MLT Notes

✍️ Prince Kumar

- week 5 to 11(half)
- concept from pyqs
- for last time revesion

Best of luck 🤞 😊

Week-5

Dataset: $\{(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)\}$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}^d$

Goal:- learn $h: \mathbb{R}^d \to \mathbb{R}$

$h(w) = w^T x \quad \forall w \in \mathbb{R}$

Actual labels $= y_i$

Predicted labels $= w^T x_i$

Sum of Squre Error (SSE) $= \sum_{i=1}^{n} (h_w(x_i) - y_i)^2$

$= \sum_{i=1}^{n} (w^T x_i - y_i)$

\# $f(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2 = \|x^T w - y\|_2^2$

Soln:- $\boxed{w^* = (XX^T)^{-1}(Xy)}$

\# $X^T w^*$: projection of labels onto the space spanned by the features.

* Loss fn $= \sum_{i=1}^{n} (wx_i - y_i)^2$

$\nabla f = 2 \sum_{i=1}^{n} x_i (wx_i - y_i)$

① Gradient Descent

② Stochastic gradient Descent

$\boxed{w^{t+1} = w^t - \eta \nabla f(w^t)}$ ← Step Size

$\nabla f(w) = 2(XX^T)w - 2(Xy)$

\# $D = \{x_1, x_2, x_3 \cdots x_n\}$

① To Sample uniformaly the data points —

$D' = \{x_1, x_7, x\}$ $\quad t = 1, 2, \cdots T$

$w^{t+1} = w^t - \eta^T \left(2(\tilde{x}\tilde{x}^T)\right) w^t - 2\tilde{x}\tilde{y}$

② $\boxed{w_{SGD}^T = \frac{1}{T} \sum_{t=1}^{T} w^t}$

$z + w^T x$

$(y|x) \sim w^T x + \varepsilon$, $\varepsilon \sim \text{Normal}(0, \sigma^2)$

$(y|x) \sim N(w^T x, \sigma^2)$

Parameter w

$$L(w) = L(w; —) = \prod_{i=1}^{n} f_{y|x}(y_1, y_2, -- y_n, \theta)$$

after Solve —

$$\boxed{w^* = \hat{w}_{ML} = (XX^T)^{\dagger} Xy}$$

Conclusion :- Maximum likelihood estimater assuming Zero Mean Gaussian Noise is same as linear regression with SQUARED ERROR.

Expected value of $\hat{w}_{ML}$ —

$$\mathbb{E}[\|\hat{w}_{ML} - w\|^2] = \sigma^2 (\text{trace}(XX^T)^{-1})$$

Eigen val. of $(XX^T)^{-1} = \{1/\lambda_1, 1/\lambda_2 ---, 1/\lambda_3\}$

So, $\mathbb{E}[\|\hat{w}_{ML} - w\|^2] = \sigma^2 \left(\sum_{i=1}^{d} 1/\lambda_i\right)$

Mean Square Error (MSE)

$\hat{w}_{new\ (MAP)} = \underbrace{(XX^T + \lambda I)^{-1}}_{\text{Trac}(\ +\ )} (Xy)$

$$= \sum_{i=1}^{d} \frac{1}{\lambda_i + \lambda}$$

So, $\boxed{\mathbb{E}(\|\hat{w}_{new\ (MAP)} - w\|^2) = \sigma^2 \sum_{i=1}^{d} 1/\lambda_i + \lambda}$

Mean Square Error (MSE)

$\boxed{MSE(\hat{w}_{new\ (MAP)}) < MSE(\hat{w}_{ML})}$

$\lambda > 0, \lambda \in \mathbb{R}$

λ (Penalty) → Cross-Validation :

80%	20%
Training set	Test set

K-fold
It will perform validation on K-models

LOOCV Leave one out Cross V.
Special case of K-fold where,
→ models will be trained during LOOCV : $N=k$
$k=N$ → data points

Bayesian Modeling

Prior = $w \sim N(0, \gamma^2 I)$, $(y|x) \sim N(w^T x, \sigma^2)$

Let $f(w) = \frac{1}{2} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \frac{1}{2\gamma^2} \|w\|^2$

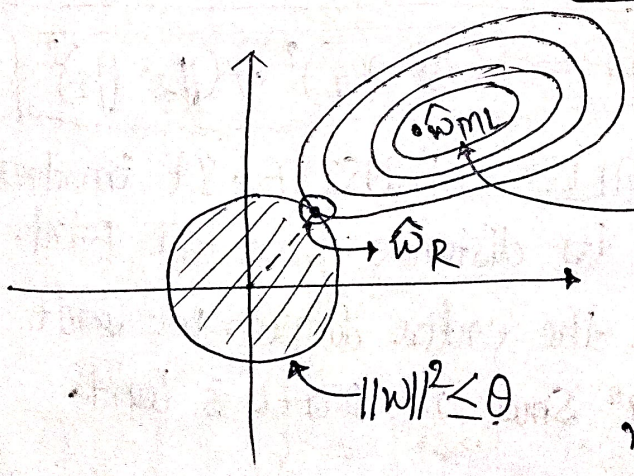
Then $\hat{w}_{MAP} = \arg\min_w f(w)$ → Solve optimization problem

$$\hat{w}_{MAP} = (XX^T + \frac{1}{\gamma} I)^{-1} xy$$

Ridge Regression

$$\hat{w}_R = \arg\min_w \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|^2$$

Loss; Regularization; Ridge penalty ($\frac{1}{\gamma^2}$)



$(w - \hat{w}_{ML})(XX^T)(w - \hat{w}_{ML}) \equiv$ Constant

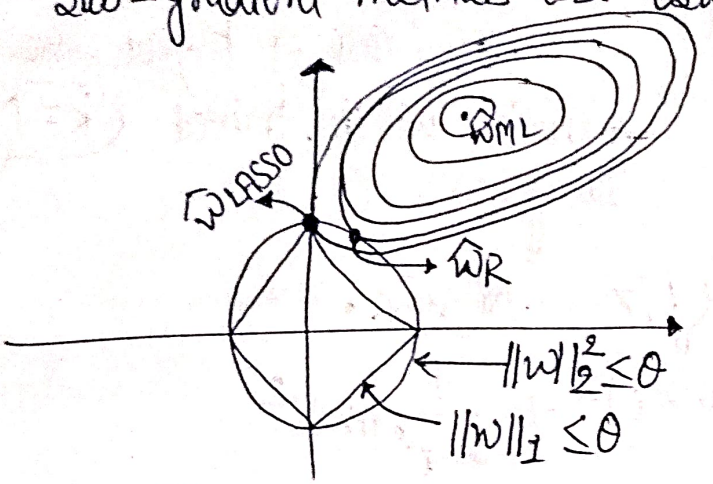
Conclusion :- Ridge Regression pushes feature values towards 0. But does not neccessary make it 0

LASSO Regression

(Least Absolute Shrinkage & Selection Operator)

→ LASSO does not have a Closed form Soln.

