* Any - 8th Feb

* Constrained Optimization problem

$f_0 : R^n \longrightarrow R$

$$\min f_0(x) \Rightarrow \text{s.t.} \quad f_i(x) \leq 0 \quad i \in [1, m]$$
$$f_j(x) = 0 \quad j \in [1, l]$$

$\lambda_1, \ldots, \lambda_m \qquad \mu_1, \ldots, \mu_l$

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

$\lambda_i \geq 0 \qquad \underbrace{\lambda_i\text{'s}, \mu_j\text{'s}}$

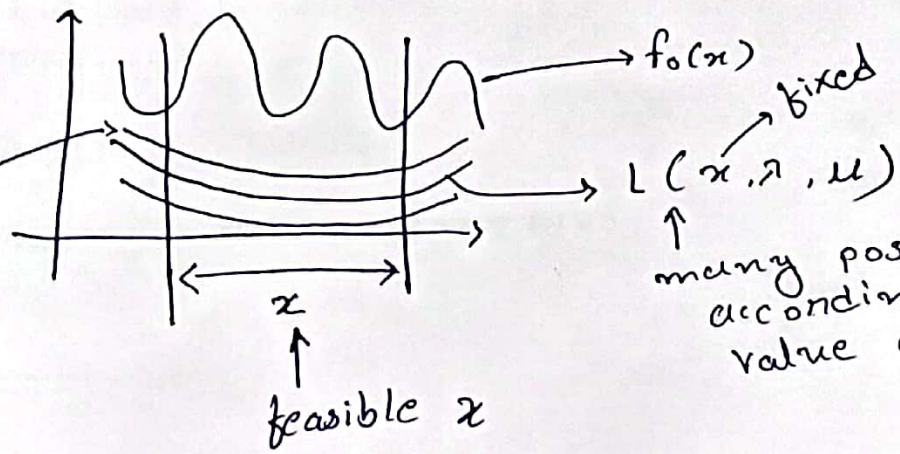
lagrangian multipliers &
Dual variable

$$L(x, \lambda, \mu) = f_0(x) + \underbrace{\sum_{i=1}^{m} \lambda_i f_i(x)}_{-ve} + \underbrace{\sum_{j=1}^{l} \mu_j f_j(x)}_{0}$$

~~which~~

x is feasible pt. $\rightarrow$ satisfies the constrained

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



get the fun^n which is near to $f_0(x)$ so we maximize it

$$\min_x L(x, \lambda, \mu) \quad \cancel{\text{s.t. } \lambda \geq 0} \leq f_0(x^*)$$
$$\text{s.t. } \lambda \geq 0$$

$$\max_{(\lambda, \mu) \in \text{Domain}} \left[\min_x L(x, \lambda, \mu) \right] \leq f_0(x^*)$$
$$\text{s.t. } \lambda \geq 0$$

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

dual fun^n

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

s.t. $\lambda \geq 0$

(λ^*, μ^*) are point of maxima

$g(\lambda, \mu) \leq \boxed{g(\lambda^*, \mu^*) \leq f_0(x^*)}$ → weak duality theorem

~~$g(\lambda^*, \mu^*) \leq$~~

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

↑ duality gap

Primal problem is convex problem $\xrightarrow{\text{usually}}$ (not always) constraints should satisfy some conditions (Regularity conditions)

& strong duality hold → $f_0(x^*) - g(\lambda^*, \mu^*) = 0$

* KKT Conditions :-

① Primal feasibility

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

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

$x^* \rightarrow$ pt. of primal optimal

$\lambda^*, \mu^* \rightarrow$ pt. of dual optimal

→ zero duality gap

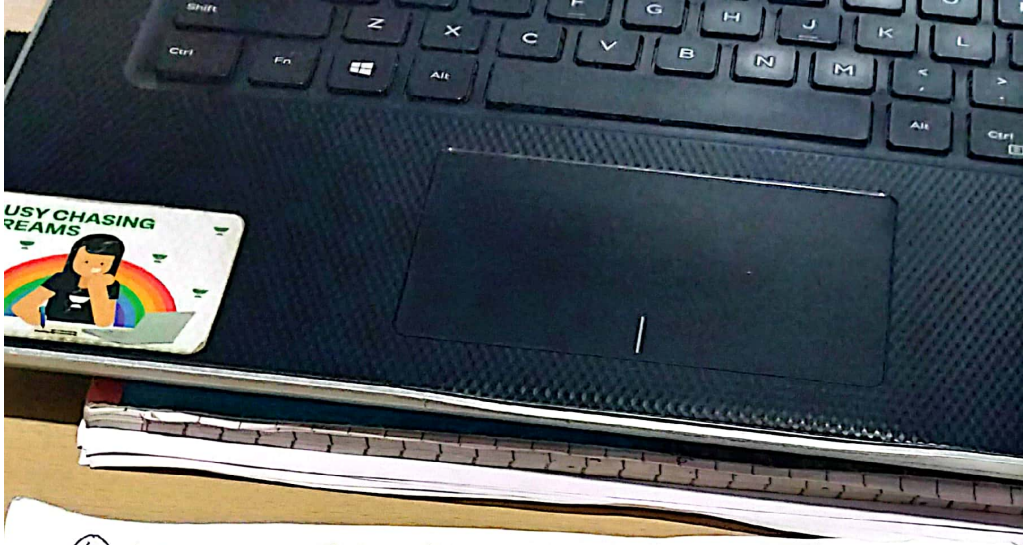
② dual feasibility

$\lambda^* \geq 0$

③ Complementary Slackness

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

$$L(x, \lambda, \mu) = f_0(x) + \underbrace{\sum_{i=1}^{m} \lambda_i f_i(x)}_{\text{also } 0 \text{ at optimum}} + \sum_{j=1}^{l} \mu_j f_j(x)$$



④ Lagrange Optimality

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

optimal

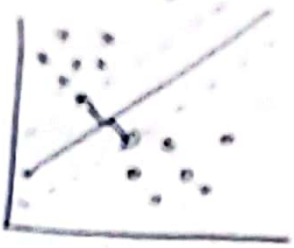
→ Any tuple (x, λ, μ) → is called a KKT point if it satisfies the KKT cond^n.

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

→ When you have zero duality gap (x^*, λ^*, μ^*) ⇓ KKT point.

