1. 教材：简老师手稿 + ppt.（无教材，参考教材 PRML）
(Bishop' 2006)

Day 1:

1. Introduction
2. Probability Distribution (data & compute)
3. Linear Regression (输出连续)
4. Linear Classification (输出离散)
5. Kernal Method. (another trick)
6. Sparse Kernal Method ≡ SVM → support vector
7. Mixture Model & EM.
8. Approximate Inference.

1. Introduction

Supervised → classification
Supervised → regression
have a teacher
label

Unsupervised → cluster
without teacher

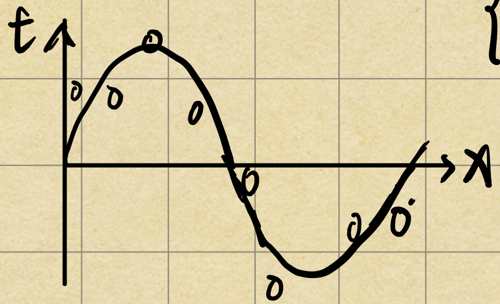
$$\underbrace{\{(x_1,t_1),(x_2,t_2),\cdots,(x_n,t_n)\}}\Rightarrow x\xrightarrow{f}t$$

mapping

1' Random varible

2' train / testing → {prediction / Generalization (泛化)}

2. Generalization: fitting Sin()



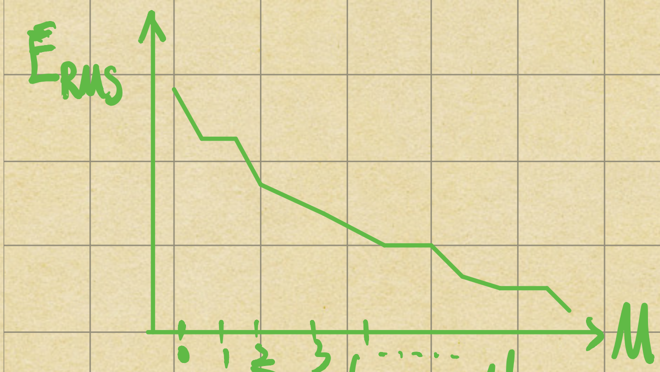
$\{x_n, t_n\}^N$ $y(x, w) \to t$

1. Assume $y(x, w) = w_0 + w_1 x + \cdots + w_M x^M$
2. Learning Criterion / Objective:
   Error function & Divergence.

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2$$

root-mean-squares:

$$\sqrt{2\cdot E(w^*)/N} = E_{RMS}$$



⇒ larger M, more complex the Model, less Error (Overfitting)

how to trade-off the overfitting and underestimating?

$$\tilde{E} = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2 + \frac{\lambda}{2}\|w\|^2$$ adjust error and w.

hyper parameter. (regularization term): λ, M

$\|w\| = w^T \cdot w$

regularization

reduce the value of w when M increase

Add probability into it:

Prior / likelihood / Posterior

$$P(w|D) = \frac{P(D|w)\cdot P(w)}{P(D)}$$

D: Data

in disperse condition: $P(D) = \sum_{w} P(w)\cdot P(D|w)$

$P(D) \to \int P(D|w)\cdot P(w)\,dw$ (evidence)

$$P(w|D) \propto P(D|w)\cdot P(w)$$

For model selection

Supplement: $\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma}\cdot \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) > 0$

$y(x, w)$, $x \in R$, $w \in R^{M+1}$

3.

with D-dimension Gaussian:

$$N(x|\mu,\Sigma)=\frac{1}{(2\pi)^{\frac{D}{2}}}\cdot\frac{1}{|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(x-\mu)^T\cdot\Sigma^{-1}(x-\mu)\right\}$$

with 1-dimension, we obtain probability of dataset:

$$p(x|\mu,\delta^2)=\prod_{n=1}^{N}N(x_n|\mu,\delta^2)$$

then likelihood:

$$\ln P(x|\mu,\delta^2)=\ln(N(x_1|\mu,\delta^2)\cdot N(x_2|\mu,\delta^2)\cdots N(x_n|\mu,\delta^2))$$

$$=-\frac{1}{2\delta^2}\sum_{n=1}^{N}(x_n-\mu)^2-\frac{N}{2}\ln\delta^2-\frac{N}{2}\ln(2\pi)$$

