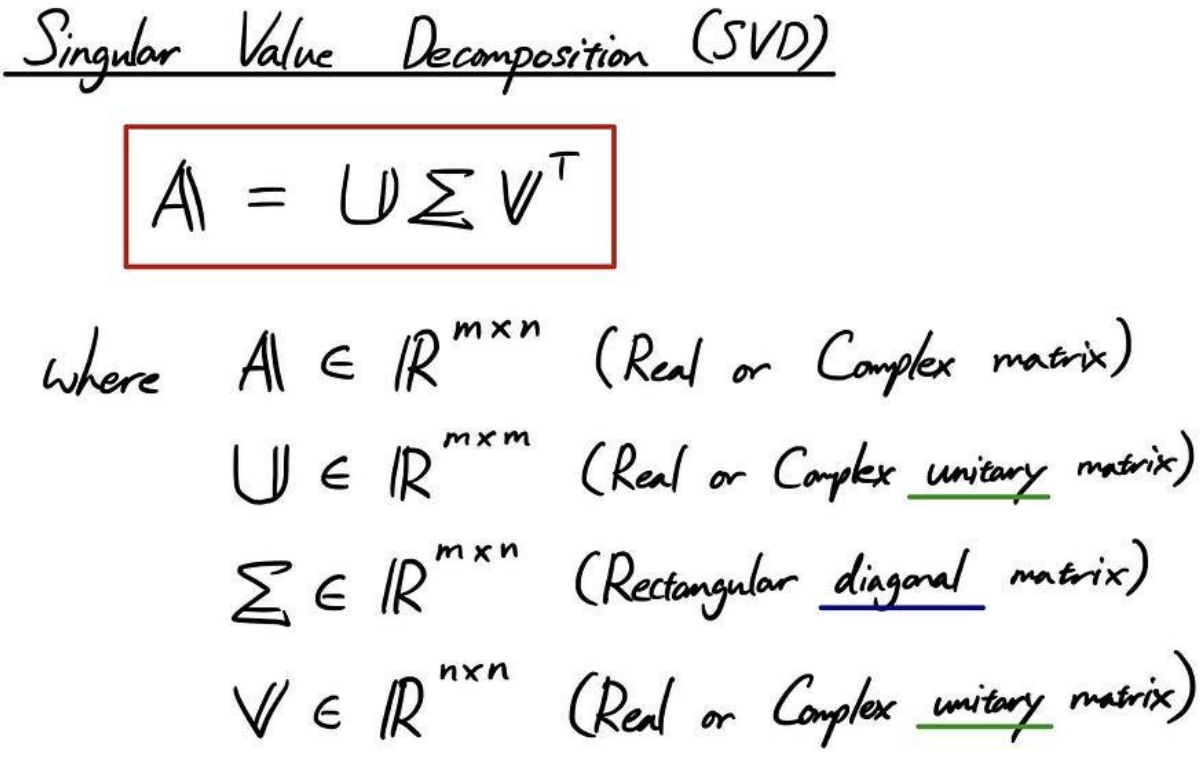
The problem with eigendecomposition is that it can be done only for square matrices,so for factorization or decomposition of non symmetric or non-square matrices,we do **singular value decomposition.**

It is very important and applicable concept having huge use in machine learning,recommendation system,data computation etc.. .
Let's understand it in mathematical manner:-

It is the decomposition of a rectangular matrix into product of two orthogonal square and a rectangular diagonal matrix.



here U and V are orthogonal matrix which means:- U'U = I and V'V = I.
So let's understand how it is done:-



Taking matrix A as

$$A = \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Taking a square matrix AA' (dim = mxm),it's eigen decomposition is done,i.e all n eigen values are taken and represented as nxn matrix(n eigen vectors each of length n as matrix dim is nxn),this matrix will be called U.

$$AA^T = \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & 4 & 0 & 0 \\ 1 & 3 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 20 & 14 & 0 & 0 \\ 14 & 10 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = W$$

$$\begin{bmatrix} 20-\lambda & 14 & 0 & 0 \\ 14 & 10-\lambda & 0 & 0 \\ 0 & 0 & -\lambda & 0 \\ 0 & 0 & 0 & -\lambda \end{bmatrix} \mathbf{x} = (W - \lambda I)\mathbf{x} = 0$$

$$U = \begin{bmatrix} 0.82 & -0.58 & 0 & 0 \\ 0.58 & 0.82 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

2. Again same step will be done with square matrix A'A(dim nxn),to again eigen decompose it to a matrix of dim nxn called as V.

$$A^T.A = \begin{bmatrix} 2 & 4 & 0 & 0 \\ 1 & 3 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$V = \begin{bmatrix} 0.40 & -0.91 \\ 0.91 & 0.40 \end{bmatrix}$$

3. The middlemost matrix is a diagonal matrix of same dimension as of A with diagonal components being square root of eigen values of AA' or A'A(both will have same eigen values).

$$S = \begin{bmatrix} 5.47 & 0 \\ 0 & 0.37 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

The SVD is used for many purpose in real life like sentiment analysis,entity recognition.