Machine Learning Notes

* Statistical Learning / Machine Learning model can be categorized as:

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

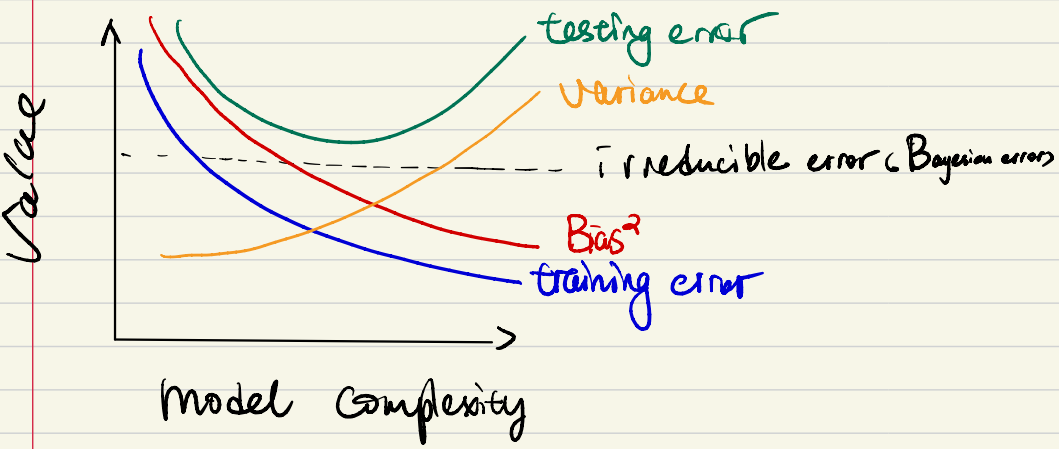
{ generative : use bayesian rule to get $Pr(Y|X)$
discriminative: don't use bayesian rule, directly calculate $Pr(Y|X)$

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

Figure of

* Training, Testing error

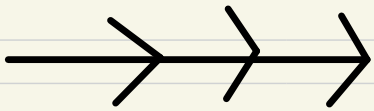
Bias, variance trade-off



* test error = bias^2 + variance + δ^2

* test error can be no smaller than Bayesian error, which is a theoritical error.

Details of each Model (/method)



Logistic Regression

predicted

$$y = \frac{1}{1+e^{-\theta x}} \in [0,1]$$

$$\frac{1}{y} = 1 + e^{-\theta x}$$

$$\Updownarrow$$

$$1 - \frac{1}{y} = e^{-\theta x}$$

$$(-1)\ \log\left(1 - \frac{1}{y}\right) = -\theta x \quad \times(-1)$$

$$\log \frac{y}{y-1} = \theta x$$

$$\frac{y-1}{y}$$

true value.

loss :

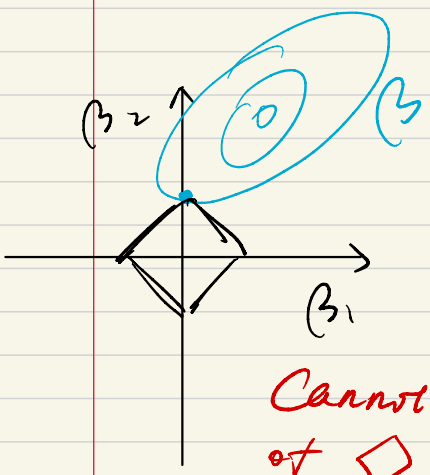
Obj. function $h_\theta(x) = \frac{1}{1+e^{\theta^T x}}$

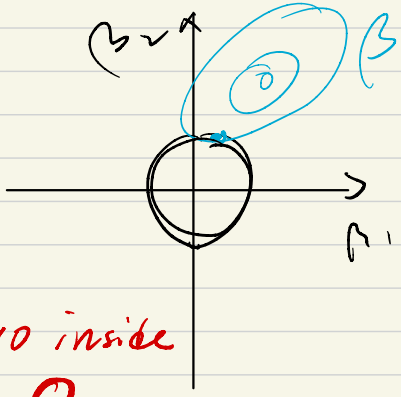
$$-\log L(\theta) = -\sum \log P(y \mid x, \theta)$$

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

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

l_1 & l_2 Penalty

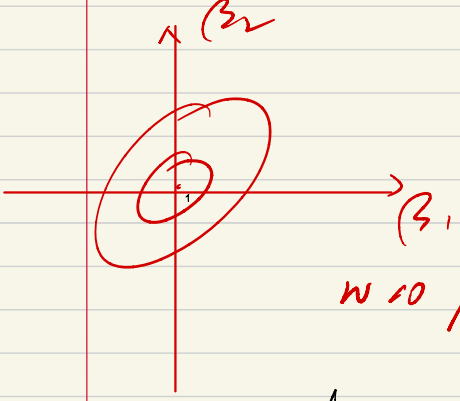




Cannot go inside of ◇ or O

l_1 Lasso

l_2 ridge



w/o penalty.

$$L(\beta) = \text{Loss Fcn: } \sum \log(y, \hat{y}) + \lambda \|\beta\|^2$$

$\lambda \|\beta\|$

Ada Boost

for each iteration

calculate classification error

$$err_m = \frac{\sum w_i 1(T_n(x_i) \neq y_i)}{\sum w_i}$$

w_i ← weight of each sample

tree weight

$$\alpha_n = \ln \frac{1 - err_m}{err_m}$$

α_n ← weight of each tree

$< err_m \uparrow \Rightarrow \ln \cdots \downarrow \Rightarrow \alpha_n \downarrow >$

