__/__/____

ADVANCED MACHINE LEARNING

9/1/23

⇒ Well known conferences for ML

- ICML
- Neurips
- ICLR
- AAAI
- CVPR
- ACL

↓

Reproduce the results in papers.

A Statistical Learning framework.

Domain set - x

Label set - y.

Training data: $S = ((x_1, y_1), (x_2, y_2), \ldots$
$\ldots (x_m, y_m))$.

from $x \times y$.

Learner's output: $h : x \to y$.

(Source: 2nd book - Learning from Data)

Data generation model

Prob. dist. over x is D.

Labelling func^n: $f : x \to y$.

Measures of success.

(Gen error) $L_{D,f}(h) = \underset{x \sim D}{P} [h(x) \neq f(x)]$ → loss func^n

↓

over entire distribution

learn → h

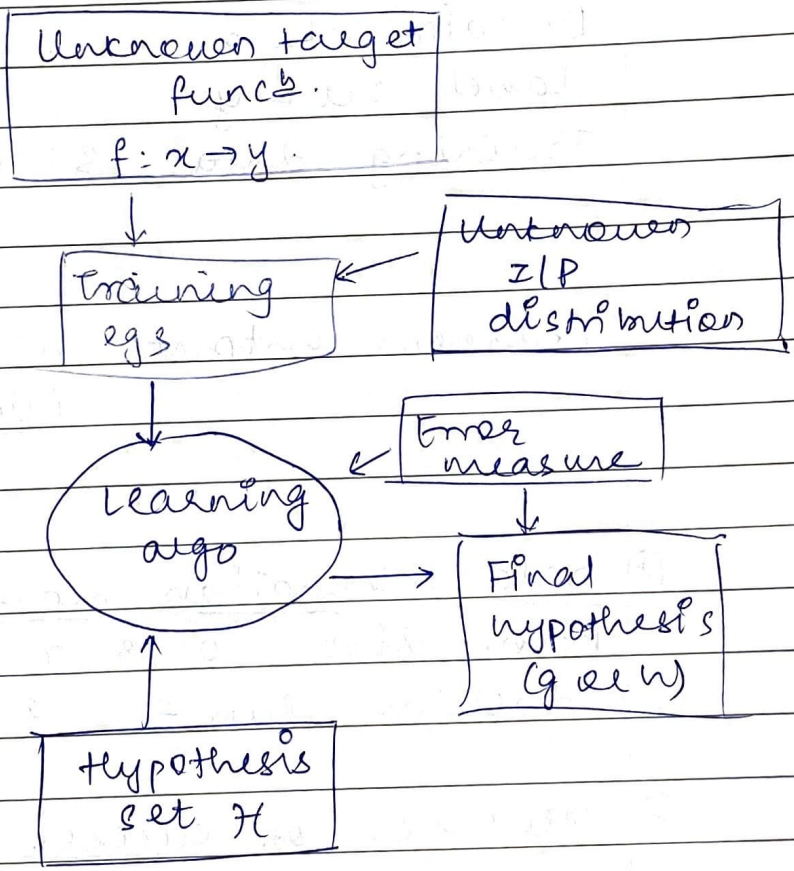
fixed → f

(M) ERM framework. (Empirical Risk Minimization)

Training error → (empirical error) $L_S(h) = \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$.

Search for a soln that works well on available data.

→ I.I.D assumption
→ Goal is that func^n perform well on test data.



16/1/23 **Lec: 2**

$$E_{in}(h) = \frac{1}{N}\sum_{j=1}^{N}[h(x_m) \neq f(x_m)]$$

$$E_{out}(h) = P[h(x) \neq f(x)]$$

① **Marble Experiment**

A bin with red & green marbles.

Prob. of picking red → μ

" green → $1-\mu$

Problem → μ is unknown

N – total no. of marbles.

V – proportion of red marbles out of N

$\mu = \frac{V}{N}$ If 'N' is large enough, then μ is near to empirical mean

② **Hoeffding's Inequality**

– type of Chernoff Bound.

$$P[|V-\mu| > \varepsilon] \leq 2e^{-2\varepsilon^2 N}, \quad \varepsilon > 0.$$

$|V-\mu|$: empirical mean, original mean; δ

→ eg. If $\varepsilon = 0.1$, what is no. of samples?

δ – failure probability

Relate this to learning. (h = hypothesis)

For fixed hypothesis,

$$P[|E_{in}(h) - E_{out}(h)| > \varepsilon] \leq 2e^{-2\varepsilon^2 N}, \quad \varepsilon > 0$$

(∞) Union Bound

Events → A & B.

$P(A \cup B) \leq P(A) + P(B)$

→ For a fixed set of hypothesis of size M

$P[|E_{in}(h) - E_{out}(g)| > \varepsilon] \leq 2M e^{-2\varepsilon^2 N}$, $\varepsilon > 0$.

(derive)

with probability atleast $1-\delta$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$$

↑ Generalization Bound

Let $2Me^{-2\varepsilon^2 N} = \delta$.

$e^{-2\varepsilon^2 N} = \frac{1}{2M}\delta$.

