__/__/____

ADVANCED MACHINE LEARNING

9/1/23

⇒ well known conferences for ML

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

↓

Reproduce the results in papers.

(1) A Statistical Learning framework.

Domain set - x

Label set - y.

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

from $x \times y$.

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

(Source: 2nd book - Learning from Data)

(2) Data generation model

Prob. dist. over x is D.

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

(3) 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

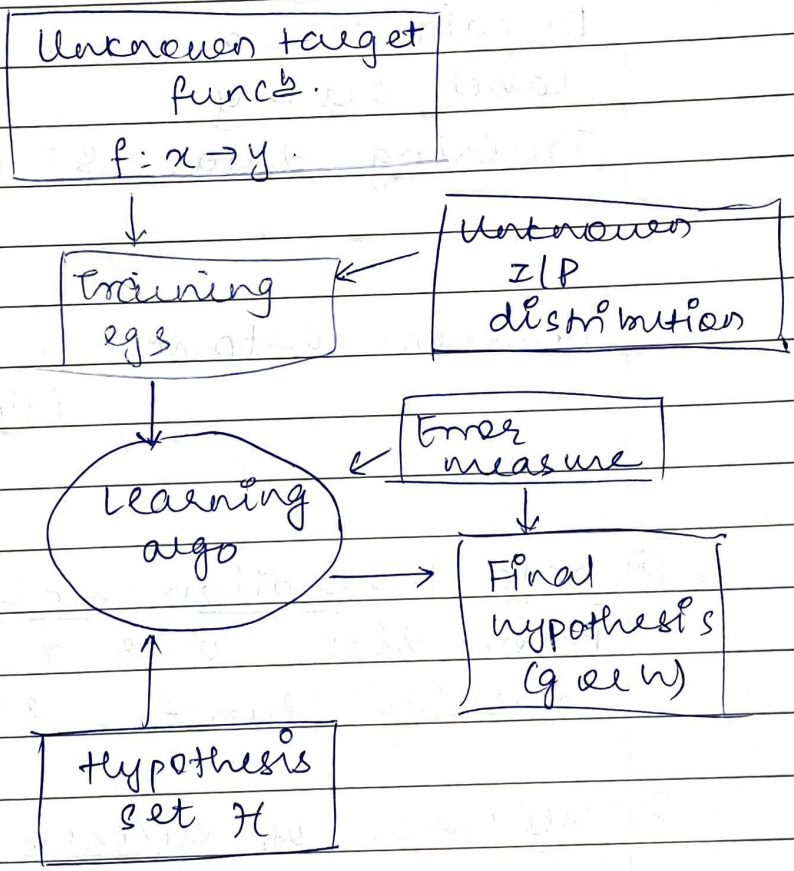
fixed → f

(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 → u

" green → $1-u$

Problem → u is unknown

N - total no. of marbles.

V - proportion of red marbles out of N

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

② **Hoeffding's Inequality**

- type of chernoff Bound.

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

V: empirical mean; u: original mean; δ: $2e^{-2\varepsilon^2 N}$

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

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

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

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

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

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

$\mathcal{H}$ – hypothesis space

(M) PAC learning → Probably Approximately correct

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

→ ~~When running the learning algo.~~

(M) Defns

Dichotomies:

$$\mathcal{H}(x_1, \dots, x_N) = \{ h(x_1), \dots, h(x_N) \mid h \in \mathcal{H} \}$$

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 $M = \infty$?

(M) Growth function – for a hypothesis set $\mathcal{H}$ is

$$m_{\mathcal{H}}(N) = \max_{x_1 \dots x_N} | \mathcal{H}(x_1, x_2, \dots, x_N) |$$

$$m_{\mathcal{H}}(N) \leq 2^N$$

(M) Shattering: If $m_{\mathcal{H}}(N) = 2^N$. (hypothesis set can generate all possible dichotomies)

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

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

② Project

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

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.

- Hoeffding's Inequality Works:-

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

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

→ 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)

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

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

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

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

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

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

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

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

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

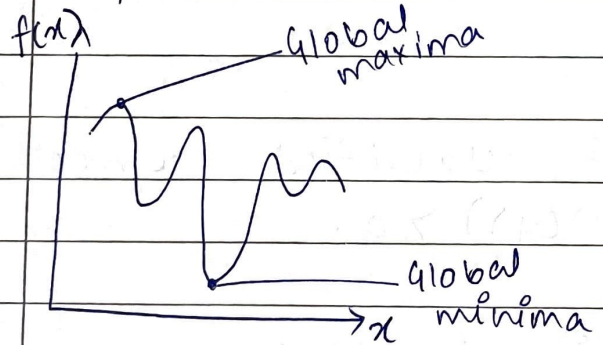
- Continuous
  - Non convex
  - Convex
- Combinational
  - Submodularity
  - Subset Selection

Gradient Descent → move in direction of -ve gradient.

[Optimization for ML – Nisharth Veshma]

→ 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^n, 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

ii) 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 $\forall x, y \in X$

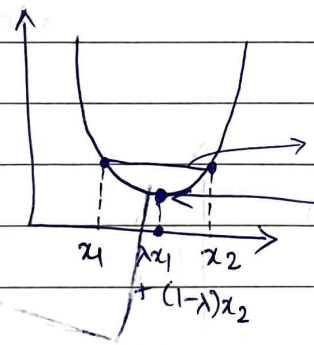
$\lambda x + (1-\lambda) y \in X \;, \forall 0 \leq \lambda \leq 1$

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

$R^n \subseteq X$

$\forall x_1, x_2$

$f(\lambda x_1 + (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$ — 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

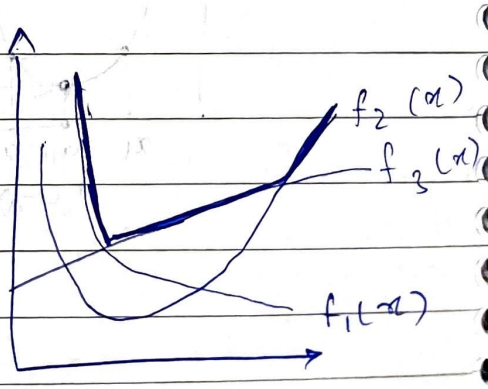
Egs
① $w^T x + b$ on R^n — convex
② -ve log likelihood func^n — "
③ quadratic func^n → $f(x) = \left(\frac{1}{2} x^T A x + b^T x + C\right)$ on R^n (A is PSD).

