PAGE NO.
DATE / /

## 1.1 Introduction to Machine Learning :-

Questions : * Is AI/ML just a fad ?
- New innovation wave - Automated pattern recog. - Automated decision making.

Input + Labels → Model | Input + Model → Labels | AI ⊇ ML ⊇ Neural Networks ⊇ DL

* Why ML now ? : - Lots of data - Lots of computatn - New frameworks & alg.

When ML not useful : - Unsupervised learning - Too little data - Model is too simple

Sweet spot for ML : - Lots of structure - Explainability not critical - Prediction accuracy >> - Complex but stationary

* What is model complexity : Linear + non linear + different inputs + fn of fn.

* Other considerations : Memory Transferability Computation Speed & parallelization

* Life stages : Need identification → Data gathering → Model choice → Train valid → Deployment → Monitoring.

Directions in ML research : - Train with unlabelled data - High level labelling
- Train with fewer labelled data - One task to another
- Explaining decisions - More cautious in recognising new scenarios.

* Myth or Reality ?- M, M, M, M, R, M, M, M, M

Takeaways : - Not silver bullet
- Still useful - Data is currency of ML
- Critically design, question & then accept ML

## 1.2 ML for smart monkeys :-

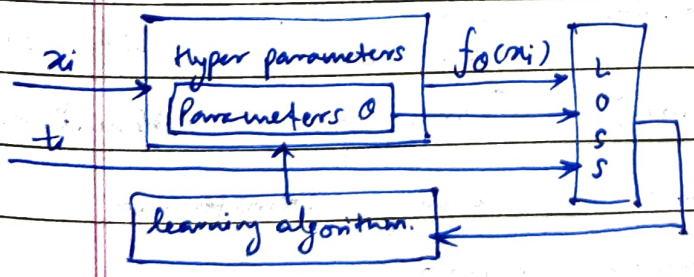
[High level recipe for a machine learning concept or model.]

Elements of a model : Input $x_i$ and function $f_\theta(x_i)$

Utility of a model : Target output $t_i$ → Bring $f_\theta(x_i) \to t_i$

Model

Aim to do this by minimizing $L(t_i, f_\theta(x_i), \theta)$ (Optimization)




Training phase.

learning is parameter ($\theta$) tuning

Hyper parameter is a design choice

ip $x_i$ → Model → $f_\theta(x_i)$ Op

(Testing phase)

Examples of hyper parameters : Random forest - No. of nodes & tree depth
Neural Network - No. of nodes in layers

₹ 14,67,882.5

PAGE No.
DATE / /

## Recipe for ML training:

(1.) Decide on type of ML problem (2.) Prepare data (3.) Shortlist ML frameworks
(4.) Prepare training, test & validation sets (5.) Train, validate, repeat (6.) Test ONCE

## (1) Broad types of ML problems:

Supervised : $x_i$, $t_i$ both given (Most advances are here)

Unsupervised : $x_i$ only given.

Supervised ML using classification / ordinal / Continuous (Ranking) (Regression or rare

Unsupervised classification / Dimension red$^n$
(clustering) (Maybe PCA?)

## (2) Preparing the data:

- Remove useless data (No variance / Not available) → May be already depends on value
  - Example (How many days spent in hospital already depends on given pt
- Reduce Redundancy (Get rid of copies of highly correlated variables)
- Handle missing data (Impute if sporadic, drop if too frequent)
- Transform variables (Convert discrete to one hot bit and normalise continuous va

## (3) Popular ML frameworks

| | Supervised | | Unsupervised | |
|---|---|---|---|---|
| Vector : (Basic data) | Logistic regression, SVM, RF, NN (classificat$^n$) | Linear regression (Regression) | Kmeans, Fuzzy Cmeans, DB SCAN (clustering) | PCA, K-PCA, LLE, ISOMAP (Dimens$^n$ reduct$^n$) |
| Series, text : | Recurrent NN, long short-term memory, Transformer, 1D CNN, Hidden Markov Models | | | |
| Images : | 2D CNN, Markov Random fields | | Probabilistic graphical models | |
| Video : | 3D CNN, CNN + LSTM, MRF | | | |

④ Preparing data for training & validation :

- Training (70%) to optimize parameters
- Validn (15%) to tune hyperparameters & check overfitting or underfitting etc.
- Testing (15%) Use only once after many rounds of validation.

