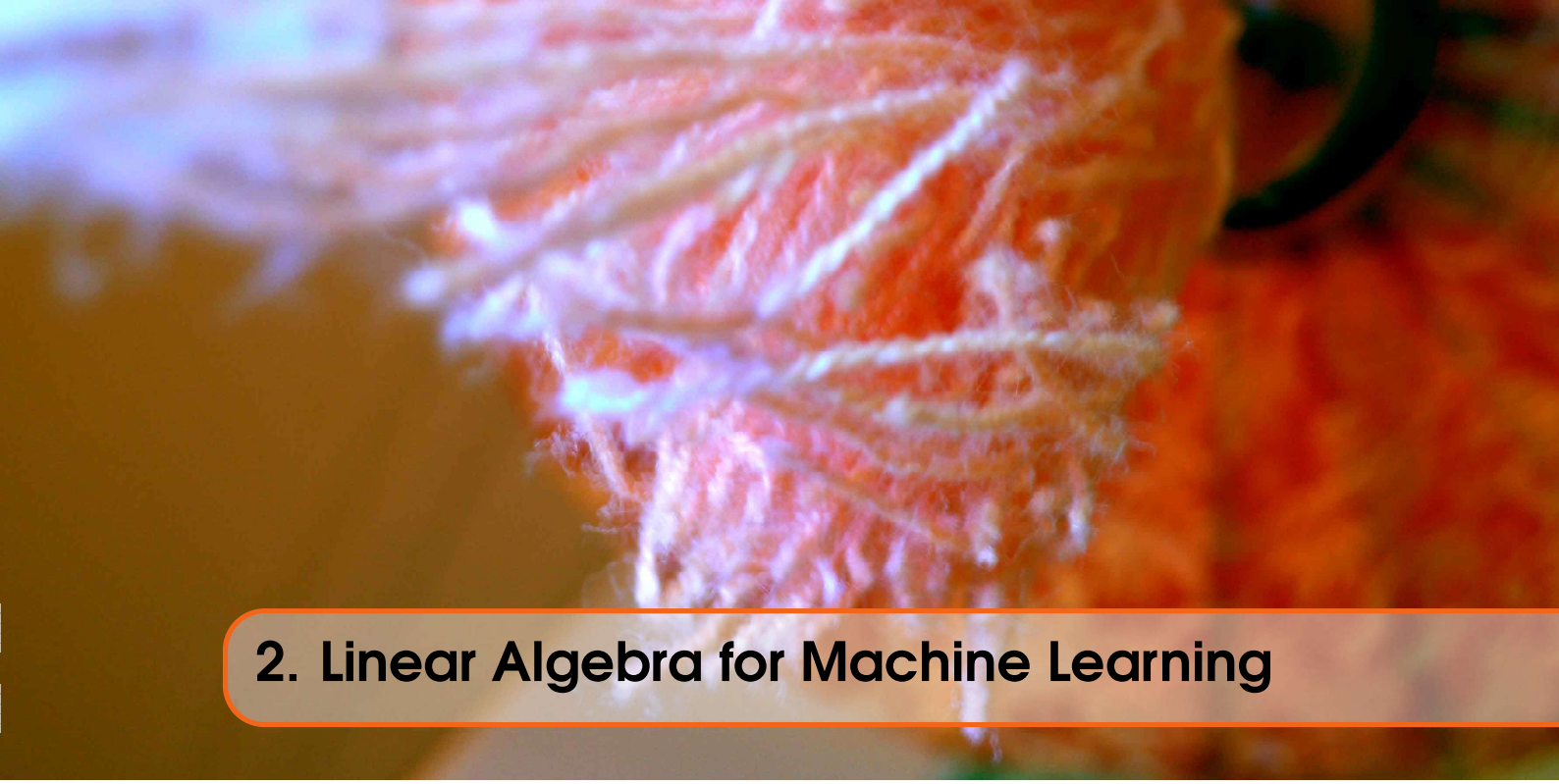

2. Linear Algebra for Machine Learning

Machine learning algorithms deal with the observation of the existing system and forecast or classify new observations for better decision-making. ML works with high dimensional and high volumes of data, observations or data are represented in vector or matrix form. Sometimes it is associated with a large number of parameters, and the set of parameters is represented as a vector or matrix. The concept of linear algebra is imperative for better understanding the theory behind the algorithm and graphical perception of the hypothesis.

