What is Machine Learning?

→ ML is a study of algorithms that improve automatically through experience.

Experience = Old data! (Training Data)

→ We look for patterns.

→ ML is a subset of AI that works by looking at prior examples.

* First ML was written by Arthur Samuel for playing checkers!

* AI ≠ ML → Helps AI

↳ Algorithms that work similarly to a cognitive level of humans.

* Machine Learning problems are of two types

① Regression Problem: 'Where does this data fit?' is the type of questions we ask here.

② Classification Problem: What type of data is this? → We ask a binary question.

* Supervised Learning: (Image Recognition) Data comes in pairs such as with one input & one output. This input data is given to the algorithm to check whether it is correct. (We teach the algorithm ourselves).

* Unsupervised Learning: (Separating two speakers from an audio clip) Here, we don't have an output rather the algorithm looks for the pattern in the data. The output is the pattern itself.

* Deep Learning (Representation Learning): We're trying to learn representation of my data.

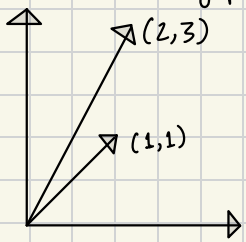
* Learning Theory: Principles! (More on that later)

* Reinforcement Learning: Agent learns optimal behaviour through interactions with its environment.

* Rank, Column Space, Null Space & Nullity

$$\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

input output



Column Span: Span of column vectors.
(Where can this matrix take us?)

$$\begin{pmatrix} 1 & 2 \\ 3 & 6 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} a+2b \\ 3a+6b \end{pmatrix}$$

$$= a+2b\begin{pmatrix} 1 \\ 3 \end{pmatrix}$$

Column Space

Or,
We are basically how its column can be mixed in different ways.

$$\underset{A}{\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}}\begin{pmatrix} a \\ b \end{pmatrix}$$

Columns: $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$ $\begin{pmatrix} 2 \\ 3 \end{pmatrix}$

These are the columns & these are linearly independent.

$$Col(A) = \mathbb{R}^2$$

Find Column Space of $\begin{pmatrix} 2 & 6 & 10 \\ 1 & 3 & 5 \end{pmatrix} = A$

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix}, 3\begin{pmatrix} 2 \\ 1 \end{pmatrix}, 5\begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

$$\therefore Col(A) = span\begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

***Rank:** Number of linearly independent columns.

*** Null Space:** Find null space of $\begin{pmatrix} 1 & 2 \\ 3 & 6 \end{pmatrix}$

$$\begin{pmatrix} 1 & 2 \\ 3 & 6 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} x+2y \\ 3(x+2y) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

now, $x+2y = 0$

$x = -2y$

So input vectors become $\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -2y \\ y \end{pmatrix}$

$\therefore$ Null Space: $\begin{pmatrix} -2 \\ 1 \end{pmatrix} = y\begin{pmatrix} -2 \\ 1 \end{pmatrix}$

Rank ⟶ Column Space

Nullity ⟶ Null Space

Rank + Nullity = # of columns.

Ex: $\begin{pmatrix} 2 & 4 & 6 \\ 1 & 2 & 3 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2(x+2y+3z) \\ x+2y+3z \end{pmatrix} = 0$

$\Rightarrow x + 2y + 3z = 0$

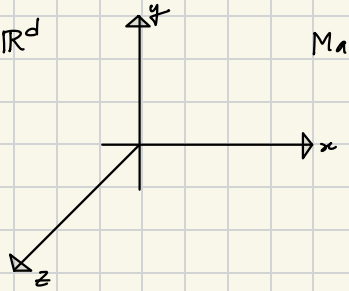
$\Rightarrow x = -2y - 3z$ so, input vector $\begin{pmatrix} -2y-3z \\ y \\ z \end{pmatrix} = \begin{pmatrix} -2y \\ y \\ 0 \end{pmatrix} + \begin{pmatrix} -3z \\ 0 \\ z \end{pmatrix}$

$= y\begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} + z\begin{pmatrix} -3 \\ 0 \\ 1 \end{pmatrix}$

$\therefore \text{Null}(A) = \text{span}\left\{\begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -3 \\ 0 \\ 1 \end{pmatrix}\right\}$

*Prerequisite 1 → Vectors

$v \in \mathbb{R}^d$



Matrix: Grid of numbers with $m \times n$ dimensions.