→ 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

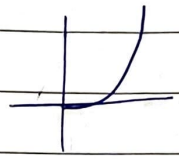
$$f(x) = \log \sum_{i=1}^{n} \exp(x_i)$$

⑤ Convex functions by composition

$h(x) \rightarrow$ h is convex

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

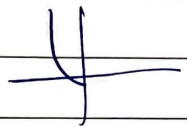
└ convex + non decreasing



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

└ concave

└ convex + non increasing



8/2/23 AML

→ Unconstrained optimization
→ Lagrange multipliers

⊛ constrained optimization problem.

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

$$\min \; f_0(x)$$
$$\text{such that } f_i(x) \leq 0 \quad i \in [1, m]$$
$$f_j(x) = 0 \quad j \in [1, l]$$

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

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

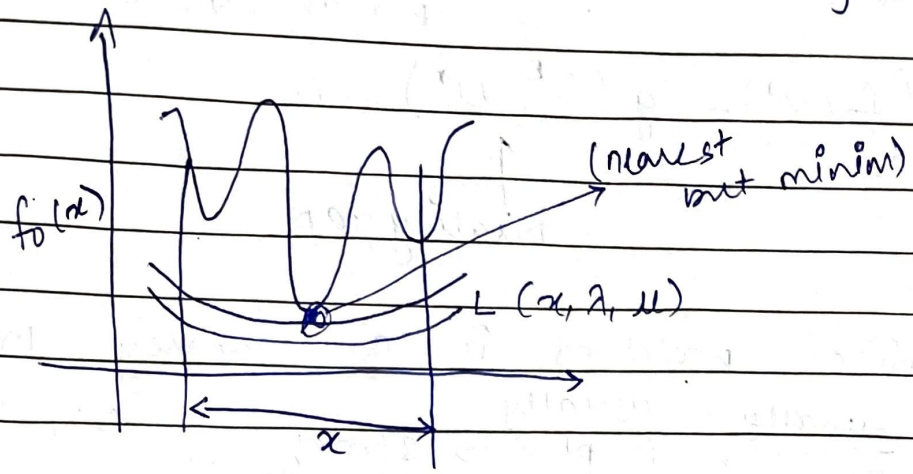
$$\|Ax - b\|_2^2 \text{ s.t. } \|x\|_2^2 \leq R$$

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

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

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

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

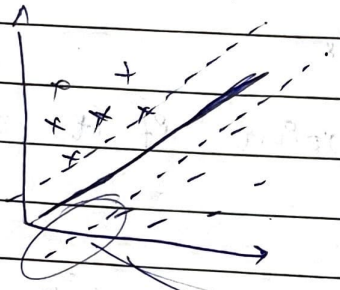
Optimization book

Scribe notes

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

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

$$\Rightarrow y_i \geq -1 + \xi_i$$

→ allows some misclassifications

Date
Page

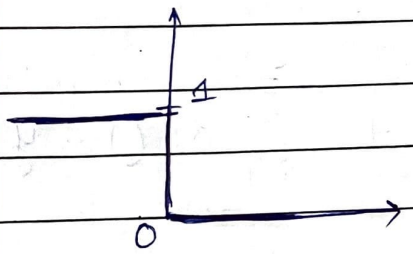
⇒ Hinge Loss ?

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

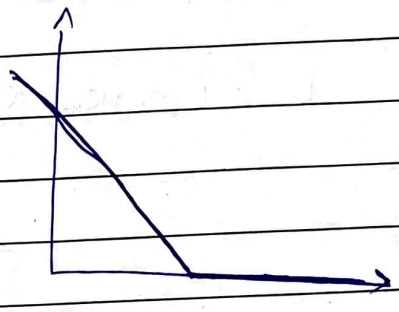
$0 < 1 - y_i (w^T x_i + w_0)$

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

→ 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 those 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

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

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

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

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

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

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

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

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

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

↑ 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 →

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

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

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

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

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

$\uparrow$

$\mathbb{R}^d$

$x_i \Rightarrow \Phi$

$z_i = \Phi(x_i)$

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

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

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

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

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

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

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

$\Rightarrow$ 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

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)

(dep 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

Date
Page

→ (orthonormal?) → k vectors,

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

→ (difference b/w orthogonal & orthonormal)

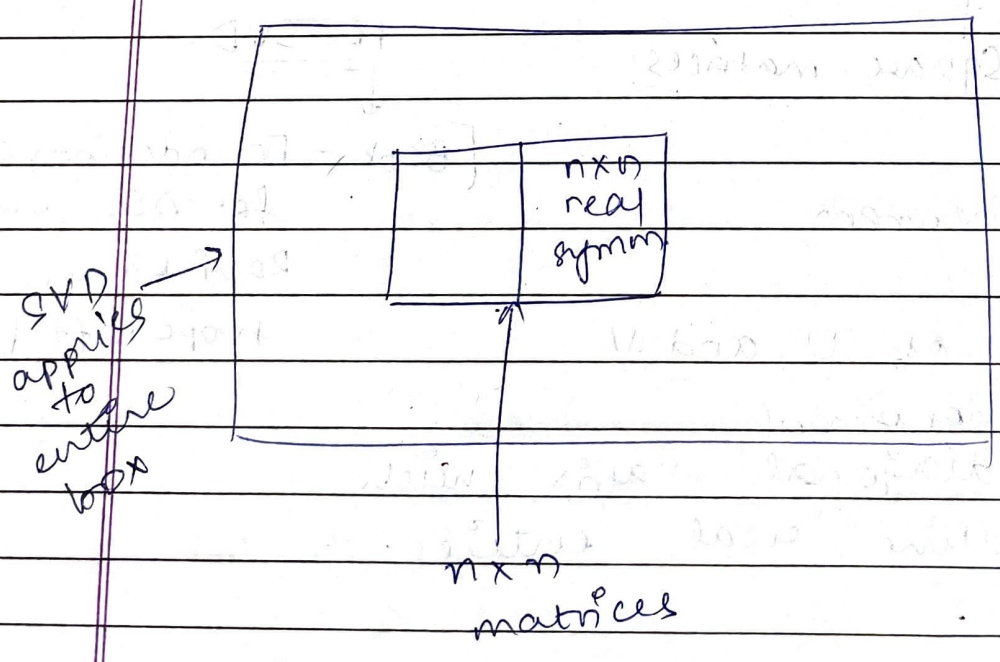
$x_i^T x_i = 1 \quad \forall \; i = i$

→ 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 $A - \lambda I$.