→ Sub-gradient methods are usually used to Solve LASSO.



* An Alternate way to regularize would then be using $\|\cdot\|_1$ norm instead of $\|\cdot\|_2^2$ norm

$\|w\|_1 = \sum_{i=1}^{n} |w_i|$

KNN → (K-Nearest Neighbours), K is hyper parameter

→ Smaller the K, Complicated the decision boundry

→ Solution : Cross-Validate for K.

Issues :- Choosing a distance function

① Manhattan distance :- $P(x_1, y_1)$ & $Q = (x_2, y_2)$

$$D = |x_1 - x_2| + |y_1 - y_2|$$

② Euclidean Distance :- $$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

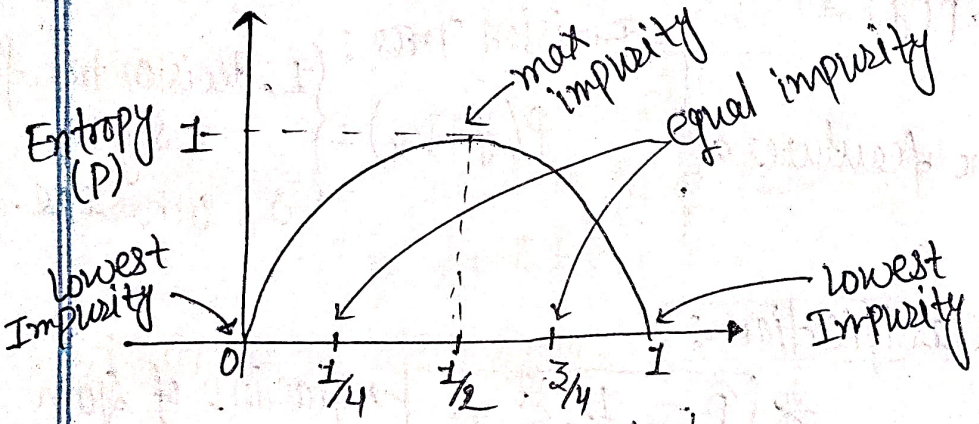
:- Prediction is COMPUTATIONALLY EXPANSIVE. It involves Sorting the datapoints acc. to its distance from test point.

:- For prediction we have to use the entire dataset. We can't throw the dataset after "learning" since no MODEL is learnt.

Decision Trees → This will Used to make the prediction

* Decisim tree Algorithm partitions the feature space (here $\mathbb{R}^2$) into horizontal & ~~by cutting the line at thresolds~~. vertical half spaces.
* D.T Algorithm partitions the space into intervals by cutting the line at thresolds. {for $\mathbb{R}^1$}
* in $\mathbb{R}^d$ → 3D Boxes (rectangular prism)

Pure Data :- If the all labels corrasponding to the features is same nature (+ve or -ve)



$$\text{Entropy}(D) = -\left[P\log_2 P + (1-P)\log_2(1-P)\right]$$

where, $P = \frac{\text{fraction of 1's}}{\text{total Data points}}$ & $(1-P)$ is fraction of another class.

D: $f_k \le \theta$ Entropy(D)

~~I.G = Entropy(D) - [∀ ∈ (Dyes)~~

yes / No

Entropy (Dyes) D_{yes} D_{No} Entropy (D_{No})

Information Gain (I.G) = Entropy(D) − [γ Entropy(Dyes) + (1−γ) Entropy(DNo)]

where, $\gamma = \frac{|D_{yes}|}{|D|}$

Types of Modelling

Generative Model

- Joint Distribution: $P(x, y)$ (features, labels)
  $= P(y/x) \cdot P(x)$
  $= P(x/y) \cdot P(y)$
- Naive Baye's
- Interested in how the features are generated

Discriminate models

Conditional dis: $P(y/x)$

- KNN: $P(y=1/x) = \begin{cases} 1, \text{ if majority of neighbours say 1} \\ 0, \text{ otherwise} \end{cases}$
- Decision Trees: $P(y=1/x) = \begin{cases} 1, \text{ decision tree for } x \text{ say 1} \\ 0, \text{ otherwise.} \end{cases}$

Naive Baye's Classification :-

$\hat{P}$

Spam=1 → $f_1, f_2 \cdots f_d$ → $P_1^1, P_2^1, \cdots P_d^1$ (d)

Non-Spam=0 → $f_1, f_2, \cdots f_d$ → $P_1^0, P_2^0, \cdots P_d^0$ (d)

* $\hat{P} = \frac{1}{n}\sum_{i=1}^{n} y_i$ → fraction of spam emails in the dataset

No. of parameter = $2d+1$

Bayes Rule :-

$$P(y^{test}=1/x^{test}) = \frac{P(x^{test}/y^{test}=1) \cdot P(y^{test}=1)}{P(x^{test})}$$

$$P(y^{test}=0/x^{test}) = \frac{P(x^{test}/y^{test}=1) \cdot P(y^{test}=0)}{P(x^{test})}$$

if $\left(\prod_{J=1}^{d}(\hat{P}_J^1)^{f_J}(1-\hat{P}_J^1)^{1-f_J}\right)\cdot\hat{P} > \left(\prod_{J=1}^{d}(\hat{P}_1^0)^{f_J}(1-\hat{P}_J^0)^{1-f_J}\right)\cdot(1-\hat{P})$

then $\hat{y}^{test}=1$, else $\hat{y}^{test}=0$

Decision fn of Naive Bays

Class: 0 & 1 pridict $y_{test}=1$, if $P(y_{test}=1/x_{test}) > P(y_{test}=0/x_{test})$

Given x_{test}

Hence, decision fn is of the form predict $y_{test}=1$ if $\frac{P(x_{test}/y_{test}=1)P(y_{test}=1)}{P(x_{test}/y_{test}=0)P(y_{test}=0)} > 1$

$\boxed{w^T x_{test} + b \geq 0}$ where, $w_i = \log\left(\frac{\hat{P}_i^1(1-\hat{P}_i^1)}{\hat{P}_i^0(1-\hat{P}_i^0)}\right)$

$\in \mathbb{R}^d$

* Conclusion :- Decision fn of Naive Bayes is LINEAR. $b = \log\left(\frac{1-\hat{P}_i^1}{1-\hat{P}_i^0}\right) + \log\left(\frac{\hat{P}}{1-\hat{P}}\right)$

Gausian Naive Bayes :- $\boxed{f[x/y] = \frac{1}{\sigma_y\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu_y}{\sigma_y^2}\right)^2}}$

