

# Machine Learning Notes

---

-Statistical Learning Theories

-Model in Mathematical Details

-Logistic Regression and Regularization

-AdaBoost

-Support Vector Machine

- Gaussian Mixture Model and EM Algorithm

# \* Statistical Learning / Machine Learning model

Can be categorized as:

	Generative	Discriminative
Parametric	Discriminant analysis (LDA, QDA) NB	Logistic Regression Neural Network
Non-Parametric		KNN - tree-based method - GMM

{ generative : use bayesian rule to get  $Pr(Y|X)$   
discriminative: don't use bayesian rule,  
directly calculate  $Pr(Y|X)$

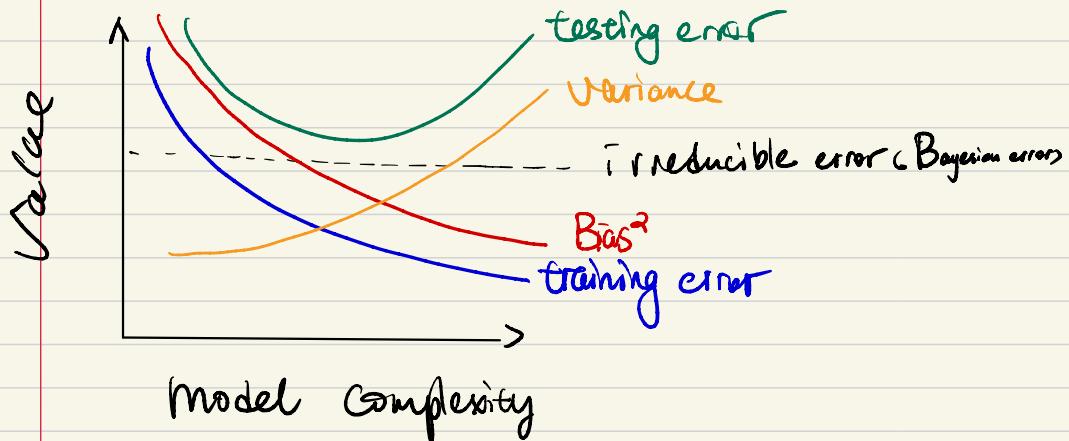
{ Parametric : has functional form ,

non-Parametric : does not have a functional form  
• # params change according to size of training