$AM(\lambda) \geq GM(\lambda)$

⇒ 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

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

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

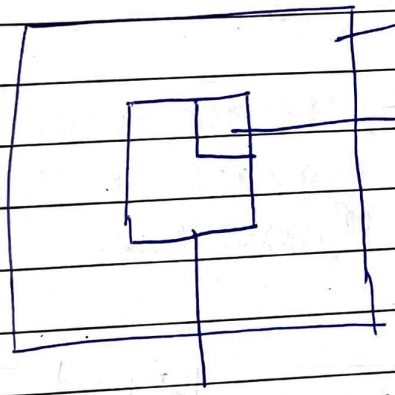
Diagonal matrix ≥ 0

(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 Hopcroft]

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

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

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

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

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

Hoffman, Kannan – book

Date / Page — chapter 4

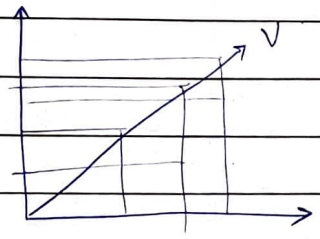
$\sigma_2(A) = \|A v_2\|_2$

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

Imp result:- Let A be n×d matrix where $v_1, \ldots, v_r$ are singular vectors. For $1 \le k \le r$, let V_k be subspace spanned by $v_1, \ldots, v_k$, then for each k, V_k is best fit k-dimensional subspace for A.

23/2/23

Ex: How many singular vectors orthogonal? = rank of matrix



$V = \{v_1, v_2\}$

$\text{span}_{\mathbb{R}}(V) = \{z \mid z = \alpha_1 v_1 + \alpha_2 v_2\}$

$V_k = \text{span}(v_1, \ldots, v_k)$

Define vectors $u_i = \frac{1}{\sigma_i(A)} A v_i$ ($n \times 1$)

$u_1, u_2, \ldots, u_d$ are called left singular vectors of A.

$\therefore A = \sum_{i=1}^{d} \sigma_i u_i v_i^T$

scalars; $n \times 1$; $1 \times d$

$A^{n \times d} = U_{n \times d} S_{d \times d} V^T_{d \times d}$

$n \times d$ (full rank) $n \ge d$

matrix formed by 1 outer product – rank 1

$Av = \sigma u$ (any Av is scaled version of u)

→ Induced Norm

→ General matrix Norm

→ Frobenius Norm :- $\|A\|_F = \sqrt{\sum_{j,k} a_{jk}^2} = \sqrt{\sum_{i=1}^{\text{rank}} \sigma_i^2}$ (singular values)

Spectral Norm (2 norm) - highest singular value of matrix

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

$\sum \sigma_i^2(A) = \|A\|_F^2$

→ Frobenius Norm used for low rank approximation.

⊛ Low-Rank Approximation Problem

$\min_{\tilde{A} \in R^{n \times d}} \|A - \tilde{A}\|_F^2$ s.t. $\text{rank}(\tilde{A}) = k$, $k < d$

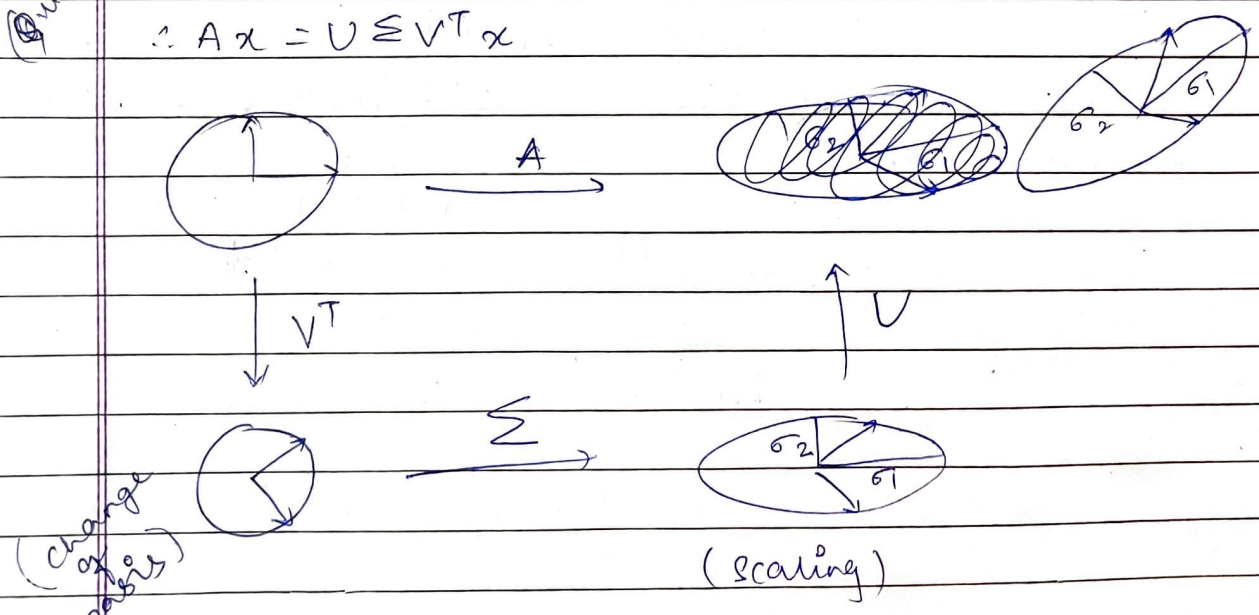
Best $\tilde{A}$

$\tilde{A} = \sum_{i=1}^{k} \sigma_i u_i v_i^T$

⊛ Geometric interpretation

$A = U \Sigma V^T$

$\therefore Ax = U \Sigma V^T x$



Date
Page

24/2/25 Related work.