CamScanner

* Hard Margin SVM



$$\min_{(\omega,\omega_0)} \frac{1}{2}\|\omega\|_2$$

s.t. $(\omega^T x_i + \omega_0)\, y_i \geq 1 \quad \forall i,\ i \in [1, n]$

$\Rightarrow$ For soft-margin SVM

$$(\omega^T x_i + \omega_0) \geq 1 - \xi_i$$

$$(\omega^T x_i + \omega_0) \leq -1 + \xi_i$$

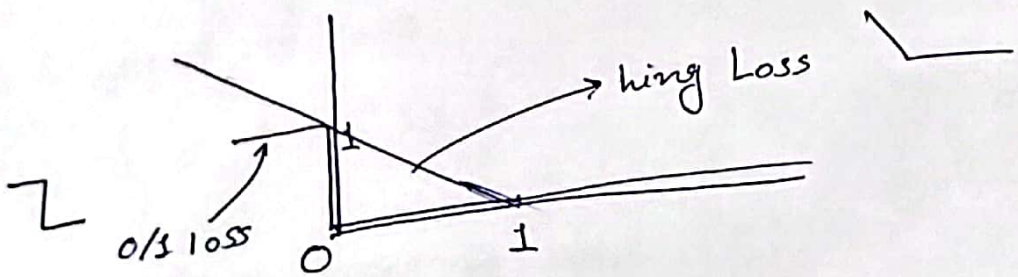
if point is misclassified

$$y_i(\omega^T x_i + \omega_0) < 1$$

$$0 < 1 - y_i(\omega^T x_i + \omega_0)$$

$\Rightarrow$ Hing Loss

$$L[(x, y), (\omega, \omega_0)] = \max\{0, 1 - y_i(\omega^T x_i + \omega_0)\}$$



$\rightarrow \frac{1}{2}\|\omega\|_2^2$

s.t. $y_i(\omega^T x_i + \omega_0) \geq 1 \quad \forall i \quad i \in [1, n]$

$\Rightarrow$ KKT conditions

1) Primal Feasibility : $1 - y_i(\omega^T x_i + \omega_0)\Big|_{\omega^*, \omega_0^*} \leq 0$

primal variables

$$L(\underbrace{\omega, \omega_0}_{x}, \underbrace{\alpha}_{\lambda}) = \underbrace{\cancel{1 - y_i(\omega^T x_i + \omega_0)}}_{\cancel{f(x)}} \underbrace{\frac{1}{2}\|\omega\|_2^2}_{f_0(x)} + \sum_{i=1}^{n} \alpha_i (1 - y_i(\omega^T x_i + \omega_0))$$

2. dual feasibility : α_i's ≥ 0 $\forall i$

3. Complementary slackness

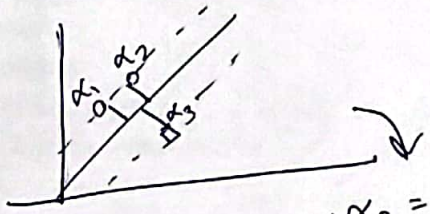
$$\alpha_i \left(1 - y_i(\omega^T x_i + \omega_0)\right) = 0$$

4. Lagrange Optimality

$$L(\omega, \omega_0, \alpha) = \underbrace{\frac{1}{2}\|\omega\|_2^2}_{\frac{1}{2}\omega^T\omega} + \sum_{i=1}^{n} \alpha_i \left(1 - y_i(\omega^T x_i + \omega_0)\right)$$

$$\nabla_\omega L(\omega, \omega_0, \alpha) = 0 \longrightarrow \omega + \sum_{i=1}^{n} \alpha_i(-y_i x_i) = 0 \Rightarrow \omega = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\nabla_{\omega_0} L(\omega, \omega_0, \alpha) = 0 \longrightarrow \sum_{i=1}^{n} \alpha_i(-y_i) = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$



$$\alpha_1 + \alpha_2 = \alpha_3$$

$$\sum_{i \in C^1} \alpha_i = \sum_{j \in C^-} \alpha_j$$

$\Rightarrow$ Support Vectors

$$\{x_i \mid \alpha_i^* > 0\}$$

$$\Rightarrow \quad g(\alpha) = \min_{\omega, \omega_0} L(\omega, \omega_0, \alpha)$$

$$\omega^* = \sum_{i=1}^{n} (\alpha_i y_i x_i) \quad \text{— (1)}$$

$$\omega_0^* \Rightarrow \sum \alpha_i y_i = 0$$

$$g(\alpha) = \frac{1}{2}(\omega^*)^T(\omega^*) + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i \omega^{*T} x_i - \underbrace{\sum_{i=1}^{n} \alpha_i y_i \omega_0^*}_{=0}$$

$$g(\alpha) = \frac{1}{2}(\omega^*)^T(\omega^*) + \sum_{i=1}^{m} \alpha_i - (\omega^*)^T \underbrace{\sum_{i=1}^{m} \alpha_i^* y_i x_i}_{\omega^* \text{ from}}$$

$$= \frac{1}{2}(\omega^*)^T(\omega^*) + \sum_{i=1}^{m} \alpha_i - (\omega^*)^T(\omega^*)$$

$$= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}(\omega^*)^T(\omega^*)$$

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

$$g(\alpha) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \underbrace{(\alpha_i \alpha_j)(y_i y_j)}_{\text{scalars}}(x_i^T x_j) + \sum_{i=1}^{m} \alpha_i$$

$g(\alpha)$ ↓ dual form of hard margin SVM

$$\max(g(\alpha))$$

$$\text{s.t.} \quad \alpha_i \geq 0 \quad \forall i$$

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

10th Feb

~~Hard Margin SVM~~

~~$\min_{(\omega, \omega_0)} \frac{1}{2}\|\omega\|^2$~~

~~s.t. $y_i(\omega^T x_i + \omega_0) \geq 1 \quad \forall i$~~

Kernel SVM

$x \in R^d$

$x \xrightarrow{\Phi} z \in R^n$

$\bar{\Phi}(x) \in R^n$, $x \in R^d$

$x_i \Rightarrow z_i = \bar{\Phi}(x)$, x_i: $m \times d$, $\bar{\Phi}(x)$: $m \times n$

$$\text{label}(x_{test}) = \text{sign}\left[\omega^{*T}\bar{\Phi}(x_{test}) + \omega_0^*\right]$$

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

$$\text{label}(x_{test}) = \text{sign}\left[\sum_{i=1}^{m} \alpha_i^* y_i \underbrace{\bar{\Phi}(x_i)^T \bar{\Phi}(x_{test})}_{\text{kernel trick}} + \omega_0^*\right]$$

CamScanner

Mercer's theorem

Kernel funⁿ :- $k : X \times X \to R$

A symmetric funⁿ $k(x_i, x_j) = k(x_j, x_i)$ can be expressed as $k(x_i, x_j) = \phi^T(x_i)\phi(x_j)$ for some $\bar{\phi}$ IFF the matrix

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

Gram matrix

PSD matrix

$$\text{label}(x_{test}) = \text{sign}\left[\sum_{i=1}^{m} \alpha_i^* y_i \underbrace{k(x_i, x_{test})}_{\bar{\phi}(x_i)^T \bar{\phi}(x_{test})} + \omega_0^*\right]$$

kernel trick

* Examples of kernel

① linear kernel $K(x_i, x_j) = x_i^T x_j$

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

③ Gaussian kernel (RBF - Radial basis funⁿ) $k(x_i x_j) = e^{-\left(\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right)}$

⇒ for making new kernel

* Soft-SVM

① L_1 - SVM

$$\min_{(\omega, \omega_0, \xi)} \frac{\lambda}{2}\|\omega\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \xi_i$$