$-2\varepsilon^2 N = \ln\left(\frac{\delta}{2M}\right)$

$$\therefore \varepsilon = \sqrt{\frac{1}{2N} \ln\left(\frac{2M}{\delta}\right)}$$

$-2\varepsilon^2 N = \ln \frac{\delta}{2M}$

$N = \frac{1}{2\varepsilon^2} \ln\left(\frac{2M}{\delta}\right)$

H - hypothesis space

(M) PAC learning → Probably Approximately correct

It is PAC learnable if
- $m_H : (0,1)^2 \to N$
- $\varepsilon, \delta \in (0,1)$, D over $X \times Y$

→ ~~When running the learning algo.~~

(N) Defns

Dichotomies:
$$H(x_1, \ldots, x_N) = \{h(x_1), \ldots, h(x_N) \mid h \in H\}$$

19/1/23 With ~~probability~~ atleast $1-\delta$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$$ ← generalization bound

→ What if $M = \infty$?

(O) Growth function - for a hypothesis set H is

$$m_H(N) = \max_{x_1 \cdots x_N} |H(x_1, x_2, \ldots, x_N)|$$

$$m_H(N) \leq 2^N$$

(P) Shattering: If $m_H(N) = 2^N$. (hypothesis set can generate all possible dichotomies)

(Q) VC-Dimension: of a hypothesis set H, denoted by $d_{VC}(H)$ or d_{VC} is the largest value of N for which $m_H(N) = 2^N$

Learning Theory

__/__/__

→ If $m_H(N) = 2^N$ for $\forall N$, $d_{VC}(H) = \infty$

VC(hyperplanes) = d + 1

⇒ Sauer's Lemma

→ VC-Generalization bound

- more complex hypothesis - doesn't generalize well

Evaluation

① Reproducibility challenge - write reproducibility report
   | paperswithcode.com/rc2022
   | Neurips 2021/22
   | ICML 2021/22
   | AAAI 2021/22
   | ACM FACCT 2021/22
   ↓
   choose any 1 paper
   - study algorithmic
   - choose papers with more implementation
   - Replicate results (on same datasets)
   - can you do something more?
     ◦ Tweak algo
     ◦ more dataset
     ◦ hyperparameter tunings
   - Replicate based on library changing

② Project

repeat of 8-10 pages

① Rep. paper
② project — Empirical
   | Theory

→ git repository for this ✓
(→ distill.pub)

Learning from data: A short course book

__/__/____

20/1/23

⑥ Bias Variance tradeoff and PLA:

VC Generalization Bound.

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4 m_H(2N)}{\delta}}$$

The VC bound is loose.

~~Hoeffdings Inequality~~ Hoeffdings Inequality Holds:-

$$m_H(N) \leq \sum_{i=1}^{d_{VC}} \binom{N}{i}$$

$$m_H(N) \leq N^{d_{VC}} + 1$$

→ Hoeffding's inequality has some slack. $m_H(N)$ gives a worst case estimate

→ Bounding $m_H(N)$ by polynomial of order

→ (Not easy to get V_C dimension of decision tree)

More V_C dimension = more complex

⇒ Rule of thumb → keep training data 10 times the V_C dimension

23/1/23

(00) Predictive & Generative Models.

Bias & Variance:-

$E_{out}(g^{(D)}) = E_x[(g^{(D)}(x) - f(x))^2]$

E_x denotes expected value w.r.t. x

Imp: $g^{(D)}$ is dependent on (D) (data)

$E_D[E_{out}(g^{(D)})]$ (E_{out} = out sample error)

$= E_D[E_x[(g^{(D)}(x) - f(x))^2]]$

$= E_x[E_D[(g^{(D)}(x) - f(x))^2]]$ ← $(a-b)^2$

$= E_x[E_D[g^{(D)}(x)^2] - 2E_D[g^{(D)}(x)]f(x) + f(x)^2]$

$E_D(g^D(x)) = \bar{g}(x)$

$E_x[E_D(g^D(x)^2) - 2\bar{g}(x)f(x) + f(x)^2]$

$= E_x[E_D(g^D(x)^2) - \bar{g}(x)^2 + \bar{g}(x)^2 - 2\bar{g}(x)f(x) + f(x)^2]$

$(\bar{g}(x) - f(x))^2$

$= E_D[(g^{(D)}(x) - \bar{g}(x))^2] + (\bar{g}(x) - f(x))^2$

$= \text{var}(x) + \text{bias}$

idea - we can break down our error into bias & variance.

→ It also has irreducible error, but we assume here that data is noise free

$\therefore E_D [E_{out}(g^{(D)})] = E_x [bias(x) + var(x)]$

here, along with H algorithm, A also matters

→ Total least square (orthogonal least square)
↳ tries to reduce the ⊥er distance between points & line

→ OLS - minimize the vertical offset.

⊛ PLA algorithm (Perceptron Learning Algorithm)

$w^{(0)} = (0, 0, 0, 0, 0)$ (initial)

while there is a misclassified point $x(t), y(t)$

$w(t+1) = w(t) + y(t) \cdot x(t)$ [$y(x) = sgn(w^T x)$]

→ If a point is classified correctly, then, $((w^T x)(y)) > 0$.

→ If data is linearly separable, PLA gives 0 errors

27/1/23.