PAGE No. / DATE / /

* Cross validatn : K folds and one test data → One for validatn K-1 for training → Rotate K times → Tune hyperparam

⊙ Reasons that model may still fail :

- Lack of training data diversity : Limited data confounders
- Proxy label leak during training : Data contains info of what to predict (Not available)
- Too much manual cleansing of training data
- Too little data, very complex models
- Concept drift : Kind of data changes since model was deployed (Age predn - Fashion change)

[No notes made for coding part. That is straight forward]

Math and actual theory :-

2.1) 2.2) Basic mathematics for ML (Refresher)

- Vectors & Matrices — Calculus and Convex opt. — Probability & Stats

① Scalar - Vector Operations

$a\mathbf{x}$, $\mathbf{x} + a = \mathbf{x} + a\mathbf{1}$ (abusing mathematical notn & going to prog. notn)

② Vector - vector operation

$\mathbf{x}+\mathbf{y}$ $\mathbf{x}.\mathbf{y} = \mathbf{x}^T\mathbf{y}$ $\mathbf{x} \odot \mathbf{y}$

③ Matrix - Matrix operations

$X, Y$ : $X+Y$, $XZ$, $X \odot Y$

Subspaces & Vector spaces

$\forall v_1, v_2 \in V$ $v_1 + v_2 \in V$

$cv_1 \in V$

for $W_1, W_2 \in W \subseteq V$ & $W_1 + W_2 \in W$

$cW_1 \in W$

Then W is a subspace of V.

2.3)

Transpose : $X^T = [x_{ji}]$ if $X = [x_{ij}]$ if $X \in \mathbb{R}^{N \times M}$

Determinant : Defined only on $A \in \mathbb{R}^{n \times n}$ — Square matrix invertible $\iff$ Full rank

Invertible matrix : $XX^{-1} = I$ for $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \rightarrow \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

Rank of matrix = No of Rows or Columns of a matrix that are linearly independent

Rank(X) $\leq \min(M, N)$ For square matrix — Rank deficient / Full rank (Not invertible) (invertible)

2.4.) Pseudo-inverse of a Matrix $(X \to X^+)$

$(XX^+)X = X$ [like identity but not identity]

PAGE No. / DATE / /

and $X^+ = (X^H X)^{-1} X^H$ (Conjugate transpose $= X^H$)

Square matrix

Eigen decomposn

$AV = \lambda V$. $V$ is a unit vector (we assume this) $\lambda_i$ is a scalar.

$N \times N \quad N \times 1$

Then we can get $A = Q \Lambda Q^{-1}$

$\lambda_1 > \lambda_2 > \lambda_3 \ldots$ for $\Lambda$ ( $Q = [v_1 \; v_2 \; v_3 \ldots v_N]$ )

Tensors $\alpha \in \mathbb{R}$ and $x \in \mathbb{R}^N$, $A \in \mathbb{R}^{N \times M}$ Tensor: $T \in \mathbb{R}^{M \times N \times P \times Q}$

Tensor transpose needs order of permutation of dimensions.

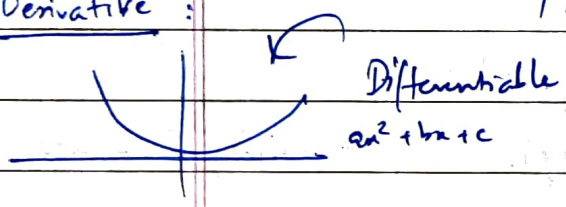
2.5) Functions $f: X \to Y$ $x \in X$ $f(x) \to y$

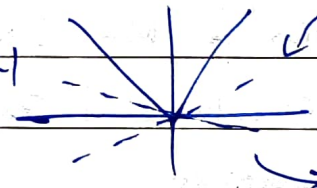
↳ May be many to one but CANNOT be one to many. (obviously)

Continuity: $x$ is continuous: $\lim_{\Delta x \to 0} f(x + \Delta x) = f(x)$ → Eg. $x^2$ – Not lipschitz cont

↳ Lipschitz continuity

Derivative: $|f(x_1) - f(x_2)| \leq K |x_1 - x_2|$ $\exists$ $K$ such that.