(10) The Class Imbalance Problem

→ No. of obsn in one class label << no. of obsn in other class label

Issue →
- Conventional classifiers designed to optimize accuracy
- Biases performance towards majority class.
- More pronounced the imbalance, more pronounced is the issue
-

(Accuracy not correct measure in this case)

Solns
① Random undersampling (of majority class).

adv:- simple, efficient
disadv:- - throws away lot of data
- not representative of test data

② Random oversampling (copies of minority class)
adv :- - no data loss
- perform better than undersampling (empirically)
disadv :- prone to overfitting

→ Ensemble learning → Bagging — Bootstrapping (with replacement)
→ Boosting

(3) SMOTE: Synthetic Minority oversampling technique:-

- Avoid overfitting due to exact replicas of minority class samples.
- Subset of minority class is taken
- New synthetic data samples similar to subset are created & added.

⇒ Others techniques/solns

- Penalization based models
- Class weight based models
- Try to solve for different evaluation measure
- Ensemble based models

⇒ Parameter estimation (X, Y)

Discriminative	Generative
Model → $P(Y\|X)$	Model $P(X, Y)$

→ Task of ML → Prediction
→ Inference → (find data patterns)

Frequentist approach → run exp. multiple times & come to soln
Bayesian approach → uses prior prob.

$P(\theta|D) \propto P(D|\theta) \cdot P(\theta)$ MAP
posterior ↓ likelihood ↓ Prior

MLE: ← tries to maximize likelihood

θ = argmax

Bayesian

27/2/2023

$P(\theta|D,\alpha) \propto P(D|\theta,\alpha) \cdot P(\theta|\alpha)$

Posterior — likelihood — prior

Prob $\in [0,1]$

≥ 1

MLE tries to maximize this.

↳ Point estimates → MAP, MLE(?)

$P(\theta | \{(x_1,y_1),(x_2,y_2)\cdots(x_n,y_n)\}, \alpha)$

$\propto P\{(y_1,y_2\cdots y_n) | (x_1,x_2,\cdots,x_n), \theta, \alpha)$

$\cdot P(\theta|\alpha)$

$P(y_1,y_2,\cdots y_n | x_1,x_2,\cdots,x_n,\theta)$

$= L(\theta)$

$= \prod_{i=1}^{n} P(y_i|x_i,\theta)$ [θ = parameter]

→ conjugate prior (?)

$\theta_{MLE} = \underset{\theta}{\text{argmax}} \prod_{i=1}^{n} P(y_i|x_i,\theta)$

log likelihood → why? → bcoz log is monotonically ↑ing function & applying log converts it to summation.

$= \underset{\theta}{\text{argmax}} \sum_{i=1}^{n} \log(P(y_i|x_i,\theta))$

$= \underset{\theta}{\text{argmin}} \sum_{i=1}^{n} -\log(P(y_i|x_i,\theta))$ ← -ve log likelihood

$$l(\theta; (x,y)) = -\log P(y|x,\theta)$$ (loss on single point)

Regression cases

$$l_{reg}(\theta, (x,y)) = (\theta^T x - y)^2$$ (MSE)

$$P(y_i^{(i)} | x_i^{(i)}, (w, \beta)) = \mathcal{N}(y_i^{(i)} | w^T x_i^{(i)}, \frac{1}{\beta})$$ (σ^2)

θ (Normal distribution)

$$\varepsilon \sim \mathcal{N}(0, \frac{1}{\beta})$$ (error)

$$\text{Likelihood} = \prod_{i=1}^{n} P(y_i^{(i)} | x_i^{(i)}, w, \beta)$$

$$= \prod_{i=1}^{n} \mathcal{N}(y_i^{(i)} | w^T x_i^{(i)}, \frac{1}{\beta})$$

$$= \prod_{i=1}^{n} \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\left\{-\left(\frac{1}{2}\beta(y_i^{(i)} - w^T x_i^{(i)})\right)^2\right.$$

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \sim \mathcal{N}(\mu_1, \sigma_1^2) \quad \sim \mathcal{N}(\mu_n, \sigma_n^2)$$

Mean vector, $\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}$

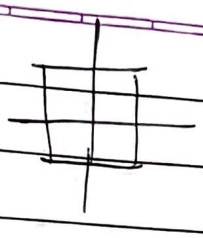
Variance-covariance matrix $= \Sigma$

Multivariate Gaussian

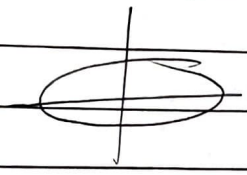
$$f_X(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(\frac{-1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

Date
Page

max norm →



Mahalanobis distance - geometrical concept



$$\max_{(w,\beta)} \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}} \sqrt{\beta} \exp\left(\frac{-1}{2} \beta \left(y_i - w^T x_i \right)^2 \right) \right)$$

$$-\text{ve log likelihood} = \sum_{i=1}^{n} \left(-\log \sqrt{\beta} + \log \sqrt{2\pi} + \frac{1}{2} \beta \left(y_i - w^T x_i \right)^2 \right)$$

min -ve log likelihood

$$\frac{\partial L(w,\beta)}{\partial w} = 0 \Rightarrow \sum_{i=1}^{n} \frac{-1}{2} \beta \cdot 2 \left(y_i - w^T x_i \right) \cdot x_i = 0$$

$$\frac{\partial L(w,\beta)}{\partial \beta} = 0 \Rightarrow$$

$$\sum_{i=1}^{n} \left[\frac{-1}{2\beta} + \frac{1}{2} \left(y_i - w^T x_i \right)^2 \right] = 0$$

$$w = (X^T X)^{-1} X^T Y$$

$$\frac{1}{\beta^*} = \frac{1}{n} \sum \left(y_i - w^T x_i \right)^2$$

← unexplained variance

9/3/23 AML

$P(y_i | x_i, w, \beta)$

$P(y_i|x_i)$ sampled from normal dist.