Maximum Likelihood:

$$\frac{d}{d\mu}(\ln P(x|\mu,\delta^2))=\frac{1}{\delta^2}\sum_{n=1}^{N}(x_n-\mu)\xrightarrow{\text{let}=0}\sum_{n=1}^{N}x_n=\sum_{n=1}^{N}\mu$$

$$\therefore \mu_{ML}=\frac{1}{N}\sum_{n=1}^{N}x_n,\quad \text{likewise}\quad \delta_{ML}^2=\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2$$

Maximum Likelihood (especially W)

$$\mu^*=\frac{1}{N}\cdot\sum_{n=1}^{N}x_n=\mu$$

Define: $\Sigma^*=\frac{1}{N-1}\sum_{n=1}^{N}(x_n-\mu^*)^2$

$$P(t|x,W,\beta)=N(t_n|y(x_n,w),\beta^{-1})$$

$\beta^{-1}=\frac{1}{\delta^2}=\frac{1}{\Sigma}$ as precision

$$\therefore P(D|W)=P(t|x,W,\beta)=\prod_{n=1}^{N}N(t_n|x,W,\beta^{-1})$$

$$\therefore \ln P(t|x,W,\beta)=-\frac{\beta}{2}\boxed{\sum_{n=1}^{N}(y(x_n,w)-t_n)^2}+\frac{N}{2}\ln\beta-\frac{N}{2}\ln(2\pi)$$

Let $\beta^{-1}=\frac{1}{N}\sum_{n=1}^{N}\{y(x_n,W_{ML})-t_n\}^2$

$$E(W)\longleftrightarrow\ln P(D|W)$$

$$W_{MAP}=\underset{W}{\operatorname{argmax}}\,P(W|X,t,\alpha,\beta)=\underset{W}{\operatorname{argmin}}\left\{\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,w)-t_n\}^2+\frac{\alpha}{2}\cdot W^T\cdot W\right\}$$

Maximum a Posterior

4.

Proof above: Given $\ln p(t|x,w,\beta)$

$$= -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,w)-t_n\}^2+\frac{N}{2}\ln\beta-\frac{N}{2}\ln(2\pi)$$

β is irrelevant to w_{ML} ★, after we calculate w_{ML}

then $\beta_{ML}=\frac{1}{N}\sum_{n=1}^{N}\{y(x_n,w_{ML})-t_n\}^2$

finally: we obtain "predictive distribution"

$$p(t|x,w_{ML},\beta_{ML})=\mathcal{N}(t|y(x,w_{ML}),\beta_{ML}^{-1})$$

Section MAP:

Given $p(w|\alpha)=\mathcal{N}(w|0,\alpha^{-1}I)=\mathcal{N}(w|0,\alpha^{-1}I)$

$$=\left(\frac{\alpha}{2\pi}\right)^{\frac{M+1}{2}}\exp\left(-\frac{\alpha}{2}w^T\cdot w\right)$$

$$\ln(p(w|\alpha))=\frac{M+1}{2}\ln\frac{\alpha}{2\pi}-\frac{\alpha}{2}w^T\cdot w$$

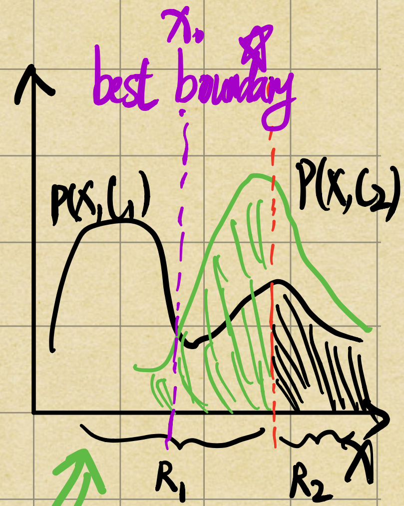
$$\therefore P(w|x,t,\alpha,\beta)\propto P(t|x,w,\beta)\cdot P(w|\alpha)$$

posterior distrib   prior distrib   likelihood

Data flow: $w_{ML}\xrightarrow{P(t|\cdot)}\beta(\text{precision})\xrightarrow{P(w|\cdot)}w_{MAP}$

Decision Theroy:

C (presence of cancer) $\begin{cases} C_1: \text{with cancer} \\ C_2: \text{without cancer.} \end{cases}$



$$\hat{C}_{MAP}=\arg\max P(C_k|x)=\arg\max\frac{P(C_k)\cdot P(x|C_k)}{P(x)}$$

1. Minimizing the misclassification Rate

$$P(\text{mistake})=P(x\in R_1,C_2)+P(x\in R_2,C_1)$$

$$=\int_{R_1}P(x,C_2)dx+\int_{R_2}P(x,C_2)dx$$

★ 2. Minimizing the expected loss

good. Loss: Minimum Bayes Risk <MBR>

Expected Loss

cancer normal

5.