λ → regularization parameter

ξ_i → slag parameter

→ more penalize (affects more) (mostly b/w $0<1$)

s.t. $y_i(\omega^T x_i + \omega_0) \geq 1 - \xi_i$

$\xi_i \geq 0 \ \forall i$

② L_2 - SVM

$$\min_{(\omega, \omega_0, \xi)} \frac{\lambda}{2}\|\omega\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \xi_i^2$$

s.t. $y_i(\omega^T x_i + \omega_0) \geq 1 - \xi_i$

$\xi_i \geq 0 \ \forall i$

ξ_i^2 → less penalize

) L_1 - CSVM

$$\frac{C}{n}\sum_{i=1}^{n} \xi_i + \|\omega\|_2^2$$

L_2 - CSVM

$$\frac{C}{n}\sum_{i=1}^{n} \xi_i^2 + \|\omega\|_2^2$$

e.g. $X = R$ $k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right)$

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

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

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

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

$$\boxed{\begin{matrix} \left\langle 1, \frac{x_i}{\sigma}, \frac{x_i^2}{\sigma^2\sqrt{2!}}, \cdots\right\rangle \\ \left\langle 1, \frac{x_j}{\sigma}, \frac{x_j^2}{\sigma^2\sqrt{2!}}, \cdots\right\rangle \end{matrix}}$$

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

CamScanner

AML 13th Feb

* Linear SVM → runs faster when more no. of data
  SVC (kernel 'linear')
  same

* Eigen Value Decomposition

$$\underset{n\times n}{A}\,\underset{n\times 1}{v} = \lambda\, \underset{n\times 1}{v} \qquad A = Q\Lambda Q^{-1}$$

⇒ $Ax = 0$ $\forall x$ satisfies this eqn they form a null space.

$$\dim\begin{pmatrix}\text{Range}\\ \text{space}\end{pmatrix} + \dim\begin{pmatrix}\text{null}\\ \text{space}\end{pmatrix} = \dim(\text{total})$$

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

n linearly independent eigenvectors → A can be factorized as $A = Q\Lambda Q^{-1}$

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

→ for every $n\times n$ real symmetric matrix → $A = A^T$
⇓
eigen values are real & Eigen vectors can be choosen ~~org~~ orthogonal → $A^TA = I$

$A = Q\Lambda Q^T$

* Algebraic Multipilicity
  No. of repititions of a particular eigen value is its algebraic multipilicity

* Geometric Multipilicity → No. of Linear Independent eigenvectors associated with it. i.e. dim of null space of $A - cI$ $\quad AM(c) \geq GM(c)$

* disadvantage
  ↳ Only works for square matrix
  ↳ may be complex eigenvalues

* SVD

Any matrix $A(n \times d)$

↓

decomposed

$A = U \Sigma V^T$ → orthonormal columns → $V^T V = I^d$

$U^T U = I^n$ ← set of orthonormal columns

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

$$\Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_d \end{bmatrix}$$

$\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \cdots \geq \sigma_d$

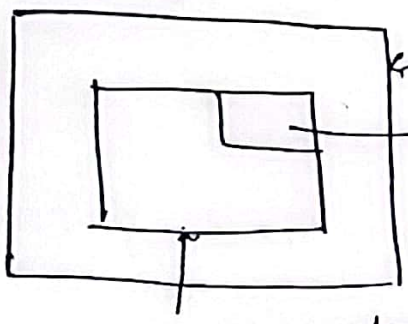
Diagonal Matrix

↓

all values ≥ 0

20-Feb

* Eigen Value decomposition



Rectangular Matrices

Real & Symmetric

↓

Eigen Value decomposition possible

square matrices

* Spectral theorem (Foundation ~~of~~ for data science - Ravi Kaman)
Hopcroft

* $A^T A x = \lambda x$

$A A^T A x = \lambda A x$

$A A^T = \lambda = U D V^T$

$$A A^T = (U \Sigma V^T)^T (U \Sigma V^T) = V \Sigma^T \overbrace{U^T U}^{\text{orthonormal}} \Sigma V^T$$

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

$$= V \Sigma^2 V^T$$

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

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

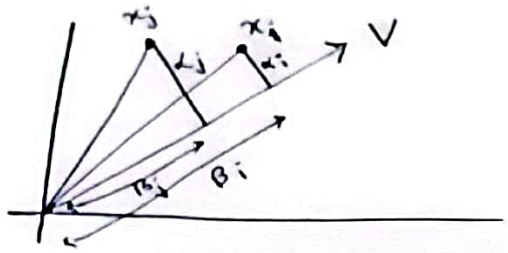
$$= U \Sigma \Sigma^T U^T$$

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

CamScanner

U → formed by eigenvectors of AA^T
V → formed by eigenvectors of A^TA

Singular Values :- square roots of eigenvalues of A^TA
→ always non-negative values



min $\sum \alpha^2$
equivalent to $\sum \beta^2$

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

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

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

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

$$\min \|Av\|_2^2 \longrightarrow \text{best fit line}$$

$$V_1 = \underset{\|v\|_2 = 1}{\operatorname{argmax}} \|Av\|_2 \longrightarrow \text{first singular vector of } A$$

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

similarly $V_2 = \underset{v \perp v_1,\ \|v\|_2 = 1}{\operatorname{argmax}} \|Av\|_2$

similarly find

$V_3, V_4, \ldots, V_d$

2 * Important Result

A :

Let A be an $m \times d$ matrix where $v_1, v_2, \ldots, v_r$ 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-dim subspace for A.

Ans 1

→ Let A be an $n \times d$ matrix where $v_1, \ldots, v_n$ are singular vectors. For $1 \le k \le n$, let V_k be subspace spanned by $v_1, \ldots, v_k$, then for each k, V_k is best fit k-dim subspace for A.

→ $v_1 = \arg\max_{\|v\|_2} \|Av\|_2$

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

$v \in R^{d \times 1}$

$Av \in R^{n \times 1}$

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

$u_i \rightarrow$ left singular vectors

$u_i = \frac{1}{\sigma_i(A)} Av_i$

→ Economical SVD & normal SVD (truncated SVD)

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

σ_i → scalar, u_i → $(n \times 1)$, v_i^T → $(1 \times d)$, $u_i v_i^T$ → $(n \times d)$ → summation of rank 1 matrices