$P(y_i|x_i) \sim N(w^T x_i, \frac{1}{\beta})$ [$y_i = w^T x_i + \varepsilon$]

ε → Gaussian noise with $N(0, \frac{1}{\beta})$

→ -ve log likelihood

$\frac{\partial L(w,\beta)}{\partial w} = 0 \Rightarrow w_{MLE} = (X^T X)^{-1} X^T Y$

$\frac{\partial L(w,\beta)}{\beta} = 0 \Rightarrow \beta = \frac{1}{n} \sum (y_i - w^T x_i)^2$

↓ (unexplained variance)

(Frequentist approach)

→ For RV, we find expectation & variance.

⇒ Gauss-Markow theorem :-

w_{MLE} is BLUE → Estimate

Best → Linear → Unbiased

(Best in terms of variance)

$E_D[w_{MLE}] = w^*$

$E_D[(X^T X)^{-1} X^T Y] = (X^T X)^{-1} X^T E_D[Y]$ — (1)

(locoz randomness only in Y)

$Y = Xw + \varepsilon$ $\quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \sim N(0, \frac{1}{\beta})$

$E_D[Y] = E[Xw] + E[\varepsilon]$ (due to linearity of expectation)

$= E[Xw] + 0 \quad (\varepsilon \sim N(0, \beta))$

Replacing in (1)

$$E_D[(X^TX)^{-1}X^TY] = (X^TX)^{-1}X^T E_D[Y]$$
$$= \underline{(X^TX)^{-1}X^TX}\, w$$
$$= I\, w$$
$$= w.$$

$\therefore E_D[w_{MLE}] = w^*$ (∵ Its unbiased dist)

$$E_D\left[\frac{1}{\beta_{MLE}}\right] = E_D\left[\frac{1}{n}(Xw_{MLE} - Y)^T(Xw_{MLE} - Y)\right]$$
$$= \frac{1}{n} E_D\left[\underbrace{w_{MLE}^T}_{\downarrow} \underline{X^TX}\, \underline{w_{MLE}} - 2w_{MLE}^T X^TY + Y^TY\right]$$
$$= \frac{1}{n} E_D\left[Y^TX(X^TX)^{-1}\cdot(X^TX)(X^TX)^{-1}X^TY - 2\underbrace{Y^TX(X^TX)^{-1}X^TY}_{I} + Y^TY\right]$$
$$= \frac{1}{n} E_D\left[-Y^TX(X^TX)^{-1}X^TY + Y^TY\right]$$
$$= \frac{1}{n} E_D\left[Y^T\underbrace{(I - X(X^TX)^{-1}X^T)}_{Z}Y\right]$$

$(x^TAx + x^TBx = x^T(A+B)x.$

$$= \frac{1}{n} E_D[Y^TZY]$$
$$= \frac{1}{n} E_D[(Xw+\varepsilon)^TZ(Xw+\varepsilon)]$$
$$= \frac{1}{n} E_D[w^TX^TZXw + \varepsilon^TZXw + w^TX^TZ\varepsilon + \varepsilon^TZ\varepsilon]$$
($\varepsilon^TZXw = 0$, $w^TX^TZ\varepsilon = 0$)
$$= \frac{1}{n}\left[w^TX^TZXw + 0 + 0 + \sum_{i,j} Z_{i,j}\, E[\varepsilon_i, \varepsilon_j]\right] \quad (2)$$

(Z is $n \times n$ matrix)

$$\varepsilon^T z \varepsilon = [\varepsilon_1 \ \varepsilon_2 \ \ldots \ \varepsilon_n] \begin{bmatrix} z_{11} & z_{12} & \cdots \\ \vdots & & \\ & & z_{nn} \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$= \sum_{i,j} z_{ij} \cdot \varepsilon_i \varepsilon_j$$

$E[\varepsilon_i, \varepsilon_j] = 0$ when $i \neq j$

$$\sum_i \sum_j Pr(\varepsilon_i \cdot \varepsilon_j) \cdot \varepsilon_i \cdot \varepsilon_j$$

$$= Pr(\varepsilon_i) Pr(\varepsilon_j)$$

$$\sum_i Pr(\varepsilon_i) \cdot \sum_j Pr(\varepsilon_j)$$

From (2), (contd)

$$= \frac{1}{n} \left[w^T x^T z x w + 0 + 0 + \frac{1}{\beta} \sum_{i=1}^{n} diag(z) \right]$$

$$= \frac{1}{n} \left[w^T x^T \left(I - X(x^T x)^{-1} x^T \right) X w + \frac{1}{\beta} \sum_{i=1}^{n} diag(z) \right]$$

$$= \frac{1}{n} \left[w^T x^T x w - w^T x^T x (x^T x)^{-1} x w + \frac{1}{\beta} \sum_{i=1}^{n} diag(z) \right]$$

$$= \frac{1}{n} \left(\frac{1}{\beta} \sum_{i=1}^{n} diag(z) \right]$$

$$= \frac{1}{n} \left[\frac{1}{\beta} tr(z) \right]$$

(sum of diagonals is trace)

$$= \frac{1}{n}\left[\frac{1}{\beta}\left(n - tr\left(X^T X (X^T X)^{-1} X^T\right)\right)\right]$$

$$= \frac{1}{n}\left[\frac{1}{\beta}\cdot\left(n - tr\left(X^T X (X^T X)^{-1}\right)\right)\right]$$

→ d dimension

$$= \frac{1}{n}\left[\frac{1}{\beta}(n-d)\right]$$

(trace of identity matrix $d \times d$ $= d$]

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & \cdots & \cdots & 1 \end{bmatrix}$$

$$E_D\left[\frac{1}{\beta_{MLE}}\right] = \left(1 - \frac{d}{n}\right)\frac{1}{\beta}$$

ADVANCED ML (CLASS NOTES)

classmate
Date 13/3/23
Page

(10) MLE for linear regression

$$E[w_{MLE}] = w.$$

$$E\left[\frac{1}{\beta_{MLE}}\right] = \left(1 - \frac{d}{n}\right)\frac{1}{\beta}$$

Gauss-Markov theorem
↓
w_{MLE} is BLUE
↓
Best Linear unbiased estimator

w_{MLE} → vector

$$var[w_{MLE}] < var[\tilde{w}]$$

where $\tilde{w}$ is any other linear estimate

$A - B \geq 0$ means $A - B$ is the semi-definite

↳ $var[w_{MLE}] - var[\tilde{w}] \geq 0$

then, $var[w_{MLE}]$ is best variance.