Convex Optimization in ML
Primal Dual Formulation
KKT conditions
} SVM problem

Optimization for ML

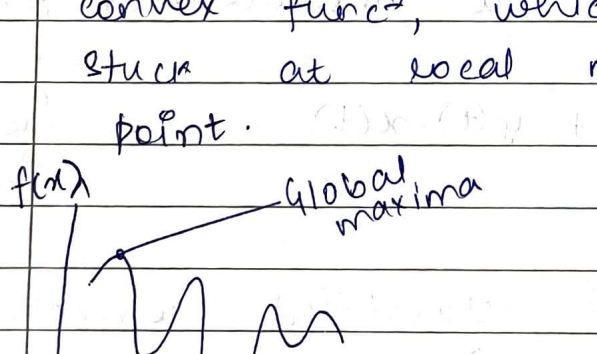
- Continuous
  - Non convex
  - Convex
- Combinational
  - Submodularity
  - Subset Selection

Gradient Descent → move in direction of -ve gradient.

[Optimization for ML – Nisharth Veshna]

→ In convex fn, there is high chance to reach global minima with more no. of iterations.

→ In practice, we use SGD in non-convex func, which can also get stuck at local minima or saddle point.



__/__/____

$f: \mathbb{R}^m \rightarrow \mathbb{R}$

Unconstrained
$\min f(x)$
such that
$x \in D = \mathbb{R}^m +$ —.

Constrained
$\min f(x)$
s.t.
$x \in \mathbb{R}^m$

Linear regression:-

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|^2$$ ← Unconstrained optimization

Ridge regression:-

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|^2 + \lambda \|x\|_2^2$$

Lasso regression:-

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2 + \lambda \|x\|$$

shape? ← Least Absolute Shrinkage & Selection Operator

Condⁿs:-

i) Let $f: \mathbb{R}^m \rightarrow \mathbb{R}$ Continuous & differentiable

iii) For local minima

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

eg. $x_1^2 + 2x_2$.

$$\nabla f(x) = \begin{bmatrix} 2x_1 \\ 2 \end{bmatrix}$$

Positive definite → ?
" semi- " → ?

Hessian matrix - matrix of partial derivatives

6/2/23

→ Eg. of unconstrained optimization problem → Linear regression

→ Eg. of constrained → Lasso, Ridge

constraints → $\|x\|$, $\|x\|_2^2$

$H(x)$ ↑ Hessian

$x^T A x$ → gives scalar

$x^T A x > 0 \;,\; \forall\, x \neq 0$

positive semi-definite

→ If all eigen values are +ve then.

$Ax = \lambda x$

⇒ Convex set :- $X \subseteq R^n$ is said to be convex if $\forall\, x, y \in X$

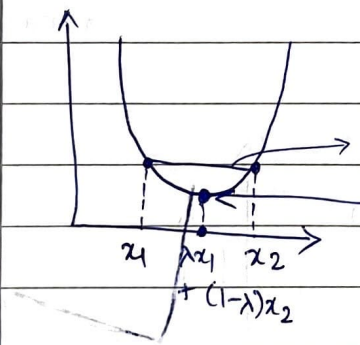
$\lambda x + (1-\lambda)\, y \in X \;,\; \forall\, 0 \leq \lambda \leq 1$

⇒ Convex function :- If $f : R^n \to R$ (defn)

$R^n \subseteq X$

$\forall\, x_1, x_2$

$$f(\lambda x_1 + (1-\lambda)\, x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \quad \text{—— (1)}$$



∴ f is convex fun^n

If in eq² (1), '<' sign → strictly convex

$\max(0, x)$ $\|x\|_p$

__/__/__

→ If $f(x)$ is convex $\Leftrightarrow$ $H(x)$ is P.S.D / $\forall x \in X$

→ $f(\cdot)$ is strictly convex $\Leftrightarrow$ $H(x)$ is PD $\forall x \in X$.

→ If func^2 is convex, local minima = global minima

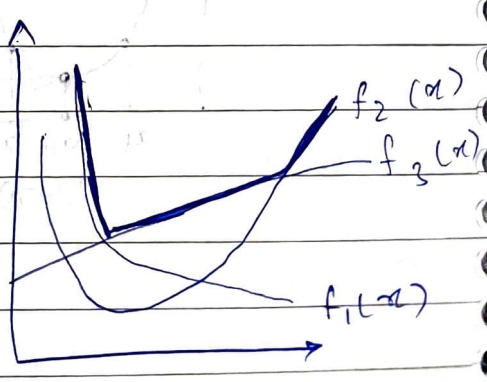
Egs
① $w^T x + b$ on R^n — convex
② -ve log likelihood func^n — "
③ quadratic func^n → $f(x) = \left(\frac{1}{2} x^T A x + b^T x + C\right)$ on R^n (A is PSD).

→ check if $\max: f(x) = \max(x_1, \ldots, x_n)$ is convex?

⑩ Operations preserving convexity
① $f(x)$ is convex
$\Rightarrow \alpha f(x)$ is also convex, $\alpha > 0$

② Sum:- $f_1(x) + f_2(x) = F(x)$
If $f_1(x)$ & $f_2(x)$ are convex, $F(x)$ is also convex.