$d \rightarrow$ rank 1 matrices

→ $Av = \lambda v$

$Av = \sigma u$

→ ~~For~~ Frobenius Norm :-

$\|A\|_F = \sqrt{\sum_{j,k} a_{jk}^2} = \sqrt{\sum_{i=1}^{Rank(A)} \sigma_i^2}$

$\|A\|_2 = \sigma_1(A) \rightarrow$ spectral norm

↳ highest singular value of matrix

→ $A^{n \times d}$ $Rank(A) = d$

$k < d$

$\tilde{A}^{n \times d}$ $Rank(\tilde{A}) = k$

↳ low rank A

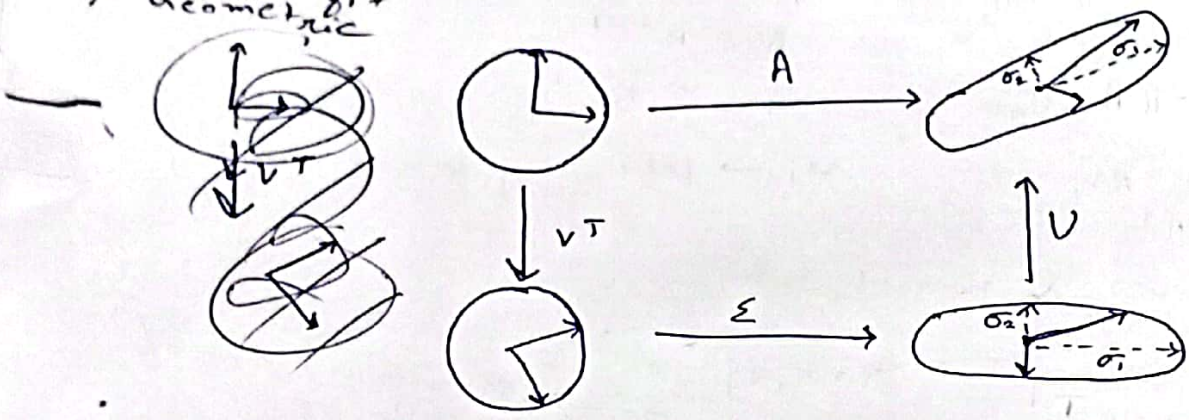
* Low-Rank Approximation

$$\min_{\hat{A} \in \mathbb{R}^{m \times n}} \|A - \hat{A}\|_F^2 \quad \text{s.t. } \operatorname{rank}(\hat{A}) = k \qquad k < d$$

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

best fit

* Geometric



$$\|Ux\|_2 = \|x\|_2$$

$$\|V^T x\|_2 = \|x\|_2$$

24-2-23

1. Random Undersampling
2. Random Oversampling
3. Synthetic Minority Over-Sampling technique

→ Penalize

* Parameter Estimation

Discriminative

Generative

21 Feb.

Bayes

$$\underbrace{P(\theta/D, \alpha)}_{\substack{\text{Posterior} \\ [0,1]}} \propto \underbrace{P(D/\theta, \alpha)}_{\substack{\text{Likelihood} \\ \geq 1}} \cdot \underbrace{P(\theta/\alpha)}_{\text{Prior}}$$

MAP

$$P((x_1,y_1)(x_2,y_2)\cdots(x_n,y_n)/\alpha) \propto \underbrace{P((y_1,y_2\cdots y_n)/(x_1,x_2,\cdots x_n,\alpha,\theta)}_{L(\theta)} \cdot \underbrace{P(\theta/\alpha)}_{\text{uniform}}$$

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

$$= L(\theta) = \prod_{i=1}^{n} P(y_i / x_i, \theta)$$

$\Rightarrow$ Conjugate Prior
$\rightarrow$ MC-MC

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

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

$$= \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} -\log P(y_i/x_i, \theta)$$

$$l(\theta, (x, y)) = -\log P(y_i/x_i, \theta)$$

$$P(y_i/x_i, \underset{\theta}{(\omega, \beta)}) = N(y_i/\omega^T x_i, \underset{\sigma^2}{\tfrac{1}{\beta}}) \quad \sim \epsilon \quad N(0, \tfrac{1}{\beta})$$

$$\text{Likelyhood} = \prod_{i=1}^{n} P(y_i/\omega^T x_i, (\omega, \beta))$$

$$= \prod_{i=1}^{n} N(y_i/\omega^T x_i, 1/\beta)$$

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

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \begin{matrix} \to N(\mu_1, \sigma_1) \\ \to N(\mu_2, \sigma_2) \\ \\ \to N(\mu_n, \sigma_n) \end{matrix} \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_n \end{bmatrix} \quad \Sigma = \begin{bmatrix} \quad \end{bmatrix}$$

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

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

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

minimize

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

$$= \sum_{i=1}^{n} \left(-\beta(y_i - \omega^T x_i) x_i\right) = 0$$

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

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

$$\| \; 0$$

$$\frac{1}{\beta} = (y_i - \omega^T x)^2$$

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

$$\beta(y_i x_i - \omega^T x^T x) = 0$$

$$\omega^T = (x^T x)^{-1}(x^T y)$$

1. Mamb

$P(y_i / x_i, \omega, \beta)$

$P(y_i / x_i) \sim N(\omega^T x_i, 1/\beta)$

$y_i : \omega^T x_i + \epsilon$

$N(0, \beta)$

* Log-likely hood

$\frac{\partial L}{\partial \omega}(\omega, \beta) = 0 \Rightarrow \omega_{MLE} = (X^T X)^{-1} X^T y$ ← random

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

⇒ Gauss-Markov theorem

ω_{MLE} is BLUE → Estimator

best Linear Unbiased

$E_D[W_{MLE}] = W^*$

$E_D[(X^T X)^{-1} X^T Y] = (X^T X)^{-1} X^T E_D[Y]$ [randomness is only in Y]

$Y = XW + \text{Error}$

$\Downarrow$

$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \sim N(0$

$E_D[Y] = E[X\omega] + E[\varepsilon]$ → Linearity of Expectation

mean = 0

$= E[X\omega] + 0$

$= \underbrace{(X^T X)^{-1} X^T X}_{I} \omega$

$= \omega$ → Only ω so unbiased

$$E_D\left[\frac{1}{\beta_{MLE}}\right] = E\left[\frac{1}{n}(X\omega_{MLE} - Y)^T(X\omega_{MLE} - Y)\right]$$

$$= \frac{1}{n} E_D\left[\omega_{MLE}^T X^T X \omega_{MLE} - 2\,\omega_{MLE}^T X^T Y + Y^T Y\right]$$

$$\omega_{MLE}^T X^T Y = Y^T X \omega_{MLE}$$

$$\cancel{= \frac{1}{n} E_D[\omega_M}$$

$$= \frac{1}{n} E_D\left[Y^T X (X^T X)^{-1} (X^T X) (X^T X)^{-1} X^T Y - 2\, Y^T X (X^T X)^{-1} X^T Y + Y^T Y\right]$$

$$= \frac{1}{n} E_D\left[-Y^T X (X^T X)^{-1} X^T Y + Y^T Y\right]$$

$$= \frac{1}{n} E_D\left[Y^T \underbrace{\left(I - X(X^T X)^{-1} X^T\right)}_{Z} Y\right]$$

$$= \frac{1}{n} E_D\left[Y^T Z Y\right]$$

$$= \frac{1}{n} E_D\left[(X\omega + \epsilon)^T Z (X\omega + \epsilon)\right]$$

$$= \frac{1}{n} E_D\left[\underbrace{\omega^T X^T Z X \omega}_{\text{as it is}} + \underbrace{\omega^T X^T Z \epsilon}_{\overset{\shortparallel}{0}} + \underbrace{\epsilon^T Z X \omega}_{\overset{\shortparallel}{0}} + \epsilon^T Z \epsilon\right]$$