$var(w_{MLE})$

$= E[(w_{MLE} - w)(w_{MLE} - w)^T]$

$= E[((X^TX)^{-1}X^TY - w)((X^TX)^{-1}X^TY - w)^T]$

$= E_D[((X^TX)^{-1}X^T(Xw + \varepsilon) - w)((X^TX)^{-1}X^T(Xw + \varepsilon) - w)^T]$

$= E_D[w + (X^TX)^{-1}X^T\varepsilon - w]$

$var(X) = E[(X - \mu)(X - \mu)^T]$

← $Y = Xw + \varepsilon$

$$= E_D\left[\left((X^TX)^{-1}X^T\varepsilon\right)\left((X^TX)^{-1}X^T\varepsilon\right)^T\right]$$

$$= E_D\left[(X^TX)^{-1}X^T(\varepsilon\varepsilon^T)X(X^TX)^{-1}\right]$$

$$= (X^TX)^{-1}X^T E(\varepsilon\varepsilon^T) X (X^TX)^{-1} \qquad (AB)^T = B^TA^T$$

$$= \underline{(X^TX)^{-1}X^T}\ \frac{1}{\beta}\ I_{n\times n}\ \underline{X(X^TX)^{-1}} \qquad \left(\varepsilon \sim N\left(0, \frac{1}{\beta}\right)\right)$$

$$\boxed{Var(w_{MLE}) = \frac{1}{\beta}(X^TX)^{-1}} \qquad \left[(X^TX)^{-1}X^TX = I\right]$$

↖ var–cov matrix

⇒ Proving the 'Best' part in BLUE.

T.P :- $-Var[w_{MLE}] + Var[\tilde{w}] \geq 0$.

$\tilde{w} = AY = A(Xw+\varepsilon) = AXw + A\varepsilon$

$E(AXw + A\varepsilon) = w \quad [\because \text{unbiased}]$

$\therefore AXw = 1w$

↑ matrix ↑ scalar (characteristic eq^n)

⇒ w is eigen vectors of Ax & corresponding eigenvalue is 1.

$Var[\tilde{w}] = E[(\tilde{w} - E(\tilde{w}))(\tilde{w} - E(\tilde{w}))^T]$

$\tilde{w} = AXw + A\varepsilon$ → $= E[(\tilde{w} - AXw)(\tilde{w} - AXw)^T]$

$= E[(A\varepsilon)(A\varepsilon)^T]$

$= E[A\varepsilon\varepsilon^TA^T]$

$= \frac{1}{\beta}(AA^T)$

$w_{MLE} = \underbrace{(X^TX)^{-1}X^T}_{\downarrow M} y$

$\text{var}(\tilde{w}) - \text{var}(w_{MLE})$ (should be PSD)

$= \frac{1}{\beta}\left(AA^T - (X^TX)^{-1}\right)$ — (1)

$\left(A = (X^TX)^{-1}X^T + B\right)$ (B')

↓ multiply X on both sides,

$AX = (X^TX)^{-1}X^TX + BX.$

$\Rightarrow (AX - I) = BX$ (prove that $BX = 0$.)

$(AX - I)w = 0$

$\Rightarrow BXw = 0$

$\text{var}[\tilde{w}] - \text{var}(w_{MLE})$

$= \frac{1}{\beta}[BB^T]$

(from (1), if we replace A, we get this)

(BB^T is the semi-def.

↓ How?

↓

($x^TAx \geq 0$ ++)

$BB^T \Rightarrow$

x^TBB^Tx

$(B^Tx)^TB^Tx$

$= \|B^Tx\|_2^2$

(which is always ≥ 0)

∴ PSD.

$((X^TX)^{-1}X^TX)((X^TX)^{-1}X^TX)$

$-(X^TX)^{-1})$

Difference between w_{MLE} & w_{MAP} ?

Posterior ∝ prior × likelihood

$$P\left(w \mid \underbrace{(y_1 \ldots y_n)(x_1 \ldots x_n)}_{\mu_0, \Sigma_0, \frac{1}{\beta}}, \frac{1}{\beta}\right)$$

$$\propto P\left[(y_1, y_2, \ldots, y_n) \mid w_1, \frac{1}{\beta}, (x_1, \ldots, x_n)\right] \cdot P(w_1 \mid \mu_0, \Sigma_0)$$

↑ prior.

16/3/2023

$$y_i = w^T x_i + \varepsilon_i$$

w: $N(\mu_0, \Sigma_0)$; ε_i: $N\left(0, \frac{1}{\beta}\right)$

$$P\left(w \mid \underbrace{(y_1 \ldots y_n)(x_1 \ldots x_n)}_{D\ \mu_0, \Sigma_0, \frac{1}{\beta}}\right) \propto P\left(y_1 \ldots y_n \mid w, \frac{1}{\beta}, (x_1 \ldots x_n)\right) \cdot P(w \mid \mu_0, \Sigma_0)$$

$$P\left(w \mid (y_1 \ldots y_n)\ (x_1 \ldots x_n), \Sigma_0, \mu_0, \frac{1}{\beta}\right)$$

$$\propto \left[\prod_{i=1}^{n} P\left(y_i \mid w, \frac{1}{\beta}, x_i\right)\right] \cdot P\left[(w \mid \Sigma_0, \mu_0)\right]$$

$$\left[\prod_{i=1}^{n} N\left(y_i \mid w^T x_i, \frac{1}{\beta}\right)\right] \cdot P\left[w \mid (\Sigma_0, \mu_0)\right]$$

$$y_i = w^T x_i + \varepsilon_i$$

$\varepsilon_i \sim N\left(0, \frac{1}{\beta}\right)$

$w \sim N(\mu_0, \Sigma_0)$

Multivariate gaussian

$$P\left(w \mid D, \Sigma_0, \mu_0, \frac{1}{\beta}\right) = \prod_{i=1}^{n} \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\left(\frac{-1}{2}(y_i - w^T x_i)^2 \cdot \beta\right) \cdot \frac{1}{(2\pi)^{d/2} \cdot |\Sigma_0|} \cdot \exp\left(\frac{-1}{2}\right.$$