③ Pointwise max:-
$f(x) = \max_{1 \le i \le k} f_i(x)$
↑ convex



Smooth func^{n}?

__/__/__

④ log sum exponential

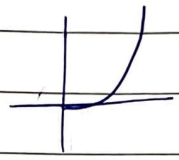
$$f(x) = \log \sum_{i=1}^{n} \exp(x_i)$$

⑤ Convex functions by composition

$h(x) \rightarrow$ h is convex

case: i $g(h(x))$

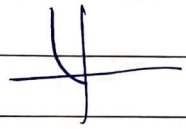
└ convex + non decreasing



case: ii $g(h(x))$

└ concave

└ convex + non increasing



8/2/23 AML

→ Unconstrained optimization
→ Lagrange multipliers

⊛ constrained optimization problem:

$f_0 : \mathbb{R}^n \to \mathbb{R}$

$\min \; f_0(x)$

such that $f_i(x) \leq 0 \quad i \in [1, m]$

$f_j(x) = 0 \quad j \in [1, l]$

no. of constraints $= m + l$

$$f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{l} \mu_j f_j(x)$$

Lagrange multipliers problem

$$\|Ax - b\|_2^2 + \lambda \|x\|_2^2, \quad \lambda > 0$$

$$\|Ax - b\|_2^2 \text{ s.t. } \|x\|_2^2 \leq R$$

↳ where λ_i, μ_j are called Lagrangian multipliers.

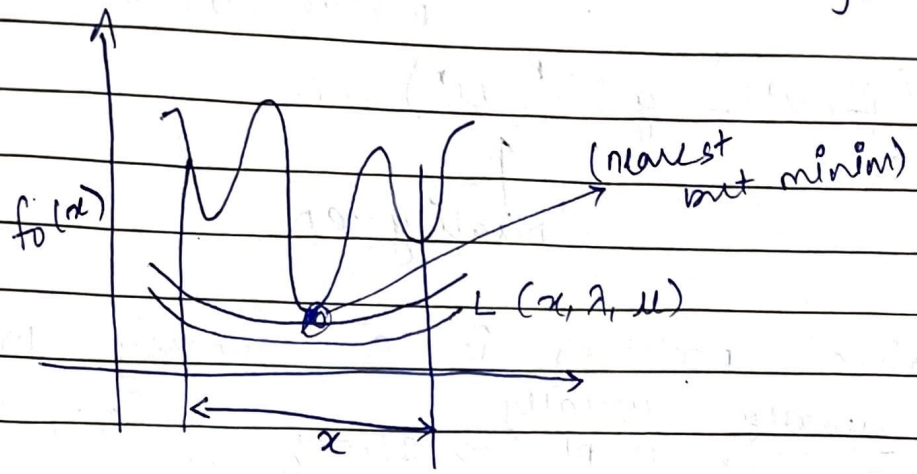
$$L(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{l} \mu_j f_j(x)$$

→ what if x is a feasible point ?

↓

(means it satisfies constraints)

For feasible x, $L(x, \lambda, \mu) \leq f_0(x)$



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

$$L(\lambda, \mu) = 2\lambda_1 \mu_1^2 + 3\lambda_2^2 \mu_2$$

$$\min L(x, \lambda, \mu) \text{ such that } \lambda \geq 0 \leq f_0(x^*)$$

lower bound on optimal solution

$$\max_{\lambda, \mu} \left[\min_x L(x, \lambda, \mu) \right] \leq f_0(x^*)$$

$(\lambda, \mu) \in \text{Domain}$ s.t. $\lambda \geq 0$

$$\min_x L(x, \lambda, \mu) = g(\lambda, \mu)$$

dual function

→ Primal problem & Dual problem

$$\max_{(\lambda, \mu)} g(\lambda, \mu)$$
$$\lambda \geq 0$$

$g(\lambda, \mu) = \min_x L(x, \lambda, \mu)$

weak duality theorem → (λ^*, μ^*) are point of maxima.

$$g(\lambda, \mu) \leq g(\lambda^*, \mu^*) \leq f_0(x^*)$$

Weak duality theorem - statement

$f_0(x^*)$ - optimal value of primal function

$$(f_0(x^*) - g(\lambda^*, \mu^*) \geq 0$$

↑ Duality gap

→ Primal problem is a convex problem

⟹ usually implies → strong duality holds (not always)

↓ (Optimum value of dual = value of primal)

$[f_0(x^*) = 0 \; ??]$

constraints should satisfy some cond^n

→ Regularity constraint

KKT conditions :-

x^* - point of primal optimal

λ^*, μ^* - point of dual optimal with zero duality gap

(i) Primal feasibility

→ x^* is feasible sol^n for primal problem.

$$f_i(x^*) \leq 0 \quad i \in [1, m]$$

$$f_j(x^*) \geq 0 \quad j \in [1, l]$$

(ii) Dual feasibility

$$\lambda^* \geq 0$$

(eg. support vectors)

(iii) Complementary Slackness :-

$$\lambda_i^* f_i(x^*) = 0 \quad \forall i$$

(iv) Lagrange optimality (func^n of x)

$$\nabla_x L(x, \lambda^*, \mu^*) = 0$$

→ Any tuple (x, λ, μ)

$x \in R^n$, $\lambda \in R^m$, $\mu \in R^l$

is called a KKT point if it satisfies KKT cond^n.