Simplest Assumption

$$P(y_i = 1/x) = \begin{cases} 1 & \text{if } w^T x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Linear Separability Assumption —

$\exists\, w \in \mathbb{R}^d$ s.t $\text{Sign}(w^T x_i) = y_i \quad \forall\, i \in [n]$

PERCEPTRON :- Input : $\{(x_1, y_1), \dots (x_n, y_n)\}$, $x_i \in \mathbb{R}^d$, $y_i \in \{+1, -1\}$

Iteration

$w^0 = 0 \in \mathbb{R}^d$

Until Convergence

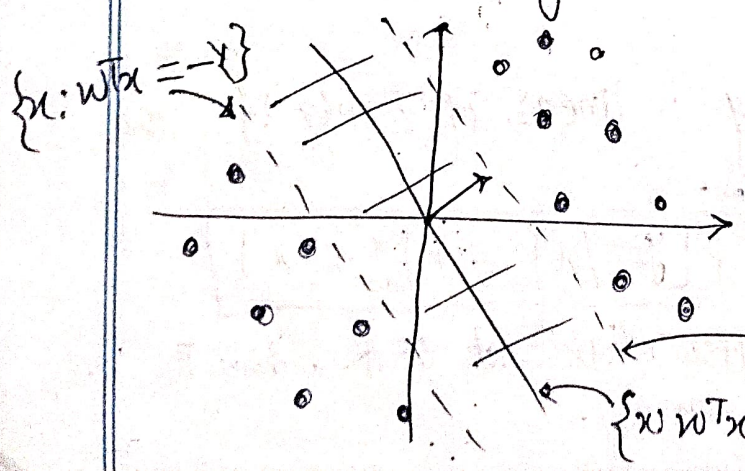
- Pick (x_i, y_i) pair from the dataset
- If $\text{Sign}(w^{tT} x_i) = y_i$
  do nothing
  ELSE $\boxed{w^{t+1} = w^t + x_i y_i}$ ← Update Rule

end

Update Rule → Update Rule pushes w in the "right" dirn for x_i if mistake $\boxed{w^{t+1} = w^t + x_i y_i}$

Assumption

① Linear Separability with γ-Margin :- A Dataset $\{(x_1, y_1), \dots \{x_n, y_n\}\}$ is linearly separable with γ margin if $\exists\, w^* \in \mathbb{R}^d$ s.t $(w^{*T} x_i) y_i \geq \gamma$ $\forall\, i$ for some $\gamma \geq 0$



② Radius Assumption

$\forall i \in D \quad \|x_i\|_2 \leq R$ for Some $R > 0$

③ without loss of generality, assume $\|w^*\| = 1$

$\gamma = w^{*T} x$, $\gamma' = \frac{\gamma}{\|w^*\|}$

Analysis of "Mistakes" of perceptron —

$\|w^{l+1}\|^2 \leq l R^2$ — ①

Mistakely

$(w^{l+1})^T w^* \geq l\gamma$ — ②

$\|w^{l+1}\|^2 \geq l^2 \gamma^2$ — ③

$w^{l+1} = w^l + x \cdot y$

for any x, y —

$(x^T y)^2 \leq \|x\|^2 \|y\|^2$

Cauchy – Schwartz

$\Rightarrow l^2\gamma^2 \leq \|w^{l+1}\|^2 \leq l R^2$

from ③ from ①

$\Rightarrow l \leq \frac{R^2}{\gamma^2}$ ← Radius Margin Bond

$l \rightarrow$ # mistakes

mistakes is bounded becoz $[\gamma > 0]$

Perceptron Convergence!

Larger the score ($z = w^T x$), more the probability of being +1.

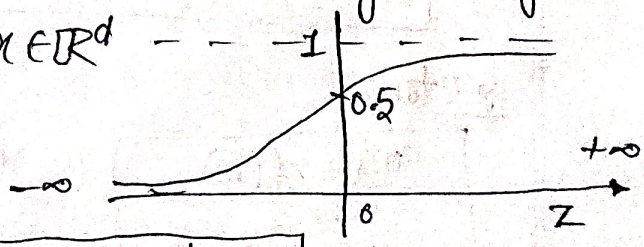
score $z \in [-\infty, +\infty]$ feature $x \in \mathbb{R}^d$

link function

$g(z) = 0.5$ if $z = 0$

$g(z) \to 1$ as $z = \infty$

$g(z) \to 0$ as $z = -\infty$



$g(z) = \frac{1}{1 + e^{-z}}$ Sigmoid/Logistic function

Model: Logistic Regression

$$P(y=1/x) = \frac{1}{1+e^{-w^T x}} = g(w^T x)$$

Gradient Update Rule:-

$$w_{t+1} = w_t + \eta_t \nabla \log L(w_t)$$

$$= w_t + \eta_t \left(\sum_{i=1}^{n} x_i \left(y_i - \frac{1}{1+e^{-w^T x}} \right) \right)$$

θ_i

$x_i \in \mathbb{R}^d$, $y_i \in \{0,1\}$, $\frac{1}{1+e^{-w^T x}} = g(w_t^T x_i)$

Kernal version

- Can argue $w^* = \sum_{i=1}^{n} \alpha_i x_i$ [formal theorem is called the Representor theorem]

Regularized Version

Cross validate Hyperparam

$$\min_w \sum_{i=1}^{n} \left[\log(1+e^{-w^T x_i}) + w^T x_i (1-y_i) + \underbrace{\frac{\lambda}{2}\|w\|^2}_{\text{Regularization}} \right]$$

Goal: To come up with a formulation that maximies "margin".

$\gamma = \frac{2}{\|w\|^2}$

$$\max_{w,\gamma} \gamma$$
$$\text{s.t } (w^T x_i) y_i \geq \gamma \quad \forall i$$
$$\|w\| = 1$$

#

$$\max_w \frac{2}{\|w\|^2} = \text{width}(w)$$
$$\text{s.t } (w^T x_i) y_i \geq 1 \quad \forall i$$

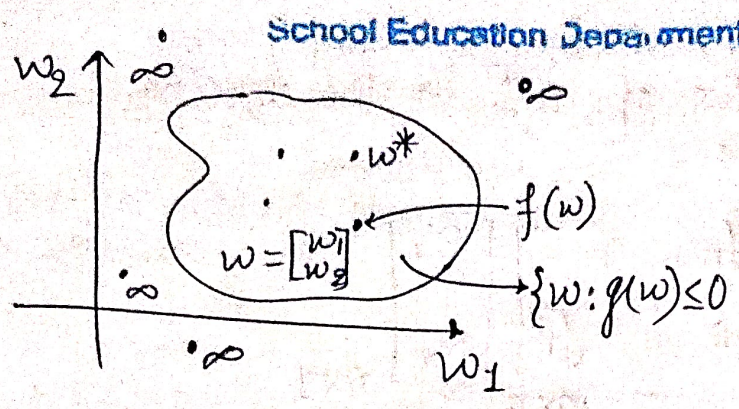
(A)