($|\Sigma_0|$ ← covariance matrix)

$\left.(w-\mu_0)^T \Sigma_0^{-1} (w-\mu_0)\right]$

$$= -\log P\left(w \mid D, \Sigma_0, \mu_0, \frac{1}{\beta}\right)$$

Posterior

$$= -n \log \sqrt{\beta} + n \log \sqrt{2\pi} + \frac{\beta}{2} \sum_{i=1}^{n} (y_i - w^T x_i) + \frac{1}{2} \log |\Sigma_0| + d \log \sqrt{2\pi} + \frac{1}{2} (w - \mu_0)^T \Sigma^{-1} (w - \mu_0)$$

$$= \log \left(\frac{1}{2\pi}\right)^{d/2} + \log \left(\frac{1}{|\Sigma_0|}\right)^{-1/2}$$

"Matrix cookbook" - book

~~$\log P(w | \Sigma_0, \mu_0, \frac{1}{\beta})$~~

$$\frac{\partial}{\partial w}\left[-\log P\left[w \mid D, \Sigma_0, \mu_0, \frac{1}{\beta}\right)\right] = 0$$

$$= \frac{\partial}{\partial w}\left[\frac{\beta}{2}(Xw - Y)^T(Xw - Y)\right] + \frac{\partial}{\partial w}\left[\frac{1}{2}(w - \mu_0)^T \Sigma^{-1}(w - \mu_0)\right] = 0$$

$$\therefore \frac{\beta}{2}\left[2(X^TX)w - 2X^TY\right] + \frac{1}{2}\left[2\Sigma_0^{-1}w - 2\Sigma_0^{-1}\mu_0\right]$$

$$(\beta(X^TX) + (\Sigma_0)^{-1})w - \beta X^TY - \Sigma_0^{-1}\mu_0 = 0$$

$$\therefore w_{MAP} = (\beta(X^TX) + \Sigma_0)^{-1})^{-1} \cdot (\beta X^TY + \Sigma_0^{-1}\mu_0)$$

prior is infinitely broad, then we get exact MLE estimate

If $\mu_0 = 0$, spherical prior, $\Sigma_0^{-1} = \frac{1}{C}I$

$$(\beta(X^TX) + \frac{1}{C}I)^{-1} \cdot (\beta X^TY)$$

$$(X^TX)^{-1}X^TY + \frac{\beta C}{\beta} X^TY$$

$$\left((X^T X) + \frac{1}{BC}\right)^{-1} X^T Y$$

(like Ridge regression)

Takeaway => (treated as L2 reg. problem)

20/3/23

Ridge Regression

$\|Ax-b\|^2 + \lambda x^T x \Rightarrow (A^TA + \lambda I)^{-1} A^T y$

Spherical prior -

$E[w_{MAP}] = E[[[X^TX + \lambda I)^{-1} X^T y]$ $(y = Xw+\varepsilon)$

$= E[(X^TX + \lambda I)^{-1} (Xw+\varepsilon)]$

$= E[(X^TX + \lambda I)^{-1} (X^TXw + X^T\varepsilon)]$

$= E[(X^TX + \lambda I)^{-1} (X^TXw) + (X^TX+\lambda I)^{-1} E(X^T\varepsilon)]$

$\mu_0 = 0$

$= \underbrace{(X^TX + \lambda I)^{-1} (X^TX w)}_{Z}$

$= ZXw$

Variance :-

$Var[w_{MAP}] = E[w_{MAP} - E(w_{MAP})]^T [w_{MAP} - E(w_{MAP})]$

$= E[Zy - ZXw)(Xy - ZXw)^T]$

$= E[E(y-Xw)(y-Xw)^T Z^T]$

$= Z\, E[\varepsilon\varepsilon^T] Z^T$

$= Z \frac{1}{\beta} I_{n\times n} Z^T$

$= \frac{1}{\beta} ZZ^T$

$= \frac{1}{\beta} (X^TX + \lambda I)^{-1} (X^TX) [X^TX+\lambda I)^{-1}]$ ← PD

SVD :- $X^TX = V\Sigma^2V^T$ $\quad (X^TX)^{-1} = V\,diag\left(\frac{1}{\sigma^2}\right)V^T$

[Orthonormal, V → Rotating]

$X^TX + \lambda I = V\Sigma^2V^T + \underline{\lambda VIV^T}$ → In place of I

$= V(\Sigma^2 + \lambda I)V^T$

$= V\,diag(\sigma_i^2 + \lambda)V^T$

$(X^TX + \lambda I)^{-1} = V\,diag\left(\frac{1}{\lambda + \sigma_i^2}\right)V^T$

$(X^TX + \lambda I)^{-1}X^TX = V\,diag\left(\frac{1}{\lambda + \sigma_i^2}\right)\frac{V^T(V\Sigma^2V^T)}{I}$

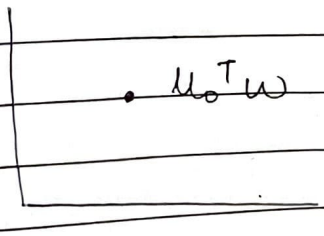
$= V\,diag\left(\frac{1}{\lambda + \sigma_i^2}\right)\Sigma_i^2V^T$ (as $\lambda \uparrow$, var $\downarrow \approx 0$)

Bias - Variance trade off.

n → no. of points
λ → Regularized parameter

For $n=0$ → WMLE doesn't exist.