Loss matrix

$$\begin{array}{l} \text{cancer} \\ \text{normal} \end{array}\begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} = L = [L_{kj}]$$

$$E[L] = \sum_k \sum_j \int_{R_j} L_{kj} \cdot \underbrace{P(x, C_k)}_{P(x) \cdot P(C_k|x)} \cdot dx \Rightarrow \hat{C}_j = \arg\min \sum_k L_{kj} P(C_k|x)$$

$$P(C_k|x) \quad \frac{P(x|C_k) \cdot P(C_k)}{P(x)} \rightarrow \text{generative model (indirect model)}$$

↓

discriminative model (direct model)

* loss function for regression:

$$E[L] = \iint L(t, y(x)) \cdot P(x,t)\, dx\, dt$$

$$= \iint \{y(x) - t\}^2 \cdot P(x,t)\, dx \cdot dt \qquad \int y(x) P(x,t) dt = \int t \cdot P(x,t) dt$$

$$\frac{\partial E[L]}{\partial y(x)} = 2 \int \{y(x) - t\} P(x,t)\, dt = 0$$

$$y^*(x) = \frac{\int t\, P(x,t) \cdot dt}{P(x)} = E_t[t|x]$$

non-prob	probability
LS (least square)	ML (Maximum likelihood)
↓ upgrade	↓
RLS (regularization least square)	MAP (Maximum a posterior)

Supplement:

$$P(t|x, X, t) = \int P(t|x,w) \cdot P(w|X,t) \cdot dw = N(t|m(x), s^2(x))$$

test data (t), training data (X, t)

$m(x) =$ ☆

$s^2(x) =$

6.

Information Theory: "Entropy" $\Leftrightarrow$ "uncertainty"

KL "divergence":

Kullback-Leibler

Defination: Two unrelated events, $h(x,y) = h(x) + h(y)$ & $h(x) = -\log_2 P(x)$ bits

$H[X] = -\sum_x P(x) \cdot \log_2 P(x)$ [H is a expection of information]

$H[p] = -\sum_i P(X_i) \cdot \ln P(X_i) = -\int p(x) \cdot \ln p(x) \cdot dx$

$= E[-\ln P(X)]$

for P is ln

for X_i is $\log_2$

Maximum Entropy (ME)

$\widetilde{H} = -\sum_i P(X_i) \ln P(X_i) + \lambda \left(\sum_i P(X_i) - 1 \right)$

Given: x is continuous random varibles:

$\Rightarrow H[X] = -\int p(x) \cdot \ln p(x) \cdot dx$ (differential entropy)

Given three constraints we have:

$-\int p(x) \ln p(x) dx + \lambda_1 \left(\int_{-\infty}^{+\infty} p(x) dx - 1 \right) + \lambda_2 \left(\int_{-\infty}^{+\infty} x p(x) dx - \mu \right) +$
$\lambda_3 \left(\int_{-\infty}^{+\infty} (x-\mu)^2 p(x) dx - \delta^2 \right)$

$\Rightarrow p(x) = \exp\{ -1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 \}$

finally we obtain $\Rightarrow p(x) = \frac{1}{\sqrt{2\pi}\delta} \exp\left(-\frac{(x-\mu)^2}{2\delta^2}\right)$ ※

then $H[X] = \frac{1}{2}\{ 1 + \ln(2\pi\delta^2) \}$

Entropy versus Pattern Recognition:

$KL(p \| q) = -\int p(x) \ln q(x) dx - \left(-\int p(x) \cdot \ln p(x) \cdot dx\right)$

$= -\int p(x) \ln \left\{ \frac{q(x)}{p(x)} \right\} dx \geq 0$

unknow true distrib — approx distrib — $\int p(x) \ln\left\{\frac{\cdot}{p(x)}\right\} dx \geq 0$

& $KL(p\|q) \neq KL(q\|p)$

proof: with Jensens Inequality we obtain:

① $f(\lambda a + (1-\lambda) b) \leq \lambda f(a) + (1-\lambda) f(b)$

$\Rightarrow f(E[x]) \leq E[f(x)]$

② $KL(p\|q) = -\int p(x) \cdot \ln\left\{\frac{q(x)}{p(x)}\right\} dx \geq -\ln \int q(x) \cdot dx = 0$