$E_D(\epsilon) = 0$

$\cancel{= \frac{1}{n} E_D[\omega^T X^T Z X \omega}$ because of unbiased

$$= \frac{1}{n}\left[\omega^T X^T Z X \omega + 0 + 0 + \sum Z_{ij} E_D[\epsilon_i \epsilon_j]\right]$$

$$[\varepsilon_1 \ \varepsilon_2 \ \cdots \ \varepsilon_n]_{1\times n} \begin{bmatrix} z_{11} & z_{12} & \\ & & \\ & & \end{bmatrix}_{n\times n} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

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

$$\cancel{[\varepsilon_1 \quad\quad\quad]_{1\times n}}$$

$$\left[\varepsilon_1 [z_{11}] + \varepsilon_2 [z_{22}] + \quad \varepsilon_1 [z_{12}] + \quad \right]_{1\times n} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

$$(\varepsilon_1 \varepsilon_1 z_{11} + \varepsilon_1 \varepsilon_2 z_{11}) \quad (\varepsilon_1 + \varepsilon$$

$$+ \varepsilon_1 \varepsilon_2 z_{11} +$$

$$= \frac{1}{n} \left[\omega^T x^T z \, x \omega + \sum z_{ij} \, E[\epsilon_i \, \epsilon_j] \right]$$

$$\epsilon_i \epsilon_j = 0 \text{ when } i \neq j$$

$$\sum_i \sum_j Pr(\epsilon_i \epsilon_j) \cdot \epsilon_i \epsilon_j$$

$$\Downarrow$$

$$Pr(\epsilon_i)(\epsilon_j)$$

$$\sum_i Pr(\epsilon_i) \epsilon_i \quad \sum_j Pr(\epsilon_j) \cdot \epsilon_j$$

$$Var(x) = E[x^2] - [E(x)]^2$$

$$E(x) = 0$$

$$\frac{1}{\beta} = E[\epsilon_i^2]$$

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

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

$$= \frac{1}{n}\left[\omega^T X^T X\omega - \omega^T \underbrace{X^T X (X^TX)^{-1} X^T}_{I} X\omega + \frac{1}{\beta}\sum_{i=1}^{n} \operatorname{diag}(z)\right]$$

$$\Downarrow$$

$$\omega^T X^T X \omega$$

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

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

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

$$\Downarrow$$

$$tr\left(\underbrace{X^TX(X^TX)^{-1}}_{I}\right)$$

$$\Downarrow$$

$$d$$

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

$$= \frac{(n-d)}{n\beta}$$

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

no. of features

no. of samples

$$\left[\frac{1}{\beta}\right]$$

13 March AML

$$\mathrm{Var}[w_{MLE}] < \mathrm{Var}[\tilde{w}]$$

any other line

$$A - B \succeq 0$$

positive semi definite

$A \succeq B$ [$\because$ A & B are matrices]

$$\mathrm{var}(x) = E\left((x-\mu)(x-\mu)^T\right)$$

$$\mathrm{Var}[W_{MLE}]$$

$$= E\left[(\omega_{MLE} - \omega)(\omega_{MLE} - \omega)^T\right]$$

$$Y = X\omega + \epsilon$$

$$= E\left[\left[(X^TX)^{-1}X^TY - \omega\right]\left[(X^TX)^{-1}X^TY - \omega\right]^T\right]$$

$$= E\left[\left[(X^TX)^{-1}X^T(X\omega + \epsilon) - \omega\right]\left[(X^TX)^{-1}X^T(X\omega + \epsilon) - \omega\right]^T\right]$$

$$= E\left[\left[(X^TX)^{-1}(X^TX\omega + X^T\epsilon) - \omega\right]\left[(X^TX)^{-1}(X^TX\omega + \epsilon) - \omega\right]^T\right]$$

$$= E\left[\left[\omega + (X^TX)^{-1}X^T\epsilon - \omega\right]\left[(X^TX)^{-1}X^T\epsilon + \omega - \omega\right]^T\right]$$

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

$$= E\left[(X^TX)^{-1}X^T\underbrace{\epsilon \cdot \epsilon^T}X\left((X^TX)^{-1}\right)^T\right]$$

$n \times d$

$(n \times n)$

$$= (X^TX^{-1})X^T \; E(\epsilon\epsilon^T) \; X\left((X^TX)^{-1}\right)^T$$

$$\Downarrow$$

$$\frac{1}{\beta} I \leftarrow \text{Identity matrix}$$

$$\hookrightarrow n \times n$$

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

CamScanner

$\tilde{\omega} = A \cdot Y = A(X\omega + \epsilon) = AX\omega + A\epsilon$

$\uparrow$ unbiased

so $E(\tilde{\omega}) = \omega$ $\quad E(AX\omega + A\epsilon) = \omega$

$E(\epsilon) = 0$

$\epsilon(0, \sigma)$

$AX\omega + 0 = \omega$

$\underbrace{AX}_{\text{matrix}}\omega = \underset{\text{scalar}}{1} \cdot \omega \Rightarrow (AX - I)\omega = 0$

$\Rightarrow \omega$ is eigenvector of AX & corresponding eigenvalue is 1

$var[\tilde{\omega}] = \cancel{E[(\tilde{\omega} - \omega}$

$= E\left[(\tilde{\omega} - E(\tilde{\omega})) \cdot (\tilde{\omega} - E(\tilde{\omega}))^T\right]$

$= E\left[(\tilde{\omega} - AX\omega)(\tilde{\omega} - AX\omega)^T\right]$

$= E\left[(A\epsilon)(A\epsilon)^T\right]$

$= E\left[A\epsilon\epsilon^T A^T\right]$

$= A \; E\left[\epsilon\epsilon^T\right] \cdot A^T$

$= A \cdot \frac{1}{\beta} I \cdot A^T$

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

$$\mathrm{var}(\hat{\omega}) - \mathrm{var}(\omega_{MLE}) \succeq 0$$

$$\frac{1}{\beta} A A^T - \frac{1}{\beta}\left((X^T X)^{-1}\right)^T \succeq 0$$

$$A = (X^T X)^{-1} X^T + B$$

$$\frac{1}{\beta}\left(BB^T\right)$$

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

$$BB^T \longrightarrow n \times n$$

$$\downarrow$$

$$C$$

$$\cancel{[x_1 \quad x_2 \quad x_n]\, C \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}$$

$$\cancel{\sum_{i=1}}$$

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

$$(AX - I) = BX$$

$$(AX - I)\omega = 0$$

$$\Rightarrow BX\omega = 0$$

only happen when $BX = 0$ because ω is not a zero

$$\Rightarrow AX - I = 0$$

$$\Rightarrow AX = I$$

$$\Rightarrow A = X^{-1}$$

$$= x^T B B^T x$$

$$= (B^T x)^T (B^T x) \succeq 0$$

$$= \|B^T x\|_2^2 \succeq 0$$

Positive semi definite

Posterior $\propto$ $\underbrace{\text{Prior} \times \text{Likelihood}}$

maximize $\Rightarrow$ maximum a posteriori

$\overset{\shortparallel}{\text{MAP}}$

$$P\Big(\omega / \underbrace{(y_1, y_2, \ldots, y_n)(x_1, \ldots, x_n)}_{\mu_0,\ \Sigma_0,\ 1/\beta},\ 1/\beta\Big)$$

$$\propto P\left[(y_1, y_2, \ldots, y_n) / \omega, \tfrac{1}{\beta}, (x_1, \ldots, x_n)\right] \cdot P\big(\omega, (\mu_0, \Sigma_0)\big)$$