## Figure of

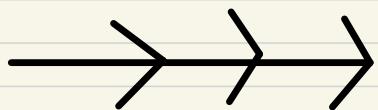
\* Training, Testing error

Bias, variance trade-off

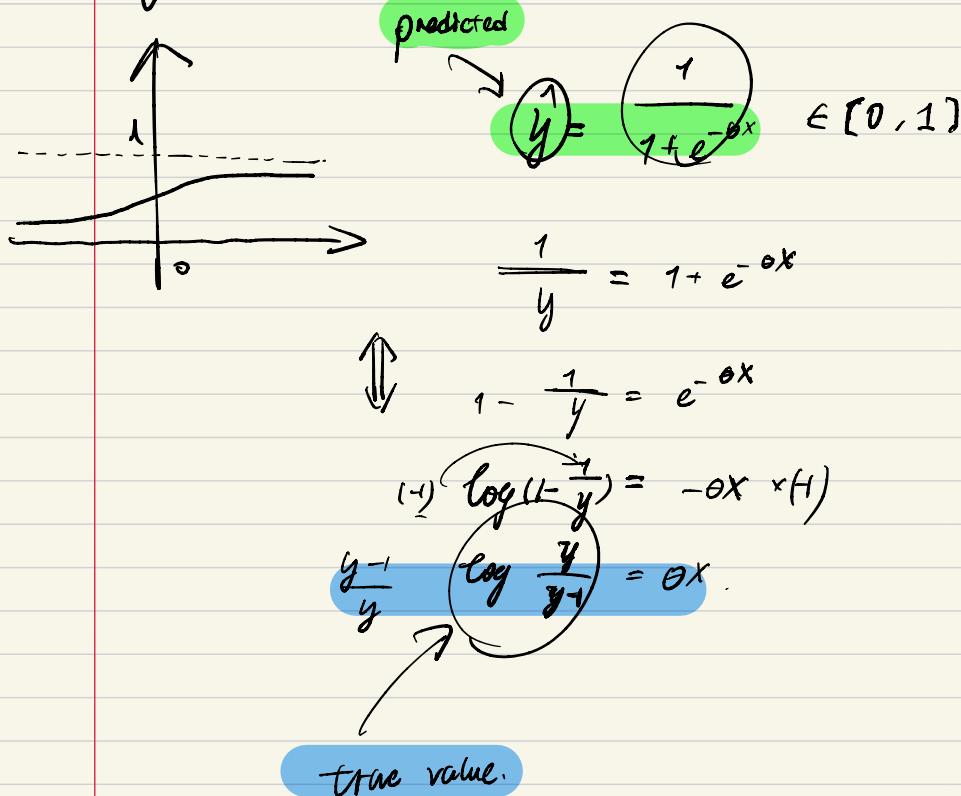


- \* test error =  $\text{bias}^2 + \text{variance} + \delta^2$
- \* test error can be no smaller than Bayesian error, which is a theoretical error.

Details of each Model / method



# Logistic Regression



loss:

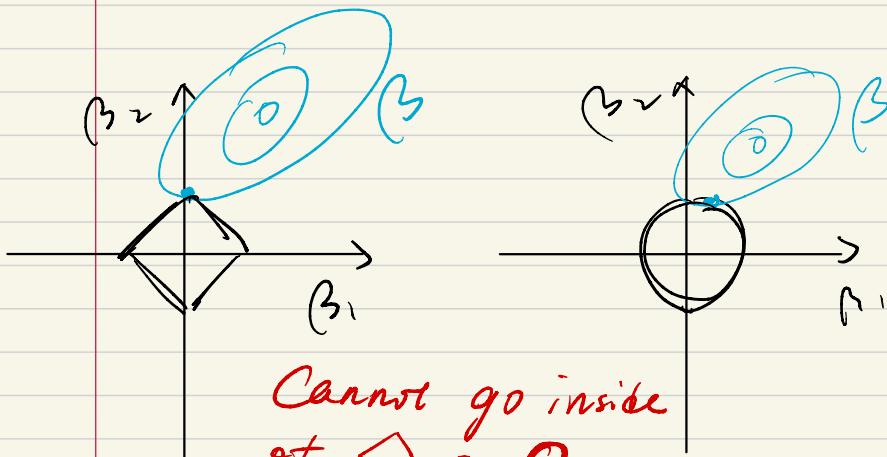
$$\text{Obj. function } h_{\theta}(x) = \frac{1}{1 + e^{\theta x}}$$

$$-\log L(\theta) = -\sum \log P(y|x, \theta)$$

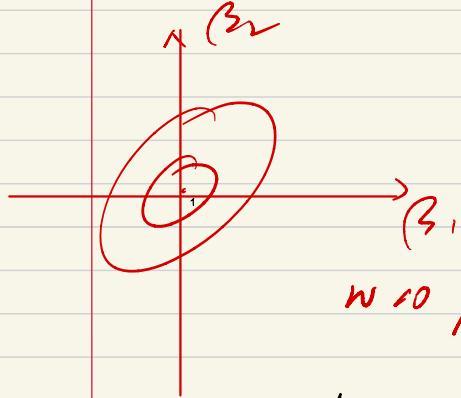
$$= -\sum [y \log(h_{\theta}(x)) + (1-y) \log(1-h_{\theta}(x))]$$

$$\sum -y \log y - (1-y) \log(1-y)$$

# $\ell_1$ & $\ell_2$ Penalty



$\ell_1$  Lasso



$\ell_2$  ridge

$$\text{Loss Func: } \underbrace{\sum \log(y_i \cdot \hat{y}_i)}_{\lambda \|\beta\|^2}$$



# Ada Boost

for each iteration

# calculate classification error

$$\text{err}_m = \frac{\sum w_i \mathbb{1}(T_m(x_i) \neq y_i)}{\sum w_i}$$

$\sum w_i$  ← weight of each sample

# tree weight

$$\alpha_m = \ln \frac{1 - \text{err}_m}{\text{err}_m} \quad \leftarrow \text{weight of each tree}$$

$$< \text{err}_m \uparrow \Rightarrow \ln \dots \downarrow \Rightarrow \alpha_m \downarrow >$$

# update sample weight

$$W_i' \leftarrow W_i e^{\alpha_m \cdot \mathbb{1}(y_i \neq T_m(x_i))} = \begin{cases} W_i & \text{if } y_i = T_m(x_i) \\ W_i e^{\alpha_m} & \text{if } y_i \neq T_m(x_i) \end{cases}$$