→ When you have zero duality gap, $(x^*, \lambda^*, \mu^*) \Rightarrow$ KKT point

If something is KKT point, it will always be optimal (if func^n is convex)

⑧ General stat.

If duality gap = 0

x^* – Primal optimal

λ^* & μ^* } dual optimal

9/3/23 AML

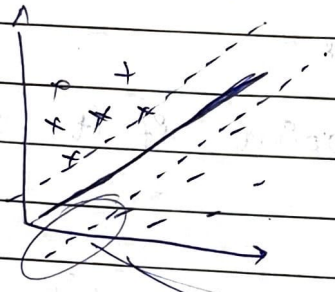
Scribe notes

Optimization book

→ Applications for SVM:-

LDL → used in NLP for topic modelling

(a) Hard Margin SVM (what is it?)



what is margin?
why we maximize margin?

Supporting hyperplanes

→ Maximum margin classifier

-1 & +1 labels are used

Hard margin SVM

→ Primal SVM

$$\min_{(w, w_0)} \frac{1}{2}\|w\|_2^2 \quad \text{such that} \quad y_i(w^T x_i + w_0) \geq 1$$

$\forall i,\ i \in [1, n]$

$\text{sign}(w^T x_{new} + w_0)$ → whatever the sign, assign that label (+1 or -1)

(b) Soft margin SVM → slack

$$y_i(w^T x_i + w_0) \geq 1 - \xi_i \quad (\text{psi})$$

$$\text{i.e. } \geq -1 + \xi_i$$

→ allows some misclassifications

Date
Page

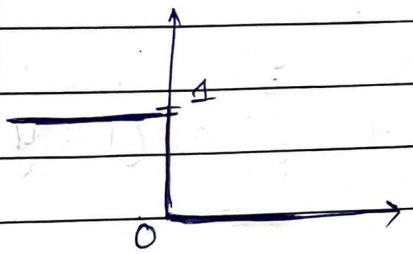
⇒ Hinge Loss ?

$y_i (w^T x_i + w_0) < 1$

$0 < 1 - y_i (w^T x_i + w_0)$

$l^{hinge} [(x, y), (w, w_0)] = \max \{0, 1 - y_i (w^T x_i + w_0)\}$

→ Zero-one loss



Hinge

⇒ Applying KKT conditions :-

$\min_{(w, w_0)} \frac{1}{2} \|w\|_2^2$ such that $y_i (w^T x_i + w_0) \geq 1$

$\forall\ i \quad i \in [1, n]$

(i) Primal Feasibility

$1 - y_i (w^T x_i + w_0) \leq 0$

w^*, w_0^*

primal variables → w & w_0

$\mathcal{L}(w, w_0, \alpha) = \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i (w^T x_i + w_0))$

Date
Page

(ii) Dual feasibility $\alpha_i^* \geq 0 \quad \forall i$

↑ (only these are support vectors that satisfy this condition)

(iii) Complementary slackness :-

$$\alpha_i^* (1 - y_i (w^{*T} x_i + w_0^*)) = 0$$

(iv) Lagrange optimality :-

$$L(w, w_0, \alpha) = \frac{1}{2} \|w\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i (w^T x_i + w_0))$$

$$= \frac{1}{2} w^T w + \sum_{i=1}^{n} \alpha_i (1 - y_i (w^T x_i + w_0))$$

$\nabla_w L(w, w_0, \alpha) = 0$: (differentiating scalar gives vector)

$$\nabla_w L(w, w_0, \alpha) = w$$

$$\sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i (w^T x_i + w_0)$$

$$\nabla_w L(w, w_0, \alpha) = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0$$ (partial derivative w.r.t. w)

$$\Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i$$ cond^n 1

gives hyperplane in terms of coeffs & S.V.S

$$L(w, w_0, \alpha) = \frac{1}{2} w^T w + \sum_{i=1}^{n} \alpha_i (1 - y_i (w^T x_i + w_0))$$

$$\nabla_{w_0} L(w, w_0, \alpha) = 0 + 0 - \sum_{i=1}^{n} \alpha_i y_i$$

$$\therefore \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{cond}^n 2$$

α_i for $i \in C^+$

α_j for $j \in C^-$

$$\sum_{i \in C^+} \alpha_i = \sum_{j \in C^-} \alpha_j$$

→ Any eg. for which dual var (α) is non-zero, will lie on supporting hyperplane.