16-Mar MAP

$$y_i = \omega^T x_i + \epsilon_i$$

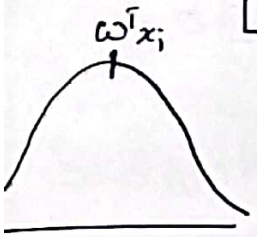
$\omega^T x_i \uparrow N(\mu_0, \Sigma_0)$, $\epsilon_i \uparrow N(0, \frac{1}{\beta})$

$$P\Big(\omega / \underbrace{(y_1, y_2, \ldots y_n)(x_1, x_2, \ldots x_n)}_{D}\Big) \propto P\big((y_1, y_2, \ldots, y_n) / \omega, \tfrac{1}{\beta}, (x_1, x_2 \ldots x_n)\big) \cdot P(\omega / \mu_0, \Sigma_0)$$

$(\mu_0, \Sigma_0, \frac{1}{\beta}) \uparrow$ Parameter

$$P\big(\omega / (y_1 y_2 \ldots y_n), (x_1 x_2 \ldots x_n), \Sigma_0, \mu_0, \tfrac{1}{\beta}\big) \cancel{\propto \prod_{i=1}}$$

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



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

Prior

$$\underbrace{(\omega / D, \Sigma_0, \mu_0, 1/\beta)}_{\text{Posterior}} = \prod_{i=1}^{n} \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(y_i - \omega^T x_i)^2}{1/\beta}\right)$$

$$\cdot \frac{1}{(2\pi)^{d/2} |\Sigma_0|^{1/2}} \exp\left(-\frac{1}{2} (\omega - \mu_0)^T \Sigma_0^{-1} (\omega - \mu_0)\right)$$

CamScanner

$$-\log P(\omega/D, \Sigma_0, \mu_0, 1/\beta)$$

$$= -n\log\sqrt{\beta} + n\log\sqrt{2\pi} - \underbrace{\sum_{i=1}^{n} \log e^{\left(-\frac{\beta}{2}(y_i-\omega^T x_i)^2\right)}}_{\frac{\beta}{2}\sum_{i=1}^{n}(y_i-\omega^T x_i)^2}$$

$$+ \frac{1}{2}\log|\Sigma_0| - \frac{d}{2}\log 2\pi$$

$$+ \frac{1}{2}\left((\omega-\mu_0)^T\Sigma_0^{-1}(\omega-\mu_0)\right)$$

$$= -n\log\sqrt{\beta} + n\log\sqrt{2\pi} + \frac{\beta}{2}\sum_{i=1}^{n}(y_i-\omega^T x_i)^2 \longrightarrow \left[\frac{\beta}{2}(X\omega-Y)^T(X\omega-Y)\right]$$

$$+\frac{1}{2}\log|\Sigma_0| - \frac{d}{2}\log\sqrt{2\pi} + \frac{1}{2}\left[(\omega-\mu_0)^T\Sigma_0^{-1}(\omega-\mu_0)\right]$$

$$\frac{\partial\left(-\log P(\omega/D, \Sigma_0, \mu_0, 1/\beta)\right)}{\partial\omega} = 0 \qquad 2(y-\omega^T x)(x^T)$$

$$\Downarrow$$

$$= 0 + 0 + \frac{\beta}{2}\left[2X^TXW - 2X^Ty\right] \qquad (\omega-\mu_0)$$

$$+ 0 + 0 + \frac{1}{2}\left[2\Sigma_0^{-1}W - 2\Sigma_0^{-1}\mu_0\right]$$

$$= \beta\left[X^TXW - X^Ty\right] + \left[\Sigma_0^{-1}W - \Sigma_0^{-1}\mu_0\right]$$

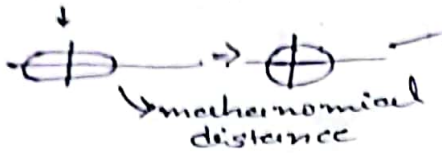
$$\left(\beta(X^TX) + \Sigma_0^{-1}\right)W - \beta X^Ty - \Sigma_0^{-1}\mu_0 = 0$$

$$(\omega-\mu_0)^T\Sigma_0^{-1} \qquad \boxed{W_{MAP} = \underbrace{\left[\beta\underset{PSD}{(X^TX)} + \underset{PD}{\Sigma_0^{-1}}\right]}_{PD}^{-1}\left(\beta X^Ty + \Sigma_0^{-1}\mu_0\right)}$$

If $n = 0 \Rightarrow \omega_{MAP} = \Sigma_0\Sigma_0^{-1}\mu_0 = \mu_0$

CamScanner

⇒ ⇒ Spherical prior → infinitely broad → ($c \to \infty$) least



→ only when eigen values are equal

20

$$\Sigma_0^{-1} = \frac{1}{c} I$$

$$\omega_{MAP} = \left(\beta (x^T x) + \frac{1}{c} I\right)^{-1} \left[\beta x^T y + \frac{1}{c} \mu_0\right]$$

if $c \to \infty$

$$= \left[\beta (x^T x)\right]^{-1} \left[\beta x^T y\right]$$

$$= \cancel{(x^T x)^{-1}} (x^T x)^{-1} \beta^{-1} \beta x^T y$$

$$= (x^T x)^{-1} x^T y$$

$$= \omega_{MLE}$$

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

$$\omega_{MAP} = \left[\beta (x^T x) + \frac{I}{c}\right]^{-1} \left[\beta x^T y + 0\right]$$

$$= (\beta x^T x)^{-1} (\beta x^T y) + \frac{cI}{\beta} x^T y$$

$$= (x^T x)^{-1} \beta^{-1} \beta x^T y + c \frac{\beta x^T y}{\beta}$$

$$= (x^T x)^{-1} x^T y + \frac{\beta c}{\beta} x^T y$$

$$= \left((x^T x)^{-1} + \frac{I}{\beta c}\right)^{-1} x^T y$$

↑ λ (ridge regression)

linear equation regression

⇒ Laplace distribution → lasso problem

$$P(\omega/\sigma_0) = \frac{1}{2\sigma_0} e^{\left(-\frac{\|\omega - \mu_0\|}{\sigma_0}\right)}$$