## Normalize

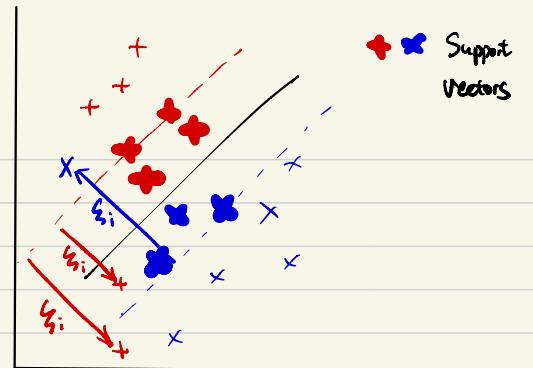
$$W_i'' \leftarrow \frac{W_i'}{\sum W_i'}$$

# SVM

- Objective function:

$$\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + \frac{C}{n} \sum_{i=1}^n \xi_i$$

s.t.  $y_i(w^T x_i + b) \geq 1 - \xi_i \quad \forall i$   
 $\xi_i \geq 0$



- first part of the equation:

Maximize margin(min distance) to separating hyperplane.

$$\max_{w, b} \min_{i=1 \dots n} \frac{|w^T x_i + b|}{\|w\|} \quad \text{s.t. } y_i(w^T x_i + b) \geq 1 \quad \forall i$$

Normalized distance \*

y\_i(w^T x\_i + b) = 1 \quad \exists i \quad \begin{cases} \text{Canonical form} \\ \text{of sep. hyp.} \end{cases}

Since the minimize inner gives 1,

$$\Rightarrow \max_{w, b} \frac{1}{\|w\|} \Rightarrow \min \frac{1}{2} \|w\|^2$$

- it's a optimal Soft margin Hyperplan problem.

where  $\xi_i$  is introduced to allow Misclassification

**error** in non-linear separable cases.

# Solution to the problem:

- Way 1: replace  $\xi$  with **Hinge Loss** and use **Stochastic Gradient Descent** to solve

$$\min \frac{1}{2} \|w\|^2 + \frac{C}{n} \sum_{i=1}^n \max(0, 1 - y_i(w^T x_i + b))$$

- Way 2: Instead of solving primal (min) problem  
we can solve it by dual (max) and then recover to primal.

benefit: ① efficient when  $d \gg n$   
(only use optimal w)

② can kernelize the problem.

- Write as Lagrangian:

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

- Objective function is then,  $\max_{\alpha \geq 0, \beta \geq 0} \min_{w, b, \xi} L(w, b, \xi, \alpha, \beta)$

- Solving inner part

$$\sum \alpha_i y_i = 0$$

$$\textcircled{1} \quad \frac{\partial L}{\partial w} = 0, \quad \textcircled{2} \quad \frac{\partial L}{\partial b} = 0, \quad \textcircled{3} \quad \frac{\partial L}{\partial \beta} = 0 \quad \Rightarrow \frac{C}{n} - 2i - \beta_i = 0$$

$$\Rightarrow \mathcal{L}_0(\alpha, \beta)$$

$$= -\frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j \langle x_i^T x_j \rangle + \sum_i \alpha_i \quad \text{if } \textcircled{2} \& \textcircled{3}$$

- $\max \mathcal{L}_0(\alpha, \beta)$

$$\underset{\alpha, \beta}{\max} -\frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j \langle x_i^T x_j \rangle + \sum_i \alpha_i \quad \text{s.t. } \textcircled{1}, \textcircled{2},$$

$$\alpha_i \geq 0, \beta_i \geq 0$$

Replacing  $\beta$

$$\underset{\alpha}{\max} -\frac{1}{2} \sum \alpha_i \alpha_j \langle x_i^T x_j \rangle + \sum \alpha_i \quad \text{s.t. } \textcircled{1}$$

$$0 \leq \alpha_i \leq \frac{C}{n}$$

- Recover to primal by KKT

Condition

primal optimal params:

$$w^* = \sum_{i=1}^n \alpha_i^* y_i x_i, \quad b^* = y_i - \sum_i \alpha_i^* y_i \langle x_i^T x_i \rangle$$

the final classifier is.

$\textcircled{1}$

$$f(x) = \text{Sign} (\langle w^* \cdot x \rangle + b) =$$

$$\text{Sign} (\sum_{i=1}^n \alpha_i^* y_i \langle x_i^T \cdot x \rangle + b^*) \quad \textcircled{2}$$

# Gaussian mixture model < model-based clustering

GMM

$$P(x) = \sum_{k=1}^K P(Z=k) f_{\text{within cluster}}(x|Z=k)$$

cluster distribution      Gaussian distribution.