⇒ Defining support vectors :-

$\{ x_i \mid \alpha_i^* > 0 \}$ SV are most difficult to classify

→ Getting dual form from primal :-

$$\min_{(w, w_0)} \frac{1}{2} \|w\|_2^2 \quad \text{s.t.} \quad 1 - y_i (w^T x_i + w_0) \leq 0 \quad \forall i$$

$$L(w, w_0, \alpha) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i (w^T x_i + w_0))$$

$$g(\lambda, \mu) = g(\alpha) = \min_{w, w_0} L(w, w_0, \alpha)$$

$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i^*$, w_0^* satisfies

$$\sum \alpha_i^* y_i^* = 0$$

$$g(\alpha) = \frac{1}{2} w^{*T} w^* + \sum_{i=1}^{n} \alpha_i^* - \sum_{i=1}^{n} \alpha_i^* y_i w^{*T} x_i^*$$

$$- \sum_{i=1}^{n} \alpha_i^* y_i^* w_0^* = 0$$

$$= \frac{1}{2} w^{*T} w^* + \sum_{i=1}^{n} \alpha_i^* - w^{*T} \sum_{i=1}^{n} \alpha_i^* y_i^* x_i^* - 0$$

$$= \frac{1}{2} w^{*T} w^* + \sum_{i=1}^{n} \alpha_i^* - w^{*T} w^*$$

$$= -\frac{1}{2} w^{*T} w^* + \sum_{i=1}^{n} \alpha_i^*$$

$$= -\frac{1}{2} \left(\sum_{i=1}^{n} \alpha_i^* y_i^* x_i^* \right)^T \cdot \left(\sum_{j=1}^{n} \alpha_i^* y_i^* x_i \right) + \sum_{i=1}^{n} \alpha_i^*$$

$$= \max \left[\frac{-1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i^* \alpha_j y_i y_j x_i^{*T} x_j^* + \sum_{i=1}^{n} \alpha_i^* \right]$$

↑ Dual form of Hard Margin SVM

s.t. $\alpha_i \geq 0 \quad \forall i$

$$\sum_{i=1}^{n} \alpha_i^* y_i^* = 0$$

✓

suibe notes end

Date 10/03/23
Page

10/2/23

Hard Margin SVM

Primal ⇒

$$\min_{(w, w_0)} \frac{1}{2} \|w\|_2^2 \quad \text{s.t.} \quad y_i (w^T x_i + w_0) \geq 1 \quad \forall i$$

Dual →

$$\max \; -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{n} \alpha_i$$

$$\text{s.t.} \; \alpha_i \geq 0 \quad \forall i \quad \& \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

→ **Kernel SVM** $\quad x \in \mathbb{R}^d$

$$x \xrightarrow{\Phi} z_i \in \mathbb{R}^n, \quad n > d$$

$$\Phi(x) \in \mathbb{R}^n$$

$\uparrow$

$\mathbb{R}^d$

$x_i \Rightarrow \Phi$

$z_i = \Phi(x_i)$

$$\text{label}(x_{test}) = \text{sign}\left[w^{*T} \Phi(x_{test}) + w_0^* \right]$$

$$w^* = \sum_{i=1}^{m} \alpha_i^* y_i \Phi(x_i)$$

m data points

$$= \text{sign}\left[\sum_{i=1}^{m} \alpha_i^* y_i \Phi(x_i)^T \Phi(x_{test}) + w_0^* \right]$$

kernel

ϕ = feature func^n

Date
Page

Q How to implement SVM in high dimension without actually getting $\phi(x_i)$

Mercer's theorem:-

Kernel func^n :- $K: X \times X \rightarrow R$

A symmetric function $K(x_i, x_j) = K(x_j, x_i)$ can be expressed as $K(x_i, x_j) = \phi^T(x_i)\phi(x_j)$ for some ϕ.

Iff the matrix,

$$K = \begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) & \cdots & K(x_1, x_m) \\ \vdots & & & \\ K(x_m, x_1) & \cdots & \cdots & K(x_m, x_m) \end{bmatrix}_{m \times m}$$

Gram matrix

K is Positive definite matrix

Replacing $\phi^T(x_i)\phi(x_j) = K(x_i, x_j)$ as $K(x_i, x_{test})$ is called Kernel trick.

(4) Examples of kernel

(1) Linear Kernel $K(x_i, x_j) = x_i^T x_j$
$= x_i^T I x_j$

(2) Polynomial kernel
$K(x_i, x_j) = (1 + x_i^T x_j)^t \; ; \; t > 0$

Date
Page

③ Gaussian kernel (RBF kernel)
Radial Basis Function

$$K(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|_2^2}{2\sigma^2}\right)$$

$K_1(x_i, x_j)$ & $K_2(x_i, x_j)$

$K = K_1(x_i, x_j), K_2(x_i, x_j)$

$K(x_i, x_j) = cK(x_i, x_j), \quad c > 0$

$\Rightarrow$ If $x_1 > x_2$, $\exp(x_1) > \exp(x_2)$
(monotonically ↑ing)

④ Soft-SVM

① L1-SVM

$$\min_{(w, w_0, \xi)} \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \xi_i$$

(λ: Regularization parameter)

$$\text{s.t. } y_i(w^T x_i + w_0) \geq 1 - \xi_i$$

$$\xi_i \geq 0 \ \forall i$$

② L2-SVM

$$\min_{(w, w_0, \xi)} \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \xi_i^2$$

③ L1C-SVM

$$\| + \frac{c}{n}\sum_{i=1}^{n} \xi_i + \frac{1}{2}\|w\|^2$$

④ L2C-SVM

$$\frac{c}{n}\sum_{i=1}^{n} \xi_i^2 + \frac{1}{2}\|w\|^2$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

