19th Jan 23

(*) Advance ML (*)

classmate Date Page

Def-A :- For $x_1, \ldots, x_n \in X$ The dichrotomies generated by $\mathcal{H}$ on these points is defined by

$$\mathcal{H}(x_1, \ldots, x_n) = \{ h(x_1), \ldots, h(x_n) \mid h \in \mathcal{H} \}$$

what is dichrotomies?

max. dichrotomies possible = 2^n

NOTE: Higher the VC dimension, ~~Higher~~ the Richer the hypothesis Space...

what is growth function of hypothesis set $\mathcal{H}$?

what is VC-dimension

Theorem :-

⇒ If $m_{\mathcal{H}}(k) < 2^k$ for some value k then

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} {}^{N}C_i$$

(Sauer's Lemma)

Higher VC dim is good for test error?
→ NO, it may overfit.

⇒ By definition of VC-dimension...

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_{VC}} {}^{N}C_i$$

Higher the VC dimen. Higher the generalization error

⇒ $d_{VC} \geq N \iff$ there exists data of size N such that $\mathcal{H}$ shatters the data.

⇒ The VC-Generalization bound for any tolerance $\delta > 0$.

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

paperswithcode.com/rc2022

distill pub

NeurIPS 2021/2022
ICML 2021/2022
AAAI 2021/2022
ACM. FACCT 2021/2022

20th Jan

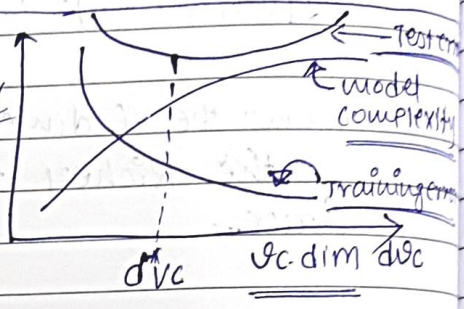
Tasks:- (1) Replicate the results. (Dataset public?).

(2) Do something more → Tweak the algo...
→ Hyperparameter Tweaking...

(3) if code already available then ~~tons~~ replicate it to another framework...

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

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



- The VC bound is loose:-
  - Hoeffding's inequality has some slack.
  - $m_H(N)$ gives worst case estimate.
  - Bounding $m_H(N)$ ~~bound~~ by polynomial of order.

- VC dimension is useful for comparing generalization performance of models.
- Loose but equally loose for all.
- Establishing the feasibility of learning for infinite hypothesis sets.
- Rule of thumb:-

  N ≥ 10× d_{VC}

(*) Bias and variance:-

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

Expected value w.r.t (x).. $g^{(D)}$ is dependent on (D).

$$E_D\left[E_{out}(g^{(D)})\right]$$
$$= E_D\left[E_x\left[\left(g^{(D)}(x) - f(x)\right)^2\right]\right]$$
$$= E_x\left[E_D\left[\left(g^{(D)}(x) - f(x)\right)^2\right]\right]$$
$$= E_x\left[E_D\left[g^{(D)}(x)^2\right] - 2E_D\left[g^{(D)}(x)\right] f(x) + f(x)^2\right]$$

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

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

25th January 23

$= E_x[$

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

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

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

$= (\bar{g}(x) - f(x))^2$ → Bias

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

⇒ Here, along with $\mathcal{H}$ algorithm $\mathcal{A}$ also matters...

(*) convex optimization in ML :-

outline:-
- convex optimization.
- Primal - Dual Formulation → KKT conditions.

} SVM problem.

⇒ if function is convex function then Gradient Descent will give global minimum.

optimization of ML
- continous
  - convex
  - Non-convex
- Discrete. (combinatorial) → submodularity.

Perceptron Learning algo.

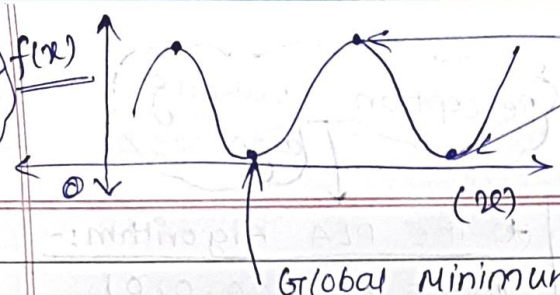
(*) The PLA Algorithm :-

$w(0) = (0, 0, 0, 0, 0)$

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

$w(t+1) = w(t) + y(t) * (t)$.