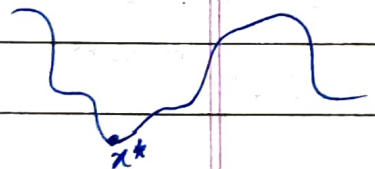

Subgradient exists
(But gradient not defined at 0)
→ Convex function / Not diff. at 0

$\lambda f(x_1) + (1-\lambda) f(x_2) \geq f(\lambda x_1 + (1-\lambda) x_2)$ $\triangleq$ Convex function.

Critical points: $f'(x) = 0$ The $x$ which satisfy this. → Maxima / Minima / Inflection

↳ Maxima: $f''(x) < 0$ | $f''(x) > 0$ : Minima | $f''(x) = 0$ Inflection point.

2.6) Gradient Descent and ascent:




Reach local minima $x^*$ We can compute $f(x)$ & $f'(x)$

If $f'(x) > 0$ – Move left / $f'(x) < 0$ – Move right

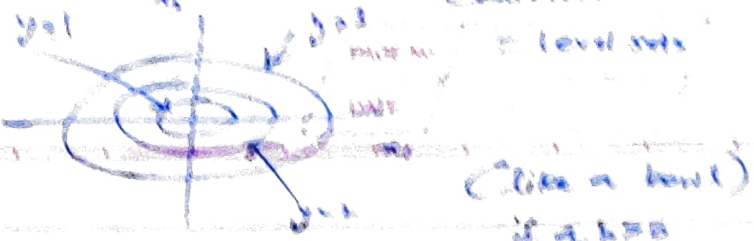
$x_1 = x_0 - \eta f'(x_0)$ or $x_n = x_{n-1} - \eta \nabla f(x_n)$

If $\eta$ too small → Too many steps needed | $\eta$ too large → Overshoots.

$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$ → If second derivative pretty deep / constant over long range.

2.4) Multivariate functions

Eg: $y = ax_1^2 + bx_2^2$ Contour plots




Gradient (Derivative of Multivariate)

Partial derivatives taken as a vector

$\nabla = [\partial y/\partial x_1 \quad \partial y/\partial x_2 \ldots \partial y/\partial x_n]$

or $f(x_1, x_2 \ldots x_n) = y$ So

$\left[\vec{x}\right] = \left[\vec{x}\right] - \eta \left[\nabla \vec{x}\right]$ → gradient descent

Newton's method for multivariate fn:

$f(x_1, x_2)$ and $\nabla f = [\partial f/\partial x_1 \quad \partial f/\partial x_2]$

The 2nd derivative : H (Hessian) = $\begin{bmatrix} \partial^2 f/\partial x_1^2 & \partial^2 f/\partial x_1 \partial x_2 \\ \partial^2 f/\partial x_2 \partial x_1 & \partial^2 f/\partial x_2^2 \end{bmatrix}$ off diagonal terms are same.

for $f(x_1, \ldots, x_n)$ $H_{ij} = \left[\partial^2 f/\partial x_i \partial x_j\right]$ — So an n×n matrix.

for eg: $f = ax_1^2 + bx_2^2 \rightarrow H = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$ - Eigen vectors tell dir of steepest ascent/descent

+ve Eigen value — There is a minima

-ve Eigen value — There is a maxima (Relate to 2nd derivative test)

Update: $\vec{x} = \vec{x} - H^{-1}\nabla f(\vec{x})$ ordinary N° (Newtons method)

Constrained Optimization Using Lagrange Multiplier :

Eg: Maximize $f(x)$, subject to $g(x) = 0$ (constraint)

How to find the tightest ? ⟹ When the contours of f are tangential to g(x).

$\therefore L(x) \stackrel{def}{=} f(x) + \lambda g(x)$

Maximize this subject to $\lambda \neq 0$ $\quad \nabla f(x) + \lambda \nabla g(x) = 0$

⟶ The derivative (gradient) of f(x) & g(x) are parallel to each other

Precise condition for having same tangent or same normal equiv.