Identity Matrix: $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ Only 1s in the primary diagonal.

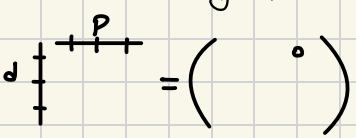
Transpose: Row ⇄ Column

Trace: $\sum_i A_{ii}$ (Sum of the primary diagonal)

Innerproduct: $x, y \in \mathbb{R}^d$ where $\sum_{i=1}^{d} x_i y_i$ → Scalar. $\langle x^T y \rangle$ {Order doesn't matter}

Cross product: We multiply two vectors & get a new vector which is perpendicular to the plane of the initial two vectors.

Outer-product: $x \in \mathbb{R}^d, y \in \mathbb{R}^d$ $xy^T \neq yx^T$ {Order Matters}



We create a matrix out of two vectors

A rank-1 matrix is made from one row vector & column vector.

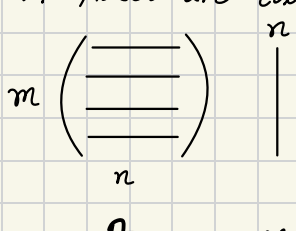
$\sum_k^2$ Rank-1 matrix = Rank-2 Matrix

$\sum_i^k$ Rank-1$_i$ = Rank-k matrix

*Increasing ranks means taking linearly independent matrices.

*Matrix-Vector Operations

→ Matrices are column vectors.



$A \in R^{m \times n}$ & $x \in R^{n \times 1}$

$Ax \in R^{m \times 1}$

$(m \times n) \times (n \times k) = (m \times k)$

Why do we need linear algebra for machine learning?

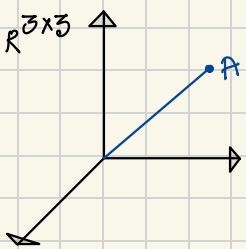
- Representing data.
- Covariance Matrices. ($S \in R^{d \times d}$)
- For calculus. (Gradients can be vectors, Hessians are matrices so is Jacobian).
- Kernel Methods.

* Geometric Interpretation

$A \in R^{m \times n}$ $\quad Ax \in R^m$

$x \in R^n$

let, $A \in R^{3 \times 3}$

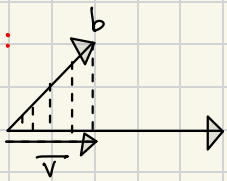


But if we take different vectors & run them through A, it's gonna give a different output.

If A is full-rank then every output will be unique for these scenario.

What happens if A is not full rank or a rank deficient matrix? It means a subspace exists & 1 to 1 mapping is still possible but within that subspace?

* Projection:



Projection on $V = (V) = \left(\frac{VV^T}{V^TV}\right) b$

*Eigen Vectors & Eigen Values

A full rank 3×3 matrix will have 3 eigenvectors & if the matrix is symmetric, those eigenvectors will be perpendicular.

Eigenvalue = Ratio of the Input & Output Vector.

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

Spectrum: Collection of eigen values.

Spectral Theorem: $A \in \mathbb{R}^{d \times d}$, $A = A^T$ have Real Valued Eigenvalues Orthonormal eigenvectors.

Hessians, Covariance Matrix, Kernel

*Quadratic Forms: $A \in \mathbb{R}^{d \times d}$ & $x \in \mathbb{R}^d$; Quadratic form: $x^T A x$

Holds for any squared matrix (we assume it is symmetric as well)

*Definiteness: $x^T A x$

- > 0 $\nexists x \neq 0$ → A is positive definite.
- ≥ 0 $\nexists x \neq 0$ → A is positive semi definite.
- < 0 " → A is negative definite.
- ≤ 0 " → A is negative semi definite.
- $>, < 0$ " → Indefinite

*Decomposing Matrices

① Singular Value Decomposition (SVD)

→ Can be done on any matrix.