27th Jan 2023



Gradient is zero (stationary points)

classmate Date Page

$f: R^n \rightarrow R$

unconstrained optimization

$\min: f(x)$

$x \in D = R^n$

constrained optimization

$\min: f(x)$

$x \subseteq R^n$

Example :-

Linear Regression

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

$B \in R^n$

$\min_{x \in R^d} \|Ax - b\|^2$

unconstrained.

Ridge Regression.

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

$\{\{$

$\min_x \|Ax - b\|^2$ such that $\|x\|_2^2 < R.$ ← constrained.

(i) Let $f_n: R^n \rightarrow R$ continous and differentiable fn.

↳ For local optimum :

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial (x_n)} \end{bmatrix}_{n \times 1} = 0$$ (zero vector)

$$\nabla^2 f(x) = H(x)$$ matrix $(n \times n)$

Hessian Matrix

$$\begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \end{bmatrix}$$

if Hessian is PSD: local minimum

else if Hessian is NSD: local maximum

(*) Positive / Negative Semi-Definite:-

PSD ⇒ $x^TAx \geq 0$ ⇒ all eigen values are ≥ 0

NSD ⇒ $x^TAx \leq 0$ $\forall x$

x^TAx : Quadratic form.

$1 \times n$ $n \times n$ $n \times 1$

1×1

$\forall x \neq 0$

if $x^TAx > 0$ then it is (P.D.) matrix.

⇒ all eigen values are greater than zero.

example: $A = I$ (identity matrix)

$\Rightarrow x^TAx > 0$

$\Rightarrow x^TIx > 0$

$\Rightarrow x^Tx > 0$

$\Rightarrow \|x\|_2^2 > 0$

$A \succ B$

↑ A is P.D. w.r.t B.

means $A - B \succ 0$

(*) constrained optimization problem :-

8th. Feb 23

$f_0 : R^n \to R$

$\min f_0(x)$

such that

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

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

Lagrangian form of C.O.P

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

original function

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

s.t. $\|x\|_2^2 \leq R$

$\lambda > 0$

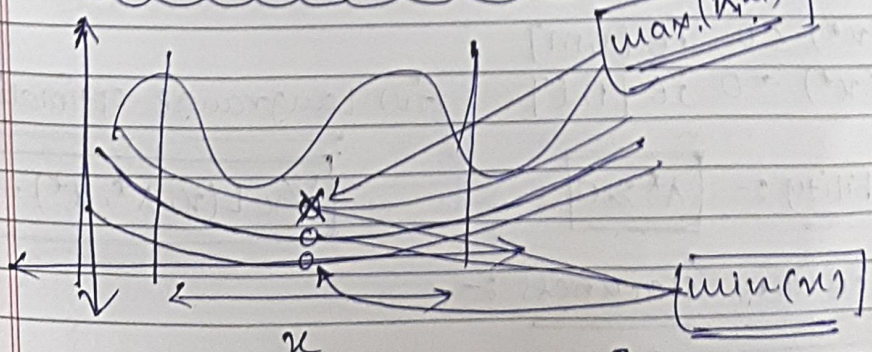
λ_i, μ_j is Lagrangian multiplier.

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

what if x is feasible point ?

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

$\min_x L(x, \lambda, \mu) \leq f_0(x^*)$ s.t. $\lambda \geq 0$



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

8th Feb 2023

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

dual function.

$$\max_{(\lambda,u)} g(\lambda, u) \leq f_0(x^*)$$

$$\lambda \geq 0$$

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

$\Rightarrow$ (λ^*, u^*) are point of ~~maximum~~ maxima

$\Rightarrow$ $g(\lambda, u) \leq g(\lambda^*, u^*)$

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

maximum value of the dual function is ~~equal to~~ less than or equal to primal function.

$f_0(x^*) - g(\lambda^*, u^*) \geq 0$ } weak duality theorem

duality gap.