20 min

~~$W_{MAP} = ($~~

~~W_{MAP}~~

$$\|Ax - b\|_2^2 + \lambda x^T x$$

$$\tilde{A} = \overbrace{\begin{bmatrix} & \\ & \end{bmatrix}}^{\tilde{A}} \begin{bmatrix} x \end{bmatrix} \sim \overset{\tilde{b}}{\begin{bmatrix} \\ \\ \\ \end{bmatrix}}$$

$$d\begin{bmatrix} \sqrt{\lambda} & & \\ & \sqrt{\lambda} & \\ & & \sqrt{\lambda} \end{bmatrix} \qquad d\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$(A^TA + \lambda I)^{-1} A^T y.$$

$$(Ax-b)^T(Ax-b) + (\sqrt{\lambda}x)^T \cdot (\sqrt{\lambda}x$$

$$\Rightarrow W_{MAP} = \left(\beta(X^TX) + \Sigma_0^{-1}\right)^{-1}\left(\beta X^T Y + \Sigma_0^{-1}\mu_0\right)$$

$$\Rightarrow E[W_{MAP}] = \cancel{E\left[(X^TX)^{-1}X^TY\right]} = E\left[(X^TX + \lambda I)^{-1}X^TY\right]$$

spherical prior & $\mu_0 = 0$

$$= \cancel{E\left[(X^TX)^{-1}X^T(X\omega + \epsilon)\right]}$$

$$= E\left[(X^TX + \lambda I)^{-1}X^T(X\omega + \epsilon)\right]$$

$$= E\left[(X^TX + \lambda I)^{-1}(X^TX\omega + X^T\epsilon)\right]$$

$$= (X^TX + \lambda I)^{-1}(X^TX\omega) + (X^TX + \lambda I)^{-1}\underbrace{E(X^T)}_{0}\epsilon$$

$$=$$

$$\mathrm{var}[w_{MAP}] = E\left[(w_{MAP} - E(w_{MAP}))(w_{MAP} - E(w_{MAP}))^T\right]$$

$$= E\left[\left(\underbrace{\cancel{(X^TX+\lambda I)^{-1}(X^Ty)}}_{Z}\right.\right.$$

$$w_{MAP} = \underbrace{(X^TX+\lambda I)^{-1}X^T}_{Z}\underbrace{y}_{(Xw+\epsilon)}$$

$$E(w_{MAP}) = ZXw$$

$$\mathrm{var}[w_{MAP}] = E\left[((Z)(y) - ZXw)(Zy - ZXw)^T\right]$$

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

$$= E\left[Z(Xw + \epsilon - Xw)(Xw + \epsilon - Xw)^T Z^T\right]$$

$$= E\left[Z\,\epsilon\,\epsilon^T Z^T\right]$$

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

$$= Z\,\underbrace{\frac{I}{\beta}}_{n\times n}\,Z^T$$

$$= \frac{1}{\beta}\left((X^TX+\lambda I)^{-1}X^TX\left[(X^TX+\lambda I)^{-1}\right]^T\right.$$

$$= \frac{1}{\beta}ZZ^T \longrightarrow \text{PD if } \lambda > 0$$

$X = U \Sigma V^T$

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

$$= V \Sigma^2 V^T$$

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

$$X^T X + \lambda I = V \Sigma^2 V^T + \lambda V I V^T$$

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

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

$V \cdot V^T = I$

$V^{-1} = V^T$

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

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

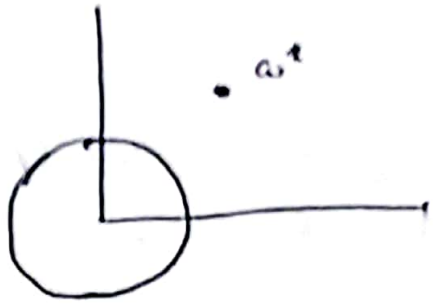
$$= V \, \mathrm{diag}\left(\frac{\sigma_i^2}{\lambda + \sigma_i^2}\right) V^T$$

As $\lambda \uparrow$ sol^n approches to 0.

CamScanner

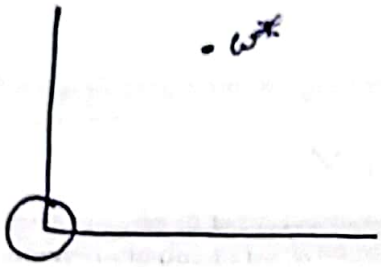
$n = 0, \lambda = 0$

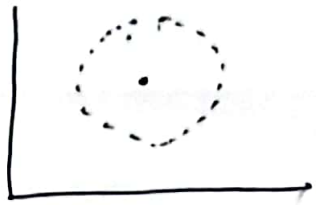
$n = 0, 0 < \lambda < \infty$



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

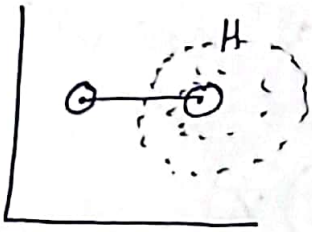
$n = 0, \lambda \to \infty$





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

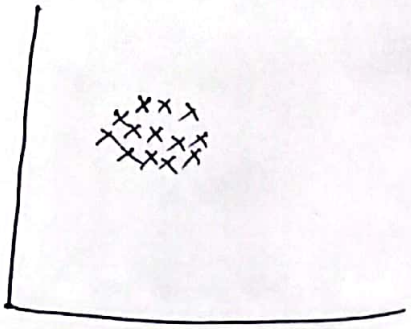
$0 < n < \infty, 0 < \lambda < \infty$ (low bias, high variance)



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

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

$E(w_{MLE}) \approx E(w_{MAP})$



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

23-March

GMM - Gaussian Mixture Models (unsupervised)

→ Gaussian distribution by itself is may not ~~suffi~~ enough in modelling real datasets

→ sometimes a linear superposition of two or more gaussian

→ For a superposition of k Gaussians

$$P(x) = \sum_{k=1}^{K} \underbrace{\omega_k}_{\text{Parameter}} \underbrace{N(x/\mu_k, \Sigma_k)}_{\text{each component of mixture}}$$

each $N(x/\mu_k, \Sigma_k)$ is called a component of mixture

Parameter ↑ mixing coefficient

we have, $\sum_{k=1}^{K} \omega_k = 1$

Also, $P(x) \geq 0$, $N(x/\mu_k, \Sigma_k) \geq 0$

$\Rightarrow \omega_k \geq 0 \quad \forall k$

$0 \leq \omega_k \leq 1$ (can be thought as probability) $[\omega_k = P(k)]$

probability X conditional. k

$$P(x) = \sum_{k=1}^{K} P(k) \, P(X/k)$$

Posterior

prior of picking kth component

$P(X/k) \to N(x, \mu_k, \Sigma_k)$ (likelihood)

Posterior probability

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

$$= \frac{p(k) \cdot p(x/k)}{\sum_l p(l)\, p(x/l)}$$

$$= \frac{\omega_k N(x/\mu_k, \Sigma_k)}{\sum_l \omega_l N(x/\mu_l, \Sigma_l)}$$

$\Rightarrow$ Parameters of GMM ω, μ, Σ

$W = (\omega_1, \ldots, \omega_N)$

$\mu = (\mu_1, \ldots, \mu_N)$

$\Sigma = (\Sigma_1, \ldots, \Sigma_N)$

$\Rightarrow$ to get the value - log of likelihood

$$\ln P(X/W, \mu, \Sigma) = \sum_{i=1}^{N} \ln \left\{ \sum_{k=1}^{K} \omega_k N(x_n/\mu_k, \Sigma_k) \right\}$$

assuming each data independen

$\Rightarrow$ Interpnete as

$Z \longrightarrow k$ - dimensional binary random variable

at a time some $Z_k = 1$ & all other co-ordinates are 0.