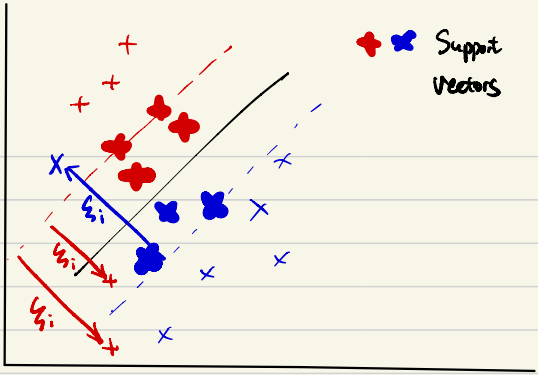
update sample weight

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

Normalize

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

SVM



- Objective function:

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

$$\text{s.t. } y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i \quad \forall i$$

$$\xi_i \geq 0$$

- first part of the equation:

  maximize margin (min distance) to separating hyperplane.

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

  → * normalized distance *

  } canical form of sep. hyp.

  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 ξ_i is introduced to allow misclassification error in non-linear seperable cases.

Solution to the problem:

- Way 1: replace ξ with **hinge Loss** and use **Stochastic Gradient Descent** to solve

$$\min \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum^{n} \max(0, 1-y_i(w^T x+b))$$

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

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

  ② can kernalize 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

$\Sigma \alpha_i y_i = 0$

$\frac{C}{n} - \alpha_i - \beta_i = 0$

$$\frac{\partial L}{\partial w} = 0 \quad (1), \qquad \frac{\partial L}{\partial b} = 0 \quad (2), \qquad \frac{\partial L}{\partial \xi} = 0 \quad (3)$$

$$\Rightarrow L_D(\alpha, \beta) = -\frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j x_i^T x_j + \sum_i \alpha_i \quad \text{if } ② \& ③$$

- $\max L_D(\alpha, \beta)$

$$= \max_{\alpha,\beta} -\frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j \langle x_i x_j \rangle + \Sigma \alpha_i \quad \text{s.t. } ①, ②, \alpha_i \geq 0, \beta_i \geq 0$$

replacing β

$$= \max_{\alpha} -\frac{1}{2}\Sigma \alpha_i \alpha_j \langle x_i, x_j \rangle + \Sigma \alpha_i \quad \text{s.t. } ① \quad 0 \leq \alpha_i \leq \frac{C}{n}$$

- Recover to primal by KKT Condition

primal optimal params:

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

the final classifier is.

$$f(x) = \text{sign}(\langle w^*, x \rangle + b) = \text{sign}\left(\sum_{i=1}^{n} \alpha^* y_i \langle x_i, x \rangle + b^*\right)$$

Gaussian mixture model <model-based clustering>

GMM

$$p(x) = \sum_{k=1}^{K} p(z=k) f(x|z=k)$$

$p(z=k)$: cluster distribution; $f(x|z=k)$: within cluster Gaussian distribution.

①

→ EM: iterate between E & M.

Estimate loglikelihood, $z_{ik} \in z_i = (z_1, \dots z_k)$. one 1, rest 0.

$\Rightarrow r_{ik} = E(z_{ik}|x_i)$

$= p_i(z_{ik}=1|x_i)$

Maximize param π_k, μ_k, Σ_k

$\Rightarrow \hat{z}_{ik} = \arg\max_k r_{ik}$ ← 0/1, by

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

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

$$\Sigma_k = \frac{\sum_i r_{ik}(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_i r_{ik}}$$

②

→ BIC select K.

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

↓ total # params

③

→ Sensitive to outliers

Σ_k assumes equal & spherical, π_k are assumed equal. => Kmeans

More on E-M Algorithm ...

E-step: Given current parameter value θ^{old},
find the posterior distribution of Z given X.

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

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

$$\log p(X|\theta^{old}) \geq Q(\theta, \theta^{old})$$
$$= \sum p(Z|X, \theta^{old}) \ln p(X, Z|\theta)$$

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

$$\theta^{new} = \arg\max_{\theta} Q(\theta, \theta^{old})$$

Mixture Guassian Case:

E:

$$\gamma(z_k) = p(z_k = 1|X) = \frac{p(z_k = 1)\, p(X|z_k = 1)}{p(X)}$$
$$= \frac{\pi_k N(X|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(X|\mu_j, \Sigma_j)}$$

M: Log-likelyhood of observing the data sample X

$$\log p(x) = \sum_{k=1}^{K} \gamma(z_k) \log p(x)$$

$$= \sum_{k=1}^{K} \gamma(z_k) \log \frac{p(x, z_k=1)}{p(z_k=1|x)}$$

$$= \boxed{\sum_{k=1}^{K} \gamma(z_k) \log p(x, z_k=1) - \sum_{k=1}^{K} \gamma(z_k) \log \gamma(z_k)}$$

(cont.)

$$\vdots$$

Maximize log likelyhood of observing the data samples

$$\max L = \sum_{n}^{N} \sum_{k}^{K} \gamma(z_{nk}) \log \pi_k + \frac{1}{2} \sum_{n}^{N} \sum_{k}^{K} \gamma(z_{nk}) \log |\Sigma_k^{-1}|$$

$$- \frac{1}{2} \sum_{n}^{N} \sum_{k}^{K} \gamma(z_{nk}) (x^{(n)} - \mu_k)^T \Sigma_k^{-1} (x^{(n)} - \mu_k)$$

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

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

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

using Lagrangian multiplier

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