Parametic Model: estimate θ

$\hat{\theta} = \arg\max_{\theta} KL(p(x) | q(x|\theta))$ & $KL(p\|q) \approx \sum_{n=1}^{N} \{-\ln q(x_n|\theta) + \ln p(x_n)\}$

Mutual Information: according to independence

$I(x,y) = KL(p(x,y) \| p(x) \cdot p(y))$

$= -\iint p(x,y) \ln\left(\frac{p(x) \cdot p(y)}{p(x,y)}\right) \geq 0$

Day 2: Probability Distribution:

Intro: Discrete: (Natural Language)

① Binomial - beta

② Multinomial - Dirichlet

likelihood - prior

conjugate - prior

Continuous:

③ Gausian - Gausian-Gamma

④ Student's t

sequential learning

⇓

Exponential Family

Background:

$x = [x_a^T\ x_b^T]^T \left(\begin{bmatrix} x_a \\ x_b \end{bmatrix}\right) \propto$ $P(x_a|x_a)$ $P(x_a|x_a)$

1. Bernoulli Distribution $x \in \{0,1\}$

Method 1: Maxmium Likelihood

In Bernoulli, Given Dataset $D=\{x_1, \cdots, x_N\}$, we obtain the likelihood function $P(D|\mu) = \prod_{n=1}^{N} P(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$

$$\Rightarrow \ln P(D|\mu) = \sum_{n=1}^{N}\{x_n \ln\mu + (1-x_n)\ln(1-\mu)\} \overset{let}{=} 0$$

$\therefore \mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$ detection: overfitting while given a small dataset.

Method 2: Posterior Bayian.

$$Bin(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m} \quad \text{where } \binom{N}{m} = \frac{N!}{(N-m)!\,m!}$$

let introduce prior probability distrib $P(\mu)$:

$P(\mu)$ must have conjugacy for the same functional formation!

Definition:

$$\underbrace{Beta(\mu|a,b)}_{\text{hyper param}} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}\cdot(1-\mu)^{b-1}$$

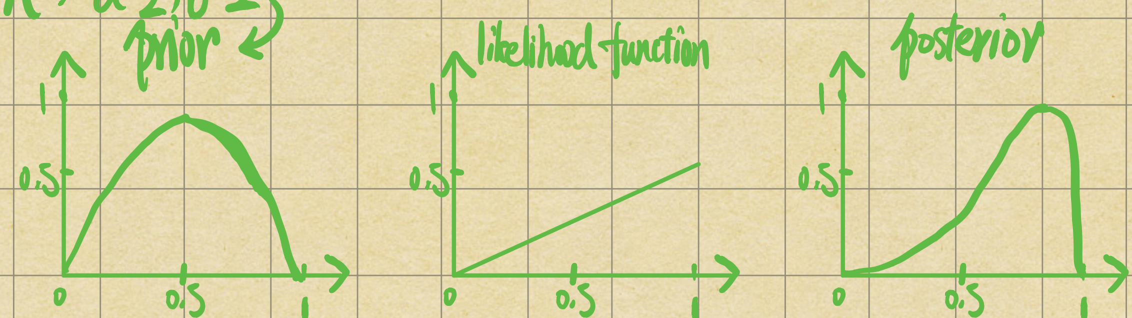
$$E[\mu] = \frac{a}{a+b}, \quad var[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$

$$P(\mu|m,l,a,b) \propto Bin(m|N,\mu)\cdot Beta(\mu|a,b)$$

$$P(\mu|m,l,a,b) \propto \mu^{m+a-1}\cdot(1-\mu)^{l+b-1} \quad \text{(without regularization)}$$

$$\therefore P(\mu|m,l,a,b) \propto \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}\cdot(1-\mu)^{l+b-1}.$$

with: $a=2, b=2$



9.