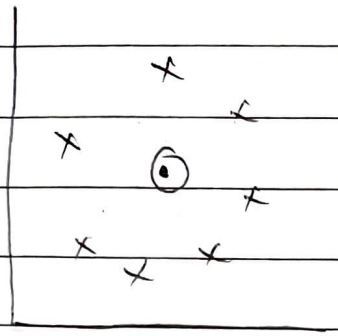
$n=0, \lambda=0$



$n=0, 0<\lambda<\infty$

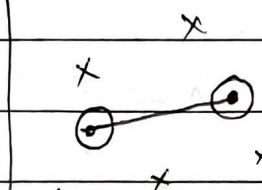
$n=0, \lambda \to \infty$

$0 < n < \infty, \lambda = 0$



$0 < n < \infty, 0 < \lambda < \infty$

Reality scenario



w_{MLE} & w^* close to each other

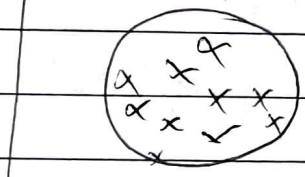
low bias, high variance.

$0 < n < \infty, \lambda \to \infty$

$n \to \infty, \lambda = 0$

$n \to \infty, 0 < \lambda < \infty$.

$n \to \infty, \lambda \to \infty$



~~For variance~~ it shrinks

(~~Question~~)

(#) Taxonomy of Info — Intents are in terms of these categories.

IR Domain - IA Select

q - query $C_{(q)}$ - set of categories of q

d - document $C_{(d)}$ - " " " " d

Assumptions :-

1) Taxonomy of info.

2) $p(c|q)$ is drawn (c = classifier)

$$\sum_{c \in C_{(q)}} p(c|q) \quad \text{— (1)} \rightarrow \text{Knowledge completeness}$$

23/3/23 ⑧ GMM:- Gaussian Mixture Models.

(Sir's notes → GMMs.pdf)

→ Gaussian distribution by itself may not be enough in modelling real datasets.

(Fourier transformation)

→ Sometimes a linear superposition of two or more Gaussians may capture data better. Such superpositions can be formulated as mixture distributions.

→ For a superposition of K Gaussians, we have,

$$p(x) = \sum_{k=1}^{K} w_k \, N(x \mid \mu_k, \Sigma_k)$$

Each $N(x \mid \mu_k, \Sigma_k)$ is called a component of the mixture. The parameters w_k are called ~~mixto~~ mixing coefficients.

$w = \{w_k\}$ $\mu = \{\mu_k\}$ $\Sigma = \{\Sigma_k\}$

$w_k \geq 0 \quad \forall k$

$\therefore 0 \leq w_k \leq 1$ can be thought of as probs.)

$$p(x) = \sum_{k=1}^{K} P(k)\, p(x|k) \qquad (w_k = p(k))$$

Think of $p(k)$ as prior of picking k^{th} component and $p(x|k) = N(x|\mu_k, \Sigma_k)$

i.e. prob. of x conditioned on k.

→ The posterior prob. $P(k|x)$ are given as

$$\gamma_k(x) = p(k|x)$$

$$= \frac{p(k)\, p(x|k)}{\sum_{l} p(l)\, p(x|l)} \qquad \text{(Bayes theorem)}$$

$$= \frac{w_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{l} w_l \mathcal{N}(x|\mu_l, \Sigma_l)}$$

→ Parameters of GMM are w, μ & Σ (rest part - refer sir's notes)

→ Topic Modelling - eg. of latent model

24/3/23

Now $\gamma(z_k)$ i.e. $P(z_k = 1 | x)$

$$= \frac{P(z_k=1)\, P(x|z_k=1)}{\sum_l P(z_l=1)\, P(x|z_l=1)}$$

$$= \frac{w_k N(x|\mu_k, \Sigma_k)}{\sum_l w_l N(x|\mu_l, \Sigma_l)}$$

→ Maximizing log likelihood of GMM is more complex problem than the case of single Gaussian. Because of summation over k inside log, log func2 does not act directly ....

EM for GMMs :- (Expectation Maximization)

↳ The log likelihood for GMM given data X is given by :-

$$\ln P(X|w, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} w_k N(x_n|\mu_k, \Sigma_k) \right\}$$

Deriving w.r.t. μ_k and equating to 0,

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{w_k N(x_n|\mu_k, \Sigma_k)}{\sum_j w_j N(x_n|\mu_j, \Sigma_j)}}_{\gamma(z_{nk})} \Sigma_k (x_n - \mu_k)$$

Multiplying by Σ_k^{-1},

$$\mu_k = \frac{1}{N_k} \cdot \sum_{n=1}^{N} \gamma(z_{nk})\, x_n$$

where, $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$

N_k = effective no. of points belonging to cluster k.

$$\Sigma_k = \frac{1}{N_k} \cdot \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T$$

Similarly,

Finally we need to maximize $\ln p(X|w, \mu, \Sigma)$ w.r.t. w_k

we consider,

$$\ln p(X|w, \mu, \Sigma) + \lambda\left(\sum_{k=1}^{K} w_k - 1\right)$$

which gives,

$$0 = \sum_{n=1}^{N} \frac{N(x_n|\mu_k, \Sigma_k)}{\sum_j w_j N(x_n|\mu_j, \Sigma_j)} + \lambda$$

Multiply both sides by w_k and sum over k, we get $\lambda = -N$

Finally, $w_k = \frac{N_k}{N}$

EM for gaussian mixtures:-

1. Initialize μ_k, Σ_k and w_k and evaluate initial value of log likelihood

2. E-step

$$\gamma(z_{nk}) = \frac{w_k \, N(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} w_j \, N(x_n | \mu_j, \Sigma_j)}$$

(Recalculate posterior)

3. $$\mu_k^{new} = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(z_{nk}) \, x_n$$

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \, (x_n - \mu_k^{new}) \, (x_n - \mu_k^{new})^T$$

$$w_k^{new} = \frac{N_k}{N}$$

where, $$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

4. Evaluate $\sum_{i=1}^{N} \ln \left\{ \sum_{k=1}^{K} w_k \, N(x_n | \mu_k, \Sigma_k) \right\}$.

Repeat 2-4 till convergence

Quiz next Monday → till GMMs. ↗ till here.