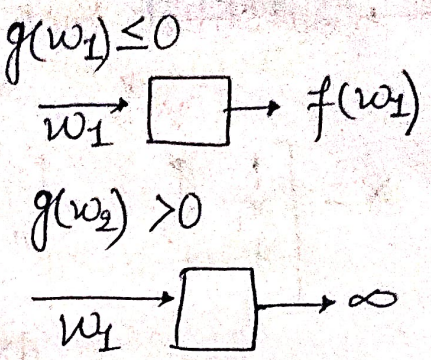
Detour

$$\min_w f(w)$$
$$\text{s.t } g(w) \leq 0$$

* $L(w, \alpha) = f(w) + \alpha g(w)$

$$\max_{\alpha \geq 0} f(w) + \alpha g(w) = \begin{cases} \infty & g(w) > 0 \\ f(w) & g(w) \leq 0 \end{cases}$$





$$\equiv \min_{w} \left[\max_{\alpha \ge 0} f(w) + \alpha g(w) \right] \quad \text{— (B)}$$

- Can we swap min & max in (B)?

Ans:- In general, No! But if f & g are "nice" function (Convex fn), then yes!

$$\min_{w} \left[\max_{\alpha \ge 0} f(w) + \alpha g(w) \right] = \max_{\alpha \ge 0} \left[\min_{w} f(w) + \alpha g(w) \right]$$ [Quadratic/Linear]

for Convex f & g.

$\|w_i'\| = \dfrac{\|w_i\|}{\gamma_i}$ *

\# $\min_{w} \frac{1}{2}\|w\|^2$ ← quadratic in w

s.t $(w^T x_i) y_i \ge 1 \quad \forall i = 1, \dots n$ ← linear in w

$\equiv \underbrace{1 - (w^T x_i) y_i \le 0}_{g_i(w)}$ # of constraints $= n$, $w \in \mathbb{R}^d$

$$L(w, \alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right)$$

$$\min_{w} \max_{\alpha \ge 0} \left[\frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right) \right] \equiv \max_{\alpha \ge 0} \min_{w} [\quad]$$

Fix Some $\alpha \ge 0$

$$\boxed{w_\alpha^* = \sum_{i=1}^{n} \alpha_i x_i y_i}$$

$\alpha \ge 0$
$y \in \{\pm 1\}$
$x \in \mathbb{R}^d$

→ Substitude back value of w^*_α in the objective

$$w^*_\alpha = XY\alpha$$

$$X = \underset{d\times n}{\begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix}} \quad \underset{n\times n}{\begin{bmatrix} y_1 & & 0 \\ & \ddots & \\ 0 & & y_n \end{bmatrix}} \quad \underset{n\times 1}{\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}}$$

$$\frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - (w^T x_i) y_i) \quad \leftarrow \text{Substitute } (w^*_\alpha = XY\alpha)$$

on Simplification

$$= \alpha^T 1 - \frac{1}{2}(XY\alpha)^T (XY\alpha)$$

where, $1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix}_{n\times}$

PRIMAL

$$\min_{w \in \mathbb{R}^d} \frac{1}{2}\|w\|^2$$
$$\text{s.t } (w^T x_i) y_i \geq 1$$

DUAL Problem

$$\max_{\alpha \geq 0} \alpha^T 1 - \frac{1}{2} \alpha^T Y^T X^T X Y \alpha$$

$X^T X \in \mathbb{R}^{n\times n}$ → Kernal K

What have we gained?

- Dual Variable dimension in $\mathbb{R}^n_+$ while primal problem dimention in $\mathbb{R}^d$
- Dual Constraints are "easier"
- More Importantly dual problem on $x^T x$ & so can be "KERNELISED"

Revisiting the langragian —

$$\min_w \underbrace{\left[\max_{\alpha \geq 0} f(w) + \alpha g(w)\right]}_{\text{Primal}} \equiv \max_{\alpha \geq 0} \underbrace{\left[\min_w f(w) + \alpha g(w)\right]}_{\text{Dual}}$$

w^* is the primal Sol^n — α^* is the dual Sol^n

$$\max_{\alpha \geq 0} f(w^*) + \alpha g(w^*) = \min_w f(w) + \alpha^* g(w)$$

After Solve :- $\boxed{\alpha^* g(w^*) \geq 0}$ ——— ①

But we already know $\alpha^* \geq 0$ & $g(w^*) \leq 0$

So, $\boxed{\alpha^* g(w^*) = 0}$ → Complementry Slackness

for multiple Constraint —

$\boxed{\alpha_i^* g_i(w^*) = 0} \; \forall i$

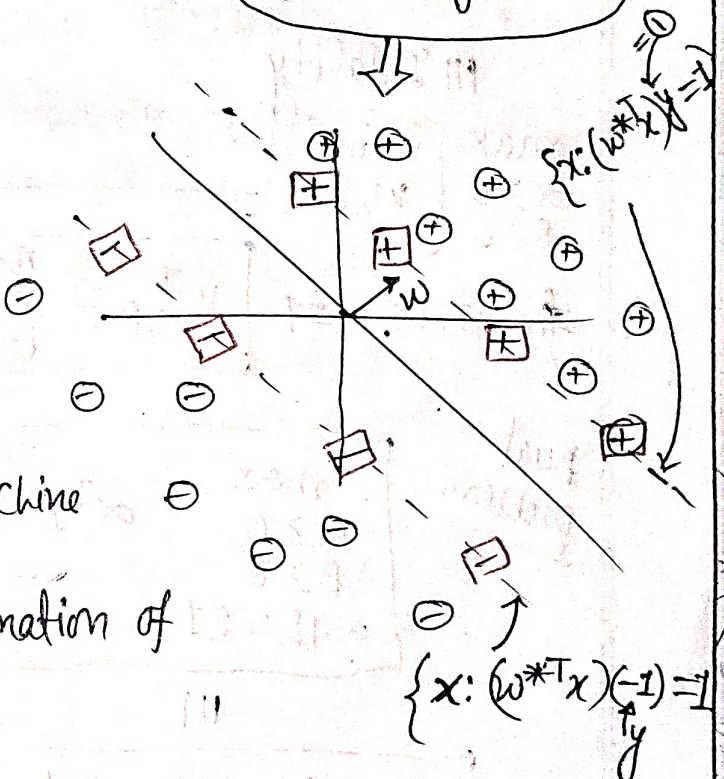
if $\alpha_i^* > 0 \xrightarrow{C.S} 1 - (w^{*T} x_i) y_i = 0 \Longrightarrow \boxed{(w^{*T} x_i) y_i = 1}$

→ Only The points that are on the "Supporting" Hyperplane can Contribute to w^*

→ These special points are called "Support Vectors".