$\Rightarrow$ Primal Problem is convex problem. (usually) $\Rightarrow$ strong duality hold (not always)

$f_0(x^*) - g(\lambda^*, u^*) = 0$

constraint should satisfy some condition

(*) KKT conditions :-

x^* ← point of primal optimal

λ^*, u^* ← point of dual optimal with zero duality gap.

(i) Primal Feasibility :-

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

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

(ii) Dual feasibility :- $\lambda^* \geq 0$

(iii) complementary slackness :-

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

(iv) Laugrange optimality :-

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

8th Feb 2023

Any tuple (x, λ, u) $\in R^n \times R^m \times R^l$ is called KKT point if it satisfies the KKT conditions

when you have $(x^*, \lambda^*, u^*) \Rightarrow$ KKT point.

if duality gap = 0

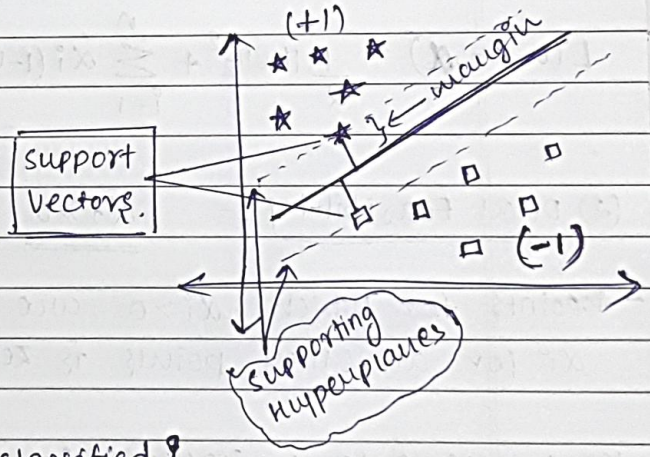
x^* } primal optimal
λ^* } dual optimal
u^* }

$\Rightarrow$ (x^*, λ^*, u^*) is KKT point

$\Leftarrow$ If Problem is convex

9th Feb. 2023

(*) Hard Margin SVM :-



$$\min_{(w, w_0)} \frac{1}{2} \|w\|_2$$

s.t

$$y_i(w^T x_i + w_0) \geq 1 \quad \forall i \in [1, n]$$

- How new data point will be classified?

$\Rightarrow$ $x_{new} \leftarrow$ new data point.

$\rightarrow$ if $(w^T x_{new} + w_0)$ is positive then positive class. else negative class...

- How to detect misclassification?

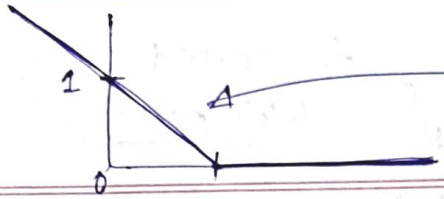
if $y_i(w^T x_i + w_0) < 1$ then point is misclassified...

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

Hinge

$$\lambda\left[(x_i y), (w, w_0)\right] = \max\left(0, 1 - y_i(w^T x_i + w_0)\right)$$

9th Feb. 2023



Hinge Loss

$$\min_{(w,w_0)} \frac{1}{2}\|w\|_2^2$$

s.t. $y_i(w^T x_i + w_0) \geq 1 \quad \forall i \in [1,n]$

} Primal form of Hard Margin.

= KKT conditions :-

(1) Primal Feasibility :-

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

w^*, w_0^*

Primal variables.

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

(2) Dual Feasibility :- $\alpha_i \geq 0 \quad \forall i \in [1,n]$

→ points for which $\alpha_i > 0$ are the support vectors.
α_i for all other points is zero.