$A = USV^T$ U, V → Orthonormal matrices S → Diagonal Matrix, D

② Eigen Value Decomposition (EVD)

$A = UDU^{-1}$

A	Step 1	Step 2	Step 3
SVD	V^T Rotation	Scaling Along the Axes	U Rotation
EVD	U^{-1} Rotation	Scaling Along the Axes (Rotation if Eigen values are complex)	U

*Matrix Calculus

$f: \mathbb{R} \to \mathbb{R}$ { f is a function that takes real valued inputs & outputs real values }

$f: \mathbb{R}^d \to \mathbb{R}$ { Vector inputs, scalar outputs } loss / func^n

- 1st Derivative → Gradient $\mathbb{R}^d$
- 2nd Derivative → $\mathbb{R}^{d \times d}$ Hessian (Also Symmetric)

$f: \mathbb{R}^d \to \mathbb{R}^D$ { Vector inputs, Vector Outputs }

- 1st Derivative → Jacobian $^{d \times p}$
- 2nd Derivative → Higher Order Tensors

Gradient is written like $\nabla_x f(x)$

$$\nabla_x f(x) = \begin{pmatrix} \frac{\delta f}{\delta x_1}(x) \\ \vdots \\ \frac{\delta f}{\delta x_n}(x) \end{pmatrix}$$

→ Direction of steepest ascent

Basically gradient tells us to move to a place where the function yields the max amount of value.

$$\nabla_A f(A) = \begin{pmatrix} \frac{\delta f}{\delta a_{11}}(A) & \cdots \\ \cdots & \frac{\delta}{\delta a_{min}} f(A) \end{pmatrix}$$

* Probability Theory

Sample space = Set of all outcomes. Ω

Favourable Outcome = Outcomes we want

Event $A \subseteq \Omega$

Conditional Probability, let B be any event such that $P(B) \neq 0$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$A \perp B$ if and only if $P(A \cap B) = P(A)P(B)$

$A \perp B$ if and only if $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$

→ Independence

* Random Variables: Outcomes to real values.

$\omega_0 = HHHTHTTTHTT$

* Random Variable is actually a function * we assign probabilities to events *

Suppose, we roll 2 fair dice

$$\Omega = \begin{Bmatrix} 1,4 & 2,4 & 3,4 & 4,4 \\ 1,3 & 2,3 & 3,3 & 4,3 \\ 1,2 & 2,2 & 3,2 & 4,2 \\ 1,1 & 2,1 & 3,1 & 4,1 \end{Bmatrix} \quad \mathbb{P}(\Omega) = \frac{1}{16}$$

$\exists$ the random variable yields a collection of that are Disjoint & Exhaustive. Disjoint → Events don't overlap

Exhaustive → Cover the entire sample space.

Z is a function of Ω

$Z = f(\Omega)$

$Z : \Omega \rightarrow \{2, 3, 4, 5, 6, 7, 8\}$

R.V maps entire outcome.

But, z can be any function like here $z(\omega) = \omega_1 + \omega_2$

or it could be $z(\omega) = \omega_1 \times \omega_2$

Here, ω_1 & ω_2 are randoms

Matrix Calculus

Approximate $f'(x)$ by finite differences

$$FD = \frac{f(x+\epsilon) - f(x)}{\epsilon} \qquad M\cdot D = \frac{f(x+\epsilon) - f(x-\epsilon)}{2\epsilon} \quad ; \epsilon = h$$

*Linearization

$$\frac{\delta y}{\delta x} = f'(x)$$

$$\Rightarrow \delta y = \underbrace{f'(x)}_{m}\ \underbrace{\delta x}_{(x-x_0)} \quad (y-y_0)$$

Calculus is all about some complicated curved surface that is locally linear.

∇ = finite perturbation

δ = infinitisimal.

$$* \ f(x) - f(x_0) \approx f'(x_0)(x - x_0)$$

* Matrix Calculation

① Elementwise product $x.*y$ or $x \odot y$ {Also called broadcasting}

$$\begin{pmatrix} 2 \\ 3 \end{pmatrix} .* \begin{pmatrix} 10 \\ 11 \end{pmatrix} = \begin{pmatrix} 20 \\ 33 \end{pmatrix}$$

Trace of A = Tr(A) = Sum of the diagonal

Confusion arises x^Tx or x^TAx

* $\frac{d}{d(\text{scalar})}(\text{vector}) = (\text{vector})$

Ex: $\bar{a} = \begin{bmatrix} x^2 \\ x^3 \\ x^5 \end{bmatrix} \quad \frac{\delta}{\delta x}\bar{a} = \begin{bmatrix} 2x \\ 3x^2 \\ 5x^9 \end{bmatrix}$