→ ALGORITHM → Support Vector Machine (SVM)

→ w^* is a sparse linear Combination of the data points



Given x_{test} : $w^*(\phi(x_{test})) = \sum_{i=1}^{n} \alpha_i^* y_i k(x_i, x_{test})$

Soft Margin formulation → used: when data Contains outliers or, not perfectly linearly separable

primal $\begin{cases} \min\limits_{w,\varepsilon} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \varepsilon_i \\ \text{s.t } (w^T x_i) y_i + \varepsilon_i \geq 1 \;\; \forall i \;\; \& \;\; \varepsilon_i \geq 0 \end{cases}$

C : Hyperparameter ≥ 0

Small C : Margin wide (High Bias & Low Var) Error Jayda Allow Karta hai

$C=0 \Rightarrow$ Bribes don't cost
$\rightarrow w=0 \in \mathbb{R}^d$ is the Soln & $C=\infty \Rightarrow$ Linear Separable case
$\hookrightarrow$ Hard margin

$$\# L(w,\xi,\alpha,\beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1-w^T x_i y_i - \xi_i) + \sum_{i=1}^{n}\beta_i(-\xi_i)$$

DUAL Problem

$$\min_{w,\xi}\left[\max_{\substack{\alpha\geq 0\\ \beta\geq 0}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1-w^T x_i y_i - \xi_i) + \sum_{i=1}^{n}\beta_i(-\xi_i)\right]$$

!!! Duality

$$\max_{\substack{\alpha\geq 0\\ \beta\geq 0}}\left[\min_{w,\xi} \text{ ——————}\right]$$

fix $w, \alpha \Rightarrow$ $\boxed{w^*_{\alpha,\beta} = \sum_{i=1}^{n}\alpha_i x y_i}$ —— ① & $\boxed{\alpha_i+\beta_i=C}$ ——— ⑪

Dual problem

$$\boxed{\begin{array}{l}\max\limits_{\substack{\alpha\geq 0\\ \beta\geq 0\\ \alpha+\beta=C1}} \quad \alpha^T 1 - \frac{1}{2}\alpha^T y^T x^T x\, y\alpha \\ \qquad\qquad\qquad\qquad \text{no } \beta \text{ term}\end{array}}$$

!!!

$$\boxed{\max_{0\leq\alpha\leq C} \alpha^T 1 - \frac{1}{2}\alpha^T y^T x^T x\, y\alpha}$$ ← Kernalizable

if $C=0$
$\Rightarrow \alpha^* = 0 \in \mathbb{R}^n$
$\Rightarrow \boxed{w^*=0}$

& $C=0 \Rightarrow$ Hard margin

\# Complementry Slackness {about Soft margin}.

Let (w^*, ξ^*) be the primal optimal Soln
& (α^*, β^*) be the dual optimal Soln.

C.S ① $\forall i \; \alpha_i^* (1 - w^{*T} x_i y_i - \xi_i^*) = 0$

② $\forall i \cdot \beta_i^* (\xi_i^*) = 0$

Various Cases

① $\alpha_i^* = 0 \Rightarrow \beta_i^* = C \xrightarrow{C.S\,②} \boxed{\xi_i^* = 0} \xrightarrow{C.S\,①} \boxed{w^{*T} x_i y_i \geq 1}$

"w^* Classifies (x_i, y_i) Correctly"

② $\alpha_i^* \in (0, C) \Rightarrow \beta_i^* \in (0, C) \xrightarrow{C.S\,②} \boxed{\xi_i^* = 0} \xrightarrow{C.S\,①} \boxed{w^{*T} x_i y_i = 1}$

③ $\alpha_i^* = C \Rightarrow \beta_i^* = 0 \xrightarrow{C.S\,②} \boxed{\xi_i^* \geq 0} \xrightarrow{C.S\,①} \boxed{w^{*T} x_i y_i \leq 1}$

Points where either x_i is incorrectly classified by w^* or Correctly classified but with margin ≤ 1.

Let's see this from the primal point of view.

Case① $w^{*T} x_i y_i < 1 \mapsto \xi_i^* \geq 1 - w^{*T} x_i y_i$ $\quad w^* x_i y_i + \xi_i^* \geq 1$

$\Rightarrow \xi_i^* > 0 \xrightarrow{C.S\,②} \boxed{\beta_i^* = 0} \Rightarrow \boxed{\alpha_i^* = C}$

Case② $w^{*T} x_i y_i = 1 \Rightarrow \xi_i^* \geq 0 \Rightarrow \boxed{\alpha_i^* \in [0, C]}$

Case③ $w^{*T} x_i y_i \geq 1 \Rightarrow \boxed{\alpha_i^* = 0}$

Goal: Meta Classifiers or Ensemble Classifies.

Weak Learner ⟶ Strong Learner

"The classifiers whose performance is slightly better than the random classifier"

Overfitting - fit noise (High Variance)

Underfitting - missing out on str thinking it is noise. (High bias)

→ error = bias + Var

Bagging → "Averaging Redues Variance"

Extra Concept & Must Remembering point —

* $\hat{w}_{MLE}: \min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - w^T x_i)^2$

* $\hat{w}_L: \min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda \|w\|$

* $\hat{w}_R: \min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda (\|w\|)^2$

MLE for Bernolli parameter $(\hat{p}) = \frac{w}{N}$ → No. of Success (fraction of ones)
→ total no. of Sample

* $X = \begin{bmatrix} | & | \\ x_1 & x_2 \\ | & | \end{bmatrix}$ ← feature
f_1 f_2

$\Rightarrow$ Solve $X^T = \begin{bmatrix} - x_1 - \\ - x_2 - \\ \vdots \end{bmatrix}$

Kernal Regression :-

$$w^{*T} \phi(x_{test}) = \sum_{i=1}^{n} \underbrace{\alpha_i^* k(x_i, x_{test})}_{\text{Kernal fn}} \quad \text{where, } \alpha^* = \underset{\in \mathbb{R}^{n \times n}}{K^{-1}} y$$

① $(1 + x_i x_j)^2$: Polynomial kernal

② $e^{\frac{-\|x_i - x_j\|^2}{2\sigma^2}}$: Gaussian kernal

* $error = \sum (h(x_i) - y_i)^2$

* $\boxed{CV_{LOOCV} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$ $\quad \boxed{CV_{k\text{-fold}} = \frac{1}{k} \sum_{i=1}^{n} (MSE)}$

* MSE (mean Square Error) $= \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x)^2$ & MAE (Mean Absolute Error) $= \frac{1}{n} \|y_i - w^T x\|$

$*\underbrace{(y - X^T w^*)^T}_{\text{Error vec}}\underbrace{(X^T w^*)}_{\text{vector prediction}} = 0$ } Error vector is $\perp^r$ (orthogonal) to predictions vector.