2.1 Scalar and Vector

A scalar is a numeric value, indicating the magnitude of something. It is a single number having no dimension and is denoted as x, a, and b etc.
The age of a person is 40, a single value, can be considered as a scalar value. Sarika's height is 5 ft, it is also a scalar value. We can write height(h)=5. A vector is a quantity that represents two things: magnitude and direction. It is a one-dimensional array of numbers arranged in order. A vector v can be represented as $(3,4)$.
$V = [1,2]$

2.1.1 Vector and its Properties

Vectors are special objects that can be added together and multiply by scalars and produce another vector.
For any two vectors $v1=\begin{bmatrix} a1 \\ b1 \end{bmatrix}$ and $v2=\begin{bmatrix} a2 \\ b2 \end{bmatrix} \in R^2$, $v1+v2 \in R^2$ and defined as

$$v1+v2 = \begin{bmatrix} a1 \\ b1 \end{bmatrix} + \begin{bmatrix} a2 \\ b2 \end{bmatrix} = \begin{bmatrix} a1+b1 \\ a2+b2 \end{bmatrix}$$

$$\text{If } v1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \text{ and } v2 = \begin{bmatrix} 4 \\ 5 \end{bmatrix} \text{ then } v1+v2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 4 \\ 5 \end{bmatrix} = \begin{bmatrix} 5 \\ 7 \end{bmatrix}$$

Vector Addition: For any two vector u and $v \in R^n$, $u+v \in R^n$ and

$u+v=v+u$
$(u+v)+w=u+(v+w)$
Additive identity $0 \in R^n$, such that $u+0=u$
For any $u \in R^n$ there exist an $v \in R^n$ a such that $u+v=0$

Vectors are special objects that can be added together and multiply by scalars and produce another vector.
For any two vectors $v1=\begin{bmatrix} a1 \\ b1 \end{bmatrix}$ and $v2=\begin{bmatrix} a2 \\ b2 \end{bmatrix} \in R^2$, $v1+v2 \in R^2$ and defined as

$$v1+v2=\begin{bmatrix} a1 \\ b1 \end{bmatrix}+\begin{bmatrix} a2 \\ b2 \end{bmatrix}=\begin{bmatrix} a1+b1 \\ a2+b2 \end{bmatrix}$$

$$\text{If } v1=\begin{bmatrix} 1 \\ 2 \end{bmatrix} \text{ and } v2=\begin{bmatrix} 4 \\ 5 \end{bmatrix} \text{ then } v1+v2=\begin{bmatrix} 1 \\ 2 \end{bmatrix}+\begin{bmatrix} 4 \\ 5 \end{bmatrix}=\begin{bmatrix} 5 \\ 7 \end{bmatrix}$$

Vector Addition: For any two vector u and $v \in R^n$, $u+v \in R^n$ and
$u+v=v+u$
$(u+v)+w=u+(v+w)$
Additive identity $0 \in R^n$, such that $u+0=u$
For any $u \in R^n$ there exist an $v \in R^n$ a such that $u+v=0$

2.2 System of Linear Equations

We have seen that the machine learning model is presented as a function and probability distribution of features. The objective is learning involves determining the optimal value of the parameter that best fits the model or minimizes the loss function. Here system of linear equations and optimization comes into action. Many machine learning algorithms require solving linear equations and finding the model parameters. Optimization algorithms deal with maximizing and minimizing an objective defined using linear algebraic equations.
Example of an optimization problem:
An optimization problem can be written as:

$$\begin{aligned} maximimize \quad & \sum_{i=1}^{N} C_i x_i \\ Subject\ to \quad & \beta_{11}x_1+\beta_{12}x_2+\ldots+\beta_{1d}x_d \leq y_1 \\ & \beta_{21}x_1+\beta_{22}x_2+\ldots+\beta_{2d}x_d \leq y_2 \\ & \beta_{N1}x_1+\beta_{N2}x_2+\ldots+\beta_{Nd}x_d \leq y_N \\ & x_1, x_2, \ldots x_N \geq 0 \end{aligned} \tag{2.1}$$