①

$\rightarrow E\text{-M}$ : iterate between  $E$  &  $M$ .

Estimate log likelihood,  $Z_{ik} \in Z_i = \{2_1, \dots, 2_K\}$ , one  $k$ , not 0  
 $\Rightarrow \pi_{ik} = P(Z_{ik}=1/x_i)$

Maximize param  $\pi_k, \mu_k, \Sigma_k$

$$\pi_k = \frac{\sum \pi_{ik}}{n}$$

$$\mu_k = \frac{\sum \pi_{ik} x_i}{\sum \pi_{ik}}$$

$$\Sigma_k = \frac{\sum \pi_{ik} (x_i - \mu_k)(x_i - \mu_k)^T}{\sum \pi_{ik}}$$

$$\Rightarrow Z_{ik} \leftarrow \underset{k}{\operatorname{argmax}} \pi_{ik}$$

②

$\rightarrow$  BIC select  $K$ .

$$\frac{2 \log l(x, \hat{\theta}) - d \log n}{6}$$

total # params

③

$\rightarrow$  Sensitive to outliers

$\Sigma_k$  assumed equal & spherical,  $\pi_k$  are assumed equal  $\Rightarrow K\text{-means}$

More on

## E-M Algorithm ...

E-step: Given current parameter value  $\theta^{old}$ ,

find the posterior distribution of  $Z$  given  $X$ .

$$P(Z|X, \theta^{old})$$

$\log P(X|\theta^{old})$  is bounded by expectation.

$$\log p(x|\theta^{old}) \geq Q(\theta, \theta^{old})$$

$$= \sum p(Z|x, \theta^{old}) \log p(x, Z|\theta)$$

M step: Locally maximize the low bound  $Q(\theta, \theta^{old})$  over the parameter  $\theta$

$$\theta^{new} = \underset{\theta}{\operatorname{argmax}} Q(\theta, \theta^{old})$$

Mixture Gaussian Case:

$$\begin{aligned} E: \quad \gamma_{CZK} &= P(C_{ZK}=1 | X) = \frac{P_{(Z_k=1)} P_{(X|Z_k=1)}}{P(x)} \\ &= \frac{\pi_k N(x | \mu_k, \Sigma_k)}{\sum_{j=1}^k \pi_j N(x | \mu_j, \Sigma_j)} \end{aligned}$$

M: Log-likelihood of observing the data sample  $x$

$$\begin{aligned}
 \log p(x) &= \sum_{k=1}^K \gamma(z_k) \log p(x) \\
 &= \sum_{k=1}^K \gamma(z_k) \log \frac{p(x, z_k)}{p(z_k | x)} \\
 &= \boxed{\sum_{k=1}^K \gamma(z_k) \log p(x, z_k)} \\
 &\quad - \boxed{\sum_{k=1}^K \gamma(z_k) \log \pi(z_k)} \quad \text{corr.}
 \end{aligned}$$

Maximize log likelihood of observing the data samples

$$\begin{aligned}
 \max L &= \sum \sum \gamma(z_{nk}) \log \pi_k + \frac{1}{2} \sum \sum \gamma(z_{nk}) \log |\Sigma_k^{-1}| \\
 &\quad - \frac{1}{2} \sum \sum \gamma(z_{nk}) (x^{(n)} - \mu_k)^T \Sigma_k^{-1} (x^{(n)} - \mu_k)
 \end{aligned}$$

$$\text{s.t. } \sum_{k=1}^K \pi_k = 1$$

$$\frac{\partial L}{\partial \mu_k} = 0 \Rightarrow \mu_k = \frac{\sum_{n=1}^N \gamma(z_{nk}) x^{(n)}}{\sum_{n=1}^N \gamma(z_{nk})}$$

$$\frac{\partial L}{\partial \Sigma_k} = 0 \Rightarrow \Sigma_k = \frac{\sum_{n=1}^N \gamma(z_{nk}) (x^{(n)} - \mu_k)^T}{\sum_{n=1}^N \gamma(z_{nk})}$$

using Lagrange multiplier

$$\underbrace{\partial (L - \alpha(\sum \pi_k - 1))}_{\partial \pi_k} = 0 \Rightarrow \pi_k = \frac{\sum_{n=1}^N \gamma(z_{nk})}{N}$$