Gini Impurity for Binary Classification —

$$G(P) = 1 - P^2 - (1-P)^2$$ → $P \leq \frac{1}{2}$ So, $P: 0 \to \frac{1}{2}$ $G(P)\uparrow$

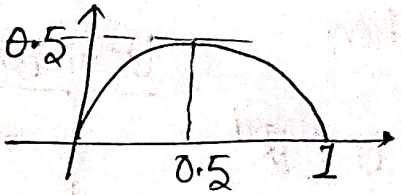
$P: \frac{1}{2} \to 1$ $G(P)\downarrow$

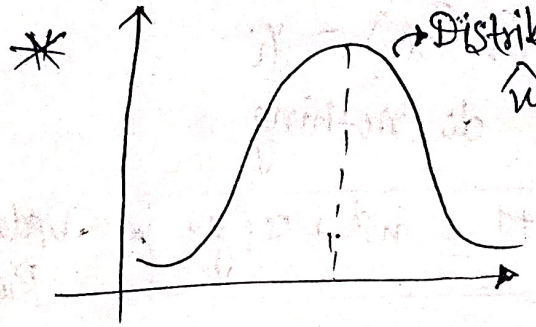
for multiple classes

$$G.I = 1 - \sum_{i=1}^{m} P_i^2$$

m = # of classes

→ $G(P)$ attains its max val. at $P = \frac{1}{2}$



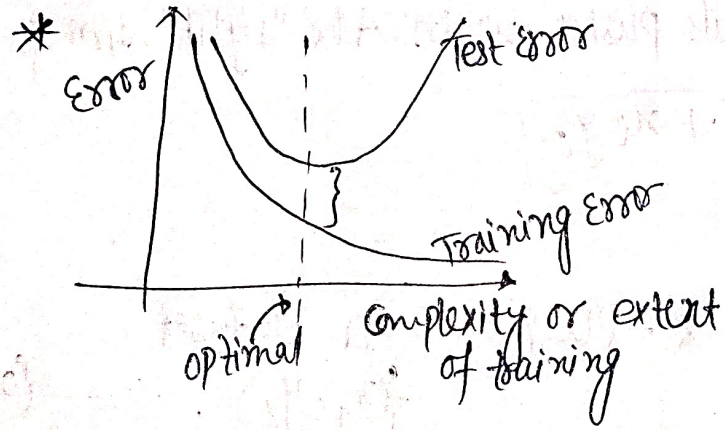


Unbised estimater of w

$E[\hat{w}_{ML}] = w$

Risk = $(\text{Bias})^2$ + varience

$E[\ ] = E[w_{ML} - w]^2 + E[(\hat{w}_{ML} - w)^2]$



weight vector will be perpendicular to data seprator line.

Theorem :- The decision boundry is linear if & only if the variance are equal for both classes.

The decision boundry is ⇒ $P[y=0/x] = P[y=1/x]$

Sub-gradient fact :- if f is differentiable at $a \in \mathbb{R}$, then A has one Sub-gradients, ELSE many.

if a matrix is diagonal then

$\begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$ → Inverse of matrix $= \begin{bmatrix} 1/a & 0 \\ 0 & 1/a \end{bmatrix}$

MLE for exponential: $\hat{\lambda}_{MLE} = \frac{n}{\sum x_i}$

with Uniform prior, MAP = MLE (if within bounds)

Exponential distribution (likelihood)

$X \sim Exp(\lambda)$

$L(\lambda/x_1, \dots x_n) = \lambda^n e^{-\lambda \sum x_i}$

Prior Dist. $\lambda \sim Uniform(0, n)$

$$P(\lambda) = \begin{cases} 1/n, & 0 < \lambda < 2 \\ 0 & \text{otherwise} \end{cases}$$

posterior dis —

$P(\lambda/x_1, \dots x_n) \propto L(\lambda/x_-) \; P(\lambda)$ (Constant)

$\propto \lambda^n e^{-\lambda \sum x_i}$

⇓

Apply log & diff & get MAP estimator

$$\hat{\lambda}_{MAP} = \arg\max_{0<\lambda<n} \left[n \ln \lambda - \lambda \sum x_i \right]$$

if Uniform(0, θ) [Likelyhood]

Likelyhood $= \begin{cases} \theta^{-n}, & \theta \geq \max(x_i) \\ 0 & \text{otherwise} \end{cases}$

$\frac{1}{(b-a)^n}$

$$\hat{\theta}_{map} = \max(x_i)$$

for Uniform prior

PCA Intution

if Rank of Cov. matrix ≤ n

then PCA will have — ~~n~~ non zero e.v

— & Remaining e.v = 0

Necessary Condition for Cov. mat —

① Symmetric ($C = C^T$)

② positive semi define (PSD)

— All Eigenval ≥ 0

— Equivalently: all principal minors ≥ 0

Sigmoid ka output —

$$P(y=+1/x) = \sigma(z) = \frac{1}{1+e^{-z}}$$

$$P(y=-1/x) = 1-\sigma(z)$$

where $z = w^T x$

Decision Rule Usually —

$$\text{predict class} = \begin{cases} 1 & \text{if } \sigma(z) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

Decision Tree Alg.

Decision trees are prone to overfitting if the maximum depth is set too deep

& Too shallow depth → underfitting

Decision tree depend on split selection purity (IG/Gini), order pe nhi.

Decision tree are sensitive [Small change/perturbations in Dataset → diff. tree (Str) means diff. best split]

⇩

इसी वजह से decision trees Unstable / high var. होता है।

Decision trees can handle both numerical & categorical fn.

numerical → threshold split (Ex: $x<5$)

categorical fn. → value based split

Ridge Regression ka golden Rule —

$$\hat{w}_{Ridge} = (xx^T + \lambda I)^{-1} xy$$

if $\lambda \uparrow$ (increase) → coefficient shrink होता है (0 ki taraf)

$\lambda \downarrow$ (decrease) → Coff. bade ho jate, tbai (ML ke pass)

All data same class then Entropy is ~~Pure~~ 0 & decision tree is pure & No further split, IG=0.

* Data (50-50 split) ⇒ Entropy is maximum

IG = Entropy (parent) − weighted Entropy (children)

Entrop (parent) ≥ weighted Entropy (child)

Always

So, IG 0 বা (+)ve হোতা Negative নহীঁ

Only Lasso can produce exactly zero coefficient.