2.3 Inner Product

Inner products facilitate the introduction of intuitive geometrical concepts, such as the length of a vector and the angle or distance between two vectors. A primary function of inner products is to determine whether vectors are orthogonal to each other. A particular type of inner product is a scalar product/dot product and is defined as

$$x^T y=\sum_{i=1}^{n} x_i y_i \tag{2.2}$$

Length and Distances: The length of a vector can be computed using the inner product.

$$||x||=\sqrt{<x,x>} \tag{2.3}$$

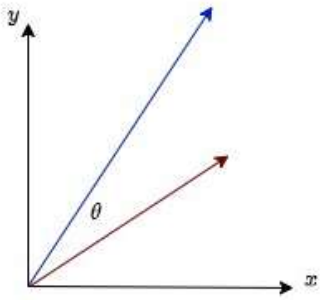

Figure 2.1: Angel between two vector.

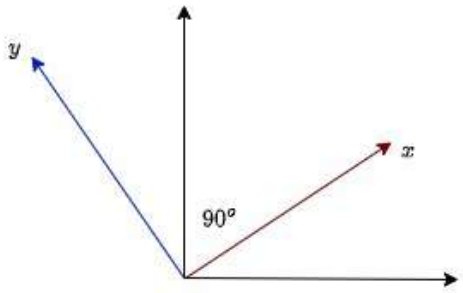

Figure 2.2: Orthogonal Vector.

Distance and Metric: Consider an inner product space $(V, <.,.>)$. Then

$$d(x,y) = ||x-y|| = \sqrt{<x-y, x-y>} \tag{2.4}$$

is called the distance between x and y for $x, y \in V$.
The mapping

$$V \times V \to d(x,y) \tag{2.5}$$

$$(x,y) \to d(x,y) \tag{2.6}$$

is called a metric.

Theorem 2.3.1 A metric d satisfies the following:
For the matrices $A, B, C \in \mathbb{R}^{m \times n}$ and α *and* $\beta \in \mathbb{R}$

d is positive definite i.e. $d(x,y) \geq 0$ for all $x, y \in V$ and $d(x,y) = 0 \Leftrightarrow x = y$

2. d is symmetric i.e. $d(x,y) = d(y,x) \forall x, y \in V$
3. Triangular inequality: $d(x,z) \leq d(x,y) + d(y,z) \forall x, y, z \in V$

Definition 2.3.1 Cauchy-Schwarz Inequality: For an inner product vector space $(V, <.,.>)$ the induced norm $||.||$ satisfies the Cauchy-Schwarz inequality

$$| < x, y > | \leq ||x||.||y|| \tag{2.7}$$

Angles: The inner product is also used to calculate the angle between two vectors. Using the Cauchy-Schwarz Inequality, we can define the angle θ between two vectors x and y.

$$-1 \leq \frac{| < x, y > |}{||x||.||y||} \leq 1 \tag{2.8}$$

There exists an unique $\theta \in [0, \pi]$ with

$$cos\theta = \frac{| < x, y > |}{||x||.||y||} \tag{2.9}$$

Orthogonal Vector: Two vectors x and y are orthogonal if and only if $< x, y >= 0$, and we can say $x \perp y$. If additionally $||x|| = 1 = ||y||$, i.e. the vectors are unit vectors, then x and y are orthonormal.

2.4 Matrices

Matrices plays an important role in machine learning. Images and all other inputs, outputs, parameters of a machine learning model is represented as a matrix. Matrix is essential to represent a system of linear equation in compact form. A matrix having m rows and n column is said to be a matrix of size/order $m \times n$ and can be written as

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \tag{2.10}$$

A matrix having 1 row and n columns is called row matrix and a matrix having n rows and 1 column is called column matrix.

Definition 2.4.1 Let $A = [a_{ij}] \in \mathbb{R}^{m \times n}$ and $B = [bij] \in \mathbb{R}^{m \times n}$ are two matrices with m rows and n columns. Then

1. the sum of two matrix A and B, denoted $A+B$, defined to be the matrix $C = [c_{ij}] \in \mathbb{R}^{m \times n}$ with $c_{ij} = a_{ij} + b_{ij} \forall i, j$.
2. the matrix addition of the matrices A and B is obtained by elementwise adding of two matrices.
3. the product of a scalar λ with the matrix A, denoted λA, defined to the matrix $C = [c_{ij}] \in \mathbb{R}^{m \times n}$ with $c_{ij} = \lambda a_{ij} \forall i, j$.
4. the scalar multiplication of a matrix is obtained by multiplying the scalar value with all the corresponding elements.