Let $X = R$

$$K(x_i, x_j) = \exp\left(\frac{-(x_i - x_j)^2}{2\sigma^2}\right)$$

$$= \exp\left(\frac{-x_i^2}{2\sigma^2} - \frac{x_j^2}{2\sigma^2} + \frac{2x_i x_j}{2\sigma^2}\right)$$

$$= \exp\left(\frac{-x_i^2}{2\sigma^2}\right) \exp\left(\frac{-x_j^2}{2\sigma^2}\right) * \exp\left(\frac{2x_i x_j}{2\sigma^2}\right)$$

$$= \exp\left(\frac{-x_i^2}{2\sigma^2}\right) \exp\left(\frac{-x_j^2}{2\sigma^2}\right) \left[1 + \frac{x_i x_j}{\sigma^2} + \frac{(x_i x_j)^2}{2!(\sigma^2)^2} + \cdots\right]$$

$$\exp\left(\frac{-x_i^2}{2\sigma^2}\right) \exp\left(\frac{-x_j^2}{2\sigma^2}\right) \left\langle 1, \frac{x_i}{\sigma}, \frac{x_i^2}{\sqrt{2!}\sigma^2} \cdots \right\rangle \left\langle 1, \frac{x_j}{\sqrt{2!}\sigma}, \cdots \right\rangle$$

$$= \exp\left(\frac{-x_i^2}{2\sigma^2}\right) \left[1, \frac{x_i}{\sigma}, \frac{x_i^2}{\sqrt{2!}\sigma^2}, \cdots\right]$$

in terms of x_i in infinite dimensions

$\phi(x_i)$

separating & supporting hyperplanes

SVC - default = RBF

Date
Page

13/2/23

SVD → Singular Value Decomposition

(*) Eigen Value Decomposition

A (non-zero) vector v of n dimension is an eigen vector of the n×n matrix A if it satisfies

$$Av = \lambda v$$

PCA → (variance - covariance matrix)

└ uses SVD.

SVD used in:-

(i) Dimensionality reduction (PCA)

(ii) Calculating pseudo-inverse $(A^TA)^{-1}A^T$

(etc)

(SVD.pdf.) - sir's notes.

$A = V \Lambda V^T$

(Λ → diagonal matrix)

(dept of eigen vector & eigen vector)

$Av = \lambda v$

v → vectors, λ = scalar (eigen value)

$A \in R^{n \times d}$

$Ax \rightarrow Ax \in R^{n \times 1}$

$x \in R^d$

↳ If A is a square matrix (n×n) with n linearly independent eigenvectors, A can be factorised as

$$A = Q \Lambda Q^{-1}$$

$$Q = \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_n \\ | & | & & | \end{bmatrix}$$

→ Linear comb^n of vectors
→ Basis (size of basis?)
→ Null space, column space (or range space)

$$Ax = \begin{bmatrix} | & | & \\ a^1 & a^2 & \cdots \\ | & | & \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \qquad \sum_{j=1}^{d} x_j \underbrace{a^j}_{\text{vectors}}$$

x_j ← scalar

→ Rank of matrix

$Ax = 0$

→ All x that satisfy $Ax = 0$ → Null space

dim (range space) + dim (null space) = total dimension

↳ Every n×n real symmetric matrix
⇒ eigen values are real & eigen vectors can be chosen orthonormal

$Av = \lambda v$

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

$\therefore A = Q \Lambda Q^T$

$A = Q \Lambda Q^T$ $(Q^{-1} = Q^T)$ — when?

(Linear algebra book) — refer

→ (orthonormal?) → k vectors,

$x_i^T x_j = 0 \quad \forall \; i \neq j$

→ (difference b/w orthogonal & orthonormal)