$\xi_i = 0$ → no bribe ← $y_i(w^T x_i + b) \geq 1$

$\xi_i > 0$ → positive bribe

If model → Simple (linear) ⇒ High bias & Underfitting

→ Complex (polinomial) ⇒ high variance & overfitting

~~MLE → β(α, β)~~

Beta (α, β) ⇒

where, n → trials

k → success

Posterior mean $= \frac{\alpha + k}{\alpha + \beta + n}$ ← Use prior

MLE $[\hat{P}_{MLE}] = k/n$ ← not Use in prior

$\hat{P}_{MAP} = \frac{\alpha + k - 1}{\alpha + \beta + n - 2}$ ← Use prior

ADA-BOOST Algorithm :-

Initialize $D_0(i) = 1/n$ ← total no. of sample point

Iteration

for $t = 0, --, T$

~~Error(ε) = Error~~

$\xi_i = \text{Error}(i) \times D_0(i)$

training Error of hi

$$\alpha_t = \ln\sqrt{\frac{1-\varepsilon_i}{\varepsilon_i}} = \frac{1}{2}\ln\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right)$$

weight ↑

Update Rule:- $D_{t+1}(i) = \begin{cases} D_t(i)\, e^{+\alpha_t} & \text{if misclassified, } \varepsilon_i \neq y_i \\ D_t(i)\, e^{-\alpha_t} & \text{if correctly classified} \end{cases}$

weight ↓

$$D_{t+1}(i) = \frac{D_{t+1}(i)}{\sum_J D_{t+1}(J)} \quad \star$$

$$h^*(x) = \text{Sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

One can prove:-

If $T \geq \frac{1}{2\gamma^2}\ln(2n)$, then Training Error = 0

→ How good is my weak learner?

Low threshold ⇒ FN = Very low & FP = No High
→ Spam, fraud, disease detection

High threshold ⇒ FN = No High & FP ⇒ Low
→ Emails filtering, Recommendation

FN → false Neg.

FP → false positive

$\hat{P}_i^y = p(x_i = 1/y) = 1 - P(x_i = 0/y)$

Var $= \frac{1}{n}\sum_{i=1}^{n} x_i x_J$ ★

Hinge loss $(L_i) = \max(0, 1 - y_i f(x_i))$ where, $f(x_i) = w^T x_i + b$

~~if $1 - y_i f(x_i) \leq 0$ then 0~~

Soft margin SVM

$$\min \frac{1}{2}||w||^2 + C\sum_i \xi_i$$

C : misclassified / margin violation ka penalty weight

Margin width $\propto \frac{1}{||w||}$

Small C :- Error zayda Allow karta hai
:- Margin wide hota hai
:- High bias, Low var.

Large C :- Error ka hard punish karta hai
→ Margin narrow (छोटा) होता है।
→ Low bias, high var.

Hinge loss - SVM
Logistic loss - Logistic regression
Modified hinge loss → perceptron
Squard loss → least square classification

for logistic Regression with threshold T, the decision boundary is — * $P(y=1/x) = \sigma(z)$

$$\sigma(w^Tx) = \frac{1}{1+e^{-w^Tx}} = T \implies w^Tx = \ln\left(\frac{T}{1-T}\right)$$

feature x_1 & x_2 → binary classification problem.

Prediction $(\hat{y}) = 1$ if $w^Tx \geq 0$
Else $(\hat{y}) = 0$ if $w^Tx \leq 0$

for Angle θ —
$w = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$

Total no. of cluster assigments possible for cluster & n-datapts.
$= k^n$

Logistic regression : $P(y=1/x) = \sigma(w^Tx)$
→ Decision boundry : $w^Tx = 0$
→ probability sirf w^Tx pe depend karti hai।
→ w की dirn में जितना आगे, prob. Utni Jayda
→ boundry ke opposite side में, probability Small.

In AdaBoost, a weak learner must have misclassification rate strictly less than 0.5.

Deep decision trees → low bias, high variance → traing accha
→ test risky

~~$\lambda_k^i = P(Z_i = K | X = x_i) = \pi_1 N(x_1;$~~

kernal matrix $= \begin{bmatrix} (1+x_1^T x_1)^2 & (1+x_1^T x_2)^2 \\ (1+x_2^T x_1)^2 & (1+x_2^T x_2)^2 \end{bmatrix}$

if $K(x_i, x_j) = (1+x_i^T x_j)^2$

Now, Centered kernal matrix $= K - K[] - []K + []K[]$

Where, $[] = \frac{1}{n}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ where $n \to$ Polynomial, ---

Distribution | Parameter(s) | MLE (max. likelyhood Estimation —

Distribution	Parameter(s)	MLE (max. likelyhood Estimation —
Bernoulli	p	$\hat{p} = \frac{1}{n}\sum x_i$
Poisson	λ	$\hat{\lambda} = \frac{1}{n}\sum x_i$
Exponential	λ	$\hat{\lambda} = \frac{n}{\sum x_i}$
Normal	μ, σ^2	$\hat{\mu} = \bar{x}$, $\hat{\sigma}^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$
Geometric	p	$\hat{p} = \frac{n}{\sum x_i} = \frac{1}{\bar{x}}$

for Voronai regions —

First find Z_1 & Z_2 then → C_1 & C_2 {mean of Z_1 Data & Z_2 Data}

(2) midpoint of line segment $\{C_1 \& C_2\}$ → $\left\{\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2}\right\}$

(3) Slope of the line segment $\{C_1 \& C_2\}$ → Solope $m_1 = \frac{y_2 - y_1}{x_2 - x_1}$

(4) Slope of perpendicular bisector :- (m_2)

$m_2 = \frac{-1}{m_1}$ → Now, eqn of $\perp^r$ bisector — $(y - y_1) = m_2(x - x_1)$

Normal (μ, σ^2) :-

$$PDF\left(f_x(k)\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)}$$

if The mean of both the clusters are the same then there will be no reassignment.

Variance of the dataset along a principal component
= eigenvalue corresponding to that component
$= \lambda_1 = V_1^T C V_1$ ★

The Trace of XX^T is equal to the trace of X^TX

any kernal of the form $(x_1^T x_2 + C)^P$ is valid kernal if $P \in N$, $P > 0$, $C \geq 0$