* $\frac{d(\text{scalar})}{d(\text{vector})} = \text{vector}$

$$f(x,y,z) = xyz^2 = x_1 x_2 x_3^2$$

$$\frac{\delta f}{\delta \bar{x}} = \begin{pmatrix} \frac{\delta f}{\delta x_1} & \frac{\delta f}{\delta x_2} & \frac{\delta f}{\delta x_3} \end{pmatrix}^T$$

* $\frac{d}{d(\text{vector})}(\text{vector}) = \text{Matrix} = \text{2nd Order Tensor}$

$$\bar{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \quad \bar{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

$$\frac{\delta \bar{a}}{\delta \bar{b}} = \begin{pmatrix} \frac{\delta a_1}{\delta b_1} & \frac{\delta a_2}{\delta b_2} \\ \frac{\delta a_1}{\delta b_1} & \frac{\delta a_2}{\delta b_2} \end{pmatrix}$$

* $\frac{\delta}{\delta \bar{x}}(\bar{x}\cdot\bar{a}) = \frac{\delta}{\delta x}(\bar{x}^T\bar{a}) = \frac{\delta}{\delta \bar{x}}(\bar{a}^T\bar{x}) = \bar{a}$

(Dot product of vector)

$$\bar{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad \bar{a} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \Rightarrow \underline{\bar{x}\cdot\bar{a}} = a_1x_1 + a_2x_2 + a_3x_3$$

$$\because \frac{\delta}{\delta \bar{x}}(\bar{a}\cdot\bar{x}) = \begin{pmatrix} \frac{\delta}{\delta x_1}(\bar{x}\cdot\bar{a}) \\ \frac{\delta}{\delta x_2}(\bar{x}\cdot\bar{a}) \\ \frac{\delta}{\delta x_3}(\bar{x}\cdot\bar{a}) \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$$

* Matrices with respect to a vector

$$\frac{\delta}{\delta x}(AB) = \frac{\delta A}{\delta x}B + A\frac{\delta B}{\delta x}$$

→ kind of like chain rule but not really.
Order of the matrices matter.

* $\bar{x}^T A \bar{x}$ → Quadratic form

$$\frac{\delta}{\delta \bar{x}}(\bar{x}^T \bar{A} \bar{x}) = (\bar{A} + \bar{A}^T)\bar{x}$$

$$= \underbrace{2Ax}$$

For symmetric matrices

$ax^2 + bxy + cy^2$

$$(x \quad y)\begin{pmatrix} a & \frac{b}{2} \\ \frac{b}{2} & c \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$

$$= (x \quad y)\begin{pmatrix} ax + \frac{by}{2} \\ \frac{bx}{2} + cy \end{pmatrix}$$

$$= ax^2 + \frac{bxy}{2} + \frac{bxy}{2} + cy^2$$

$$= ax^2 + bxy + cy^2$$

→ $\bar{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \bar{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

$$\bar{x}^T \bar{A} \bar{x} = (x_1 \quad x_2)\begin{pmatrix} x_1 a_{11} + x_2 a_{12} \\ x_1 a_{21} + x_2 a_{22} \end{pmatrix}$$

$$= x_1^2 a_{11} + x_1 x_2 a_{12} + x_1 x_2 a_{21} + x_2^2 a_{22}$$

$$\therefore \frac{\delta}{\delta \bar{x}} \bar{x}^T \bar{A} \bar{x} = \begin{pmatrix} \frac{\delta}{\delta x_1} \bar{x}^T \bar{A} \bar{x} \\ \frac{\delta}{\delta x_2} \bar{x}^T \bar{A} \bar{x} \end{pmatrix} = \begin{pmatrix} 2x_1 a_{11} + x_2 a_{12} + x_2 a_{21} + 0 \\ 0 + x_1 a_{12} + x_1 a_{21} + 2x_2 a_{22} \end{pmatrix}$$

$$= \begin{pmatrix} 2x_1 a_{11} + x_2(a_{12} + a_{21}) \\ x_1(a_{12} + a_{21}) + 2x_2 a_{22} \end{pmatrix} = \underbrace{\begin{pmatrix} 2a_{11} & a_{12} + a_{21} \\ a_{12} + a_{21} & a_{22} \end{pmatrix}}_{A + A^T}\underbrace{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}}_{\bar{x}}$$

* Maximisation = Minimisation of $-f(x)$

$x^* = \arg\min f(x)$