Conjugate Prior:

1. Sequential Learning: update.
2. Prediction: $\int_0 \cdots dw$ integral.

Example:

$$\lim_{m,l\to\infty} P(x=1|D) = \int_0^1 P(x=1|\mu)\cdot P(\mu|D)\cdot d\mu = \int_0^1 \mu\cdot P(\mu|D)d\mu$$

$$= E[\mu|D] = \frac{m+a}{m+a+l+b} \Rightarrow \frac{m}{N}$$

predictive distrib

2. Multinomial Variables:

$$Mult(m_1, m_2, \cdots, m_k|\mu, N) = \binom{N}{m_1, m_2, \cdots m_k}\prod_{k=1}^{K}\mu_k^{m_k}$$, where

$$\binom{N}{m_1 m_2 \cdots m_k} = \frac{N!}{m_1! m_2! \cdots m_k!}$$ and $\sum_{k=1}^{K} m_k = N$

Dirichlet distribution: (prior for mult varibles)

$$P(\mu|\alpha) \propto \prod_{k=1}^{K}\mu_k^{\alpha_k - 1} \quad 0 \le \mu_k \le 1 \quad \sum_k \mu_k = 1$$

to simplex: → (a bounded linear manifold)

$$Dir(\mu|\alpha) = \underbrace{\frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)}}_{\text{Normalization Term.}}\prod_{k=1}^{K}\mu_k^{\alpha_k - 1}$$

$$P(\mu|D,\alpha) \propto P(D|\mu)\cdot P(\mu|\alpha) \propto \prod_{k=1}^{K}\mu_k^{\alpha_k + m_k - 1}$$

$$\Rightarrow P(\mu|D,\alpha) = Dir(\mu|\alpha + m) = \frac{\Gamma(a_0 + N)}{\Gamma(a_1 + m_1)\cdots\Gamma(a_k + m_k)}\prod_{k=1}^{K}\mu_k^{\alpha_k + m_k - 1}$$

3. Gaussian Distribution:

$$\mathcal{N}(x|\mu,\Sigma)=\frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\underbrace{(x-\mu)^T\cdot\Sigma^{-1}(x-\mu)}_{\Delta}\right\}$$

where Σ is a $D\times D$ variance matrix $|\Sigma|$ is determinant

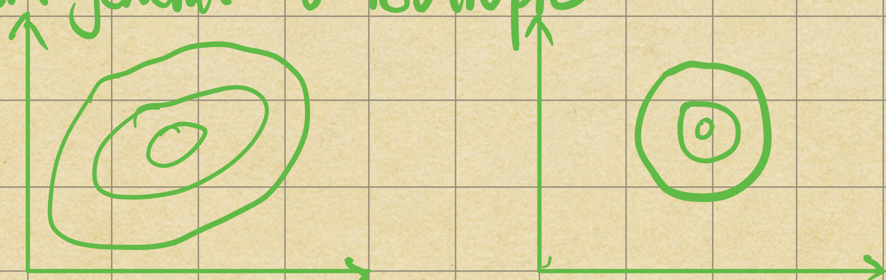
Δ: Mahalanobis distance (二次型的形式)

For Eigenvector Equation:

$$\Sigma u_i=\lambda_i u_i$$

symmetric eigenvector eigenvalue

from general to isotropic



Conditional Gaussian Distribution: $P(x_a|x_b)$

Given: $\mathcal{N}(x|\mu,\Sigma)$, $x=\begin{pmatrix}x_a\\x_b\end{pmatrix}$ $\mu=\begin{pmatrix}\mu_a\\\mu_b\end{pmatrix}$ $\Sigma=\begin{pmatrix}\Sigma_{aa}&\Sigma_{ab}\\\Sigma_{ba}&\Sigma_{bb}\end{pmatrix}$

Let $\Lambda=\Sigma^{-1}$ (precision matrix)

$$\begin{aligned}-\frac{1}{2}(x-\mu)^T\cdot\Sigma^{-1}\cdot(x-\mu)=&-\frac{1}{2}(x_a-\mu_a)^T\Lambda_{aa}(x_a-\mu_a)\\&-\frac{1}{2}(x_a-\mu_a)^T\Lambda_{ab}(x_b-\mu_b)\\&-\frac{1}{2}(x_b-\mu_b)^T\cdot\Lambda_{ba}(x_a-\mu_a)\\&-\frac{1}{2}(x_b-\mu_b)^T\cdot\Lambda_{bb}(x_b-\mu_b)\end{aligned}$$

After complicative equation, we obtain:

$$\mu_{a|b}=\mu_a+\Sigma_{aa}\Sigma_{bb}^{-1}(x_b-\mu_b)$$

$$\Sigma_{a|b}=\Sigma_{aa}-\Sigma_{aa}\Sigma_{bb}^{-1}\cdot\Sigma_{ba}$$

Every quadratic equation could be expressed by a form of Gaussian Distribution via x^2 for Σ, x^1 for mean.