RBF/Gaussian for valid kernal: $e^{-\frac{1}{2\sigma^2}||x_1 - x_2||^2}$

$X \rightarrow$ Geo(P) $P(X=K) = (1-P)^{K-1} \cdot P$
↳ required to obtained

# Binomial distribution (n, P) where n = total no. of sampled
P = Success denotes
x → no. of Success from Samples

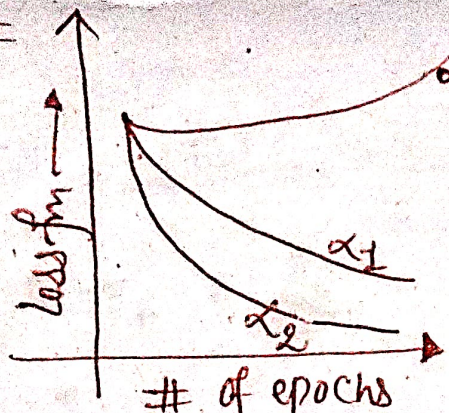
$P = P^x (1-P)^{n-x}$ → likelyhood

Beta(a, b) → $prior = P^{a-1}(1-P)^{b-1}$

$E = \frac{a}{a+b}$

$C = \frac{1}{n} X_c X_c^T$

#



where, α_1, α_2 & α_3 → diff learning rate

$\alpha_3 > \alpha_2 > \alpha_1$

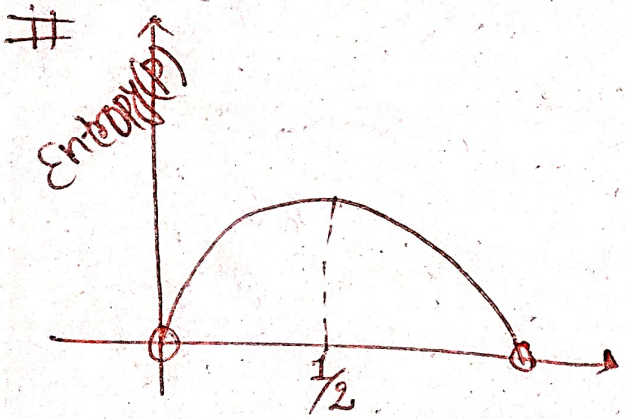
$$w_{t+1} = w_t - \alpha \nabla f_w$$

α → learning rate

$\|E[\theta_1] - \theta\|^2 > \|E[\theta_2] - \theta\|^2$

θ is parameter

θ_2 is a better estimator for θ when compared to θ_1 because it is having the lesser bias.



P → fraction of 1's in y

$L(w; x, y) = \prod_{i=1}^{n} f(y_i | x_i)$

$$L(w; x, y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_i - \mu)^2}$$

$\sigma = 1$

Gaussian Distribution :-

$$f(x \mid y \cdots) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{(x-\mu_k)^2}{\sigma_k^2}\right)}$$

Posterior $\propto$ prior $\times$ likelihood

$$\propto f(w) \times \prod_{i=1}^{n} f(y_i \mid x_i)$$

$$\propto e^{\frac{-1}{2\gamma^2}\|w\|^2} - \frac{1}{2\sigma^2}\|y - x^T w\|^2$$

🔷 Hard-margin vs Soft-margin SVM (Complete Table)

Feature	Hard-margin SVM	Soft-margin SVM
Data assumption	Data perfectly linearly separable	Data not perfectly separable (noise / outliers allowed)
Margin violation	❌ Allowed nahi	✅ Allowed
Misclassification	❌ Allowed nahi	✅ Allowed
Slack variable (ξ_i)	❌ Not used	✅ Used
Constraint	$y_i(w^T x_i + b) \geq 1$	$y_i(w^T x_i + b) \geq 1 - \xi_i$
Objective function	$\min \frac{1}{2}\|w\|^2$	$\min \frac{1}{2}\|w\|^2 + C\sum \xi_i$
Regularization parameter C	❌ Not present	✅ Present
Support vectors lie where?	Only on supporting hyperplanes	On hyperplanes, inside margin, or even wrong side
Points inside margin	❌ Impossible	✅ Possible
Misclassified points	❌ Impossible	✅ Possible
Robust to outliers	❌ No	✅ Yes

11. Bias-Variance Tradeoff

Model Type	Bias	Variance	Error Behavior
Underfitted	High	Low	Poor train/test performance
Overfitted	Low	High	Good train, poor test
Well-generalized	Balanced	Balanced	Good train/test

14. Bagging & Boosting

Technique	Parallel?	Base Learners
Bagging	Yes	High variance (e.g., deep trees)
Boosting	No	Weak learners

C. Ensemble Methods

Technique	Key Characteristics	Use Case / Purpose
Bagging	- Uses **deep decision trees** (high variance models) - **Parallel execution** (models are independent) - Bootstrap sampling (with replacement) - All models usually have equal weight	- **Reduces variance** - Handles overfitting - Example: **Random Forest**
Boosting	- Uses **decision stumps / weak learners** - **Sequential execution** (each model learns from previous errors) - Data-point weights change over iterations - Models have different importance	- **Reduces bias** - Improves weak models - Examples: **AdaBoost, Gradient Boosting**

Contrast with Boosting (important 🔥)

Method	Parallel training?	Reason
Bagging	✅ Yes	Models independent
Random Forest	✅ Yes	Bagging + feature randomness
AdaBoost	❌ No	Each model depends on previous

Why? (Bias–Variance intuition)

Model complexity aur test error ka relation **monotonic nahi hota.**

- Jab model **bahut simple** hota hai
  → **High bias**
  → Underfitting
  → **High test error**
- Complexity **badhaane par**
  → Bias kam hota hai
  → Test error **pehle kam hota hai**
- Complexity **bahut zyada** ho jaaye
  → **High variance**
  → Overfitting
  → **Test error phir badhta hai**

📊 Activation Functions – General Conditions & Formula

Activation Function	Formula	Output Range	Typical Condition / Property	Common Use
ReLU	$f(z) = \max(0, z)$	$[0, \infty)$	Negative input → 0, Positive → same	Hidden layers
Leaky ReLU	$f(z) = \max(\alpha z, z)$	$(-\infty, \infty)$	Small slope for negative values	Hidden layers
Sigmoid	$f(z) = \frac{1}{1+e^{-z}}$	$(0, 1)$	Output interpretable as probability	Binary output
tanh	$f(z) = \tanh(z)$	$(-1, 1)$	Zero-centered output	Hidden layers
Softmax	$f(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$	$(0, 1)$, sum = 1	Outputs sum to 1	Multi-class output
Linear	$f(z) = z$	$(-\infty, \infty)$	No non-linearity	Regression output
Step	$f(z) = 1[z \geq 0]$	$\{0, 1\}$	Hard threshold	Perceptron

🔑 Sabse pehle clear distinction

🔹 Hard-margin SVM

- **No point** margin ke andar hota hai
- **No misclassification**
- Constraints:

$$y_i(w^T x_i + b) \geq 1$$

🔹 Soft-margin SVM

- Points **margin ke andar bhi ho sakte hain**
- Misclassification allowed
- Constraints:

$$y_i(w^T x_i + b) \geq 1 - \xi_i$$

🫧 Ab tumhare statement par directly aate hain

> **"kuchh points to support hyperplane ke beech me bhi hota hai"**

✅ **Ye baat soft-margin SVM me true hai**

❌ **Lekin hard-margin SVM me false hai**