Theorem 2.4.1 Properties of Matrices For the matrices $A, B, C \in \mathbb{R}^{m \times n}$ and α *and* $\beta \in \mathbb{R}$

1. $A+B = B+A$
2. $(A+B)+C = A+(B+C)$
3. $\alpha(A+B) = \alpha A + \alpha B$
4. $(\alpha+\beta)A = \alpha A + \beta A$
5. $\alpha\beta(A) = \alpha(\beta A)$

Consider $A = \begin{bmatrix} 2 & 4 & 9 & 10 \\ 3 & -1 & -5 & 2 \\ -1 & 6 & 3 & -2 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 2 & 3 & 14 \\ 6 & 1 & 7 & 3 \\ 1 & -7 & 3 & 2 \end{bmatrix}$ then

$$A+B = \begin{bmatrix} 2+1 & 4+2 & 9+3 & 10+14 \\ 3+6 & -1+1 & -5+7 & 2+3 \\ -1+1 & 6-7 & 3+3 & -2+2 \end{bmatrix} = \begin{bmatrix} 3 & 6 & 12 & 24 \\ 9 & 0 & 2 & 5 \\ 0 & -1 & 6 & 0 \end{bmatrix}$$

Definition 2.4.2 Let $A = [a_{ij}] \in \mathbb{R}^{m \times p}$ and $B = [bij] \in \mathbb{R}^{p \times n}$ are two matrices with m rows and n columns. Then the matrix multiplication of two matrices A and B, denoted AB and defined to be the matrix $C = [c_{ij}] \in \mathbb{R}^{m \times n}$ with $c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}$ *for all* j, j
The matrix multiplication of two matrices is obtained by multiplying i^{th} row of first matrix and j^{th} column of second matrix and sum them up.

Identity Matrix: An identity(multiplicative) matrix is a square matrix defined as

$$\mathbb{I}_n = a_{ij} = \begin{cases} 1 & for\ i = j \\ 0 & for\ i \neq j \end{cases} \tag{2.11}$$

An identity matrix of size 3 is defined as $\mathbb{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

Definition 2.4.3 Let $A \in \mathbb{R}^{n \times n}$ be a square matrix. Then the matrix $B \in \mathbb{R}^{n \times n}$ is called inverse of A, if

$$AB = BA = \mathbb{I}_n \tag{2.12}$$

The inverse of the matrix A is denoted as A^{-1}.
Every matrix does not poses a inverse. The there exist an inverse A^{-1} for the matrix A, then the matrix A is called nonsingular matrix.

Definition 2.4.4 Orthogonal Matrix:
A square matrix $A \in \mathbb{R}^{n \times n}$ is an orthogonal matrix if its columns are orthonornal so that

$$AA^T = I = A^T A \tag{2.13}$$

which implies that

$$A^{-1} = A^T \tag{2.14}$$

Transformations by orthogonal matrices are special because the length of a vector x is not changed when transforming it using an orthogonal matrix A. Therefore,

$$||Ax||^2 = (Ax)^T (Ax) = x^T A^T Ax = x^T Ix = x^T x = ||x||^2 \tag{2.15}$$

When any two vectors x, y are transformed by the same orthogonal matrix A, then the angle between the two vectors calculated by the inner product is unchanged.

$$cos\theta = \frac{(Ax)^T (Ay)}{||Ax||.||Ay||} = \frac{x^T A^T Ay}{\sqrt{(Ax)^T (Ax)}\sqrt{(Ay)^T (Ay)}} = \frac{x^T A^T Ay}{\sqrt{x^T A^T Axy^T A^T Ay}} = \frac{x^T y}{x^T xy^T y} \tag{2.16}$$

$cos\theta$ gives the angle between x and y which is exactly same as the angle between Ax and Ay. It means that orthogonal matrices with $A\ A^(-1) = A^T$ preserves both direction and magnitude.