2.5) Probability Theory (Basics)

1. Random Variable $(X = x)$ x is a value that X can take

Eg: $X \in \{0, 1\}$ (Heads or tails toss)

Probability Mass function $P(X = x) = P_X(x)$ — fn called prob. mass fn

$\sum_x P_X(x) = 1$ Depicted using a bar chart. (Discrete RV)

Common PMFs (Univariate)

① Bernoulli $x \in \{0, 1\}$ $E[X] = p$ $\left(E[X] = \sum_x x P_X(x)\right)$

$Bern(x|p) = p^x (1-p)^{1-x}$ (Nice to write in a succinct form)

$\mathbb{E}[f(x)] = \sum_x f(x) P_x(x)$ | $\text{Var}(X) = E((X - E(X))^2)$ | $p-p^2$

$\text{Var}(X) = E[X^2] - E[X]^2$ (Trick) $(q-q^2)$

PAGE NO. DATE / /

② Binomial distribution $\sim (q, N)$ [N Bernoulli iid variables]

Distribution: $P_X(x) = {}^N C_x \, p^x (1-p)^{N-x}$ $x \in \{0, \dots, N\}$

$\mathbb{E}[X] = np$ $\text{Var}(X) = np(1-p)$ (E22)

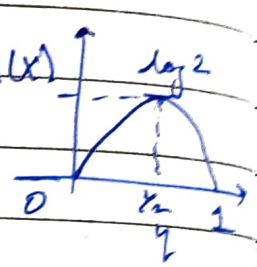
2.9.) Entropy of a Random Variable

$H(X)$ - High when variable more random / Less when less random

$\therefore H(X) \triangleq -\mathbb{E}[\log(P_X(x))]$

$= -\sum_x P(x) \log P(x)$

For Bernoulli RV: $H(X) = \log 2$ (fair coin) | $H(X) = 0$ (Completely biased)




Multiple Random Variables: (Joint / Condit$^{n}$al / Marginal)

$x$ - Dice outcome | $y \in \{0,1\} \Rightarrow$ Even | $z \in \{0,1\}$ Less than 3 (not 1,2 | 3,4,5,6)

$\therefore P(y,z) =$

| | $z=0$ | $z=1$ |
|---|---|---|
| $y=0$ | 1/6 | 2/6 |
| $y=1$ | 1/6 | 2/6 |

Joint PDF

Conditional: $\rightarrow P(y,z) / P(z) = P(y|z)$ (Obviously)

Marginal distribution: Of style variable $P(z) = \sum_y P(y,z)$

2.10.) Continuous Random Variable and pdfs (Prob. DENSITY fn) [26:18]

$f_X(x) \Rightarrow P(X = x) = 0 = f_X(x)\,dx$ $f_X(x) \geq 0 \ \forall x$

$\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ $f_X(x)$ can be $> 1$ also

$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$ (Cumulative Density function)

Examples: ① Gaussian $\sim N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]}$

$H(x) = \int_{-\infty}^{\infty} \log(f_X(x)) f_X(x)\,dx$ $\mathbb{E}[f(x)] = \int_{-\infty}^{\infty} f(x) f_X(x)\,dx$

In general.

② Beta distribution : $(f : [0,1] \to \mathbb{R}^+)$

$$\beta(x \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\, x^{a-1}(1-x)^{b-1}$$

PAGE NO. / DATE / /

③ Uniform RV ↖ Normalic$^n$ $x \in [0,1]$

— a-b $f_X(x) = 1/b-a$ for $f : [a,b] \to \mathbb{R}$

## Multivariate PDFs

$f(x_1, x_2)$ (Probability density) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2)\, dx_1\, dx_2 = 1$

or $\boxed{\int_{\infty} f(\bar{x})\, d\bar{x} = 1}$ In general.

$$P(x_2 \mid x_1 = a) = \frac{P(x_1 = a, x_2)}{P(x_1 = a)}$$ (Conditional density)

$$P(x_1) = \int_{x_2} P(x_1, x_2)\, dx_2$$ (Marginal density)

① Multivariate Gaussian Distribution : $\bar{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$ or $[y_1 \dots y_n]$

$\Sigma = [\sigma_{ij}]$ Symmetric matrix

$$f(\bar{x}) = \frac{1}{\sqrt{2\pi}^{K}\sqrt{\det(\Sigma)}}\, e^{-\frac{1}{2}\left((\bar{x}-\bar{\mu})^T \Sigma^{-1} (\bar{x}-\bar{\mu})\right)}$$

(done)