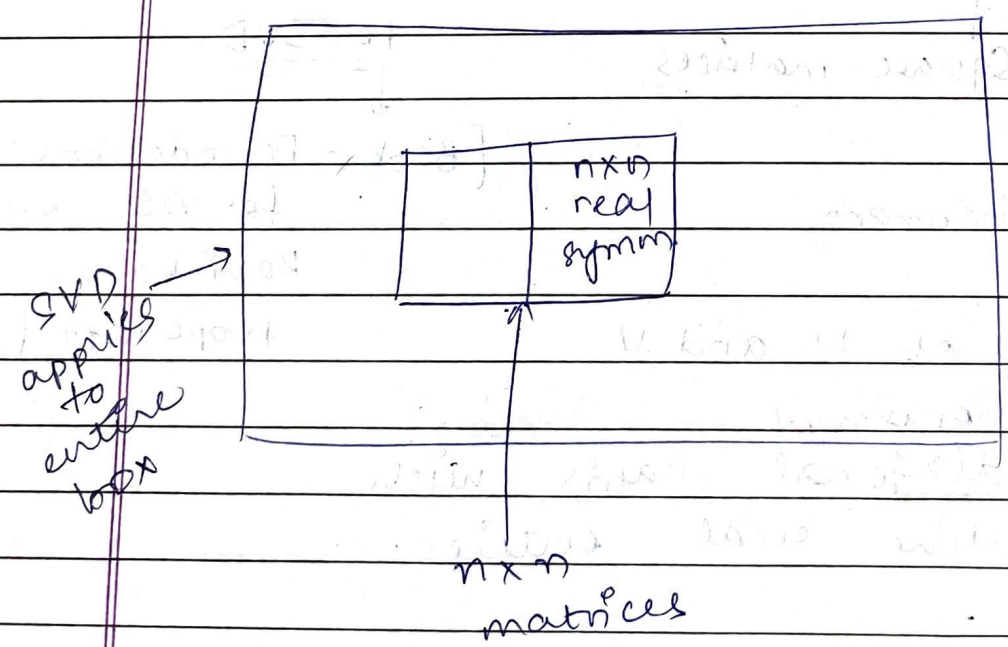
$x_i^T x_i = 1 \quad \forall \; i = i$

→ Algebraic multiplicity :- No. of repititions of a particular eigen value is its algebraic multiplicity.

→ Geometric multiplicity :- No. of linearly independent eigen vectors associated with it. i.e. dimension of null space of $A - \epsilon I$.

$AM(\epsilon) \geq GM(\epsilon)$

⇒ Issues with EVD :- (eigen value decomposition)
- Only applicable to square matrices
- May be complex eigenvalues



SVD

Any matrix A $(n \times d)$ can be decomposed as

$$A = U \Sigma V^T$$

$U_{n \times n}$	$U_{n \times d}$
$\Sigma_{n \times d}$	$\Sigma_{d \times d}$
$V^T_{d \times d}$	$V^T_{d \times d}$

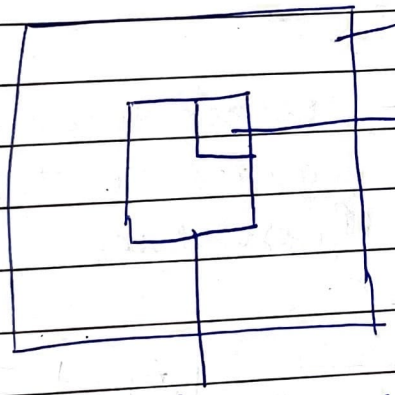
Diagonal matrix ≥ 0

(singular values always non-negative)

$U^T U = I_n$

$V^T V = I_d$

20/2/23



For SVD.

↓

[Book – Foundations for DS by Ravi Kannan Microsoft]

SVD (contd.)

Spectral theorem

→ Columns of U and V are orthonormal

– Σ is diagonal matrix with non-negative real entries.

$A^{10\times3}$ $(A = U\Sigma V^T)$

$U^{10\times3}$ $\Sigma^{3\times3}$ $V^{T\,3\times3}$

$$\begin{bmatrix} U \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & \sigma_3 \end{bmatrix} \begin{bmatrix} V^T \end{bmatrix}$$

→ Power iteration methods.

$(Av = A^2v)$

A^TA → square symm.

Suppose we consider the square symm matrix $- A^TA$.

Let x be eigenvector of A^TA and λ be its eigenvalue.

$$A^TAx = \lambda x.$$

Multiplying by A on both sides.

$$AA^T(Ax) = \lambda(Ax) \Rightarrow (AA^T(y) = \lambda y)$$

$$A^TA = VDV^T$$

$$AA^T = UDU^T$$

$$A^TA = \underbrace{(U\Sigma V^T)^T}_{A^T} \underbrace{(U\Sigma V^T)}_{A}$$

$(AB)^T = B^TA^T$

$$= V\Sigma^T U^T U\Sigma V^T$$

$$= V\Sigma^T\Sigma V^T$$

$$AA^T = (U\Sigma V^T)(U\Sigma V^T)^T$$

$$= U\Sigma V^T V\Sigma^T U^T$$

$$= U\Sigma^T\Sigma U^T$$

$$= U\Sigma^2 U^T$$

U - formed by eigenvectors of AA^T

V - formed by " " A^TA

SVD&20(1).pdf

Rank of matrix
Full rank matrix

Date
Page

Singular values → square roots of eigenvalues of ATA.

Singular vectors:-
Consider rows of A as n points in d dimensions. Consider best fit line through origin. Let v be unit vector along the line.

The length of projection of a_i (ith row of A) onto v is $|a_i \cdot v|$

∴ Sum of length of squared projections is $\|Av\|^2$

$$\sum \beta^2 = \sum_{i=1}^{n} (a_i v)^2$$

→ write in terms of matrix A

Maximizing $\|Av\|_2^2$ is best fit line.

First singular vector of A

$$v_1 = \arg\max_{\|v\|_2 = 1} \|Av\|_2$$

$$\sigma_1(A) = \|Av_1\|_2$$

Why, $v_2 = \arg\max_{v \perp v_1, \|v\|_2 = 1} \|Av_2\|_2$

Similarly find $v_3, v_4, \ldots, v_d$

$\sigma_2(A) = \|A v_2\|_2$

$A \in R^{n \times d}$

Imp result:- Let A be $n \times d$ matrix where $v_1, \ldots, v_r$ are singular vectors. For $1 \leq k \leq r$, let V_k be subspace spanned by $V: v_1, \ldots, v_k$. then for each k