(3) Complementary slackness :-

$$\alpha_i (1 - y_i(w^T x_i + w_0)) = 0 \quad \forall i \in [1,n].$$

(4) Lagrange optimality

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

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

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

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

9th Feb 2023

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

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

α_i for $i \in C^+$
α_j for $j \in C^-$
$\boxed{\sum \alpha_i = \sum \alpha_j}$

- **Support vectors :-**

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

$$\min_{(w, w_0)} \frac{1}{2} \|w\|_2^2$$

such that $1 - y_i(w^T x_i + w_0) \leq 0 \quad \forall i \quad i \in [1, n]$

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

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

$w_i^* = \sum_{i=1}^{n} \alpha_i y_i x_i$ $\quad$ w_0 satisfies $\sum_{i=1}^{n} \alpha_i y_i = 0$

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

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

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

$$= \frac{1}{2} w^{*T} w_0^* + \sum_{i=1}^{n} \alpha_i - w^{*T} w^*$$

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

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

9th Feb 2023

Max $g(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$

s.t. $\alpha_i \geq 0$ ← Dual form of Hard margin SVM

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

10th Feb 23

(*) Kernel SVM :-

$x \in R^d$

$x \longrightarrow z \in R^h$

s.t. $n > d$

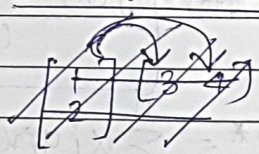
$x_i \Rightarrow z_i \Rightarrow \phi(x_i)$

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

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

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

- How to implement SVM in higher dimension without actually getting $\phi(x)$:-



Mercers Theorem :-

Kernel fun :-

$$K : X \times X \longrightarrow R$$

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

IFF,

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

Gram matrix.

- Example of Kernels :-

- Linear Kernel :-

$$K(x_i, x_j) = x_i^T x_j$$

classmate
Date
Page
10th Feb 2023

- Polynomial Kernel. :-

$$K(x_i, x_j) = (1 + x_i x_j)^t \quad ; t > 0$$

- Gaussian Kernel. (RBF Kernel) :-

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

NOTE: A necessary and sufficient condition to make our custom kernel is that the Gram Matrix should be positive definite.

(*) Soft SVM :-

Regularization term

L1-SVM :- $\min_{(w, w_0, \xi_i)} \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{2}\sum_{i=1}^{n} \xi_i$

L2-SVM :- $\min_{(w, w_0, \xi_i)} \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{2}\sum_{i=1}^{n} \xi_i^2$

L1 C SVM :- $\min_{(w, w_0, \xi_i)} \frac{C}{n}\sum_{i=1}^{n} \xi_i + \frac{\lambda}{2}\|w\|_2^2$

L2 C SVM :- $\min_{(w, w_0, \xi_i)} \frac{C}{n}\sum_{i=1}^{n} \xi_i^2 + \frac{\lambda}{2}\|w\|_2^2$

Example :- Let $x \in R$.

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

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

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

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

10th Feb. 2023

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

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

$\phi(x_i)$ $\phi(x_j)$

(*) (*)

(*) **Low Rank Approximation :-**

$$\min_{\hat{A} \in R^{n \times d}} \|A - \hat{A}\|_F^2$$

subject to $\text{rank}(\hat{A}) = k$

$k << d$

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

orthonormal.

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

solution

(*) (*)

(*) **SMOTE :-** Synthetic Minority Over-Sampling Technique.

→ Avoid overfitting due to exact replicas of minority class samples.
→ Subset of minority class is taken.
→ New synthetic data samples ~~are~~ created from subsets and added.

No overfitting. Reduced overfitting	mainly introduce noise.

27th Feb 2023

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

Posterior — Likelihood — Prior

Max

Probability [0,1]

Likelihood ≥ 1 (can be)

$D = \{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$

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

$$P((y_1,y_2,\ldots,y_n)|x_1,x_2,\ldots,x_n,\theta) = L(\theta)$$

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

⇒ Form of prior similar to posterior then it is called as conjugate prior.

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

$$= \arg\max_{\theta} \sum_{i=1}^{n} \log P(y_i|x_i,\theta)$$

⇒ we use log likelihood since, it is montonically increasing.

$$= \arg\min_{\theta} \sum_{i=1}^{n} -\log P(y_i|x_i,\theta)$$

→ negative log likelihood.

loss at single point → $\ell(\theta; (x, y)) = -\log P(y|x, \theta)$

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

27th Feb 2023

$P(y_i | x_i, (w, \beta)) = N\left(y_i | w^T x_i, \frac{1}{\beta}\right)$

$\theta = (w, \beta)$, $\beta = \frac{1}{\sigma^2}$

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

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

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

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

for multivariate,

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