$\therefore Z_k \in \{0, 1\}$ & $\sum_k Z_k = 1$

k possible states of the vector z

$\rightarrow$ Try to define joint distribution $P(X/z)$ in terms of $P(z)$ & $P(x/z)$

Now $P(Z_k = 1) = \omega_k$ $\left(\begin{array}{l} 0 \le \omega_k \le 1 \\ \& \sum_k \omega_k = 1 \end{array} \right.$

$$P(z) = \prod_{k=1}^{K} \omega_k^{Z_k}$$

Also, $P(X/Z_k = 1) = N(x/\mu_k, \Sigma$

$$p(x/z) = \prod_{k=1}^{K} N(x/\mu_k, \Sigma_k)^{z_k}$$

Now.

$$p(x) = \sum_z p(z) \cdot p(x/z)$$

$$= \sum_{k=1}^{K} \omega_k N(x/\mu_k, \Sigma_k)$$

$\therefore$ Marginal of x is GMM.

→ If we have observations $x_1, \ldots, x_N$ and as $p(x) = \sum_z p(x, z)$ for every observed data point x_n corresponding latent variable is z

Now $\gamma(z_k)$ is $p(z_k = 1/x)$

responsibility

$$= \frac{p(z_k = 1) \cdot p(x/z_k = 1)}{\sum_l p(z_l = 1) \cdot p(x/z_l = 1)}$$

$$= \frac{\omega_k N(x/\mu_k, \Sigma_k)}{\sum_l \omega_l N(x/\mu_l, \Sigma_l)}$$

~~⇒ / pdf~~

24 March

* Expectation maximization for GMM

log likelihood for GMM given data X is given by

$$\ln p(X/\omega, \mu, \Sigma) = \sum_{i=1}^{N} \ln \left\{ \sum_{k=1}^{K} \omega_k N(x_n/\mu_k, \Sigma_k) \right\}$$

derivating w.r.t. μ_k and equating to 0

$$0 = -\sum_{i=1}^{N} \underbrace{\frac{\omega_k N(x_n/\mu_k, \Sigma_k)}{\sum_j \omega_j N(x_n/\mu_k, \Sigma_k)}}_{\gamma(z_{nk})} \cdot \Sigma_k (x_n - \mu_k)$$

$$e^{-(x-\mu_k)^T \Sigma (x-\mu_k)} \rightarrow (x-\mu_k)^2 \Sigma$$

Multiplying by Σ_k^{-1}

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

where

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

effective no. of points in the particular k^{th} cluster

mean μ_k of kth Gaussian $\longrightarrow$ weighted mean of all points in the kth Gaussian

similarly

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

Finally we need to maximize

$$\ln(X/\omega, \mu, \Sigma) \text{ w.r.t. } \omega_k$$

we consider

$$\ln P(X/\omega, \mu, \Sigma) + \lambda \left(\sum_{k=1}^{K} \omega_k - 1 \right)$$

which gives

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

multiply both side by ω_k & sum over k, we get $\lambda = -N$

finally $\omega_k = \frac{N_k}{N} \longrightarrow$ Ratio of no. of points in the cluster to total no. of points

⇒ EM for Gaussian Mixture

1. Initialize μ_k, Σ_k & ω_k → evaluate initial value of log likelihood

2. E - step → Expectation step

$\gamma(z_{nk})$ → find out

3. M - step

μ_k^{new}

Σ_k^{new}

ω_k^{new}

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

repeate 2 - 4 till convergence

3 April

LDA & Topic Modeling

→ Every document - mixture of topics

Topic - mixture of words on dist. over words

→ Latent Dirichlet Allocation is a generative model.

⇒ Multinomial distribution

$P(x_k = 1) = \mu_k$ → $P(x/\mu) = \prod_{k=1}^{K} \mu_k^{x_k}$

where $\mu = (\mu_1, \ldots, \mu_k)$ & $\mu_k \geq 0$

$\sum_k \mu_k = 1$

dataset D of N independent observations

$X_1, X_2, \ldots, X_N$

likelihood $P(D/\mu) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}}$

$$= \prod_{i=1}^{K} \mu_k^{\left(\sum_n x_{nk}\right)}$$

$$= \prod_{i=1}^{K} \mu_k^{m_k} \quad \left[m_k = \sum_n x_{nk} \right]$$

no. of observations of $x_k = 1$

Now, the maximum likelihood soln for μ is

$$\mu_k^{ML} = \frac{m_k}{N}$$

We consider the joint distribution of quantities $m_1, m_2, \ldots, m_K$ conditioned on μ & N

$$\text{mult}(m_1, m_2, \ldots, m_K / \mu, N)$$

$$= \binom{N}{m_1\ m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

where $\binom{N}{m_1\ m_2 \ldots m_K} = \dfrac{N!}{m_1!\, m_2! \ldots m_K!}$

& $\sum_{k=1}^{K} m_k = N$

⇒ Dirichlet Distribution

- Conjugate prior for multinomial distribution

$P(\mu/\alpha)$ varies as $\prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$

where

$$0 \le \mu_k \le 1$$

$$\sum_k \mu_k = 1$$

CamScanner

[illegible] α_k are the parameters of the distribution

$$\alpha = (\alpha_1 \ldots \alpha_k)$$

Dirichlet distribution

$$Dir(\mu/\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} \prod_{i=1}^{k} \mu_k^{\alpha_k - 1}$$

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k$$

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} \, du$$

$\Gamma(x+1) = x\Gamma(x)$

$\Gamma(1) = 1$

extension of factorial to the real no's

Now,

$$P(\mu/D, \alpha) \propto P(D/\mu) \, P(\mu/\alpha) \rightarrow \text{prior}$$

likelihood

$$\propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}$$

Thus

$$P(\mu/D, \alpha) = Dir(\mu/\alpha + m_k)$$

$$= \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1) \ldots \Gamma(\alpha_k + m_k)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k}$$

$$m = (m_1 \ldots m_k)$$

CamScanner