11.

Marginal Gaussian Distribution:

$$P(x_a) = \int P(x_a, x_b)\, dx_b$$

We are integrating out x_b, we find terms involve x_b:

$$-\frac{1}{2} x_b^T \Lambda_{bb} \cdot x_b + x_b^T \{ \underbrace{\Lambda_{bb} \mu_b - \Lambda_{ba}(x_a - \mu_a)}_{m} \}$$

$$= -\frac{1}{2}(x_b - \Lambda_{bb}^{-1} m)^T \Lambda_{bb} (x_b - \Lambda_{bb}^{-1} m) + \frac{1}{2} m^T \Lambda_{bb}^{-1} m$$

$$\Rightarrow \int \exp\{-\frac{1}{2}(x_b - \Lambda_{bb}^{-1} m)^T \Lambda_{bb} (x_b - \Lambda_{bb}^{-1} m)\}\, dx_b$$

then picking up terms depending on x_a:

$$\Rightarrow P(x_a) = N(x_a \mid \mu_a, \Sigma_a)$$

☆☆ Sequential estimation:

Sequential methods allow data points to be processed one at a time and then discarded $\Rightarrow$ on-line application.

Given Conjugate Prior: $P(\mu) = N(\mu \mid \mu_0, \delta_0^2)$

Bayesian inference for Gaussian:

Case 1: unknow mean, known variance (δ^2)

$$\Rightarrow P(X \mid \mu) = \prod_{n=1}^{N} P(x_n \mid \mu) = \frac{1}{\sqrt{2\pi}\delta} \exp\{-\frac{\sum_{n=1}^{N}(x_n - \mu)^2}{2\delta^2}\}$$

$\Rightarrow$ Posterior distribution $P(\mu \mid X) \propto P(X \mid \mu) \cdot P(\mu)$

$$= N(\mu \mid \mu_N, \delta_N^2)$$

$$\mu_N = \frac{\delta^2}{N\delta_0^2 + \delta^2}\mu_0 + \frac{N\delta_0^2}{N\delta_0^2 + \delta^2} \cdot \mu_{ML}$$

$$\frac{1}{\delta_N^2} = \frac{1}{\delta_0^2} + \frac{N}{\delta^2}$$

Case 2: known mean, unkown variance.

$$\Rightarrow \text{likelihood}: P(X \mid \lambda) = \prod_{n=1}^{N} N(x_n \mid \mu, \lambda^{-1})$$

12.

$$\propto \lambda^{\frac{N}{2}} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

then Conjugate prior: In case of using $Gam(\lambda|a_0,b_0)$ }

we have $P(\lambda|x) \propto \lambda^{a_0-1}\lambda^{\frac{N}{2}} \exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right.$

$$= Gam(\lambda|a_N, b_N)$$

where $a_N = a_0 + \frac{N}{2}$, $b_N = b_0 + \frac{N}{2}\sigma_{ML}^2$

Case 3: unknown mean, unknown variance.

$\Rightarrow$ likelihood: $P(X|\mu,\lambda) = \prod_{n=1}^{N}\left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\frac{\lambda}{2}(x_n-\mu)^2\right\}$ }.

$$\propto \left[\lambda^{\frac{1}{2}}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right.$$

Conjugate prior: $P(\mu,\lambda) \propto \left[\lambda^{\frac{1}{2}}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^{\beta}$

$$\exp[c\lambda\mu - d\lambda]$$

$$= \underbrace{\exp\left\{-\frac{\beta\lambda}{2}(\mu - c/\beta)^2\right\}}_{\text{normal}} \lambda^{\beta/2}$$

$$\exp\left\{-\left(d-\frac{c^2}{2\beta}\right)\lambda\right\}$$

☆☆ Student's t-distribution

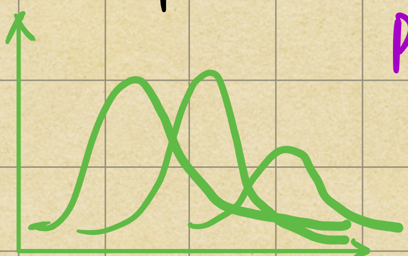
$$P(X|\mu,a,b) = \int_0^{\infty} N(x|\mu,\tau^{-1})\, Gam(\tau|a,b)\, d\tau$$

$= St(x|\mu,\lambda,\nu)$ where: λ is precision & ν is degree of freedom.

when $\nu \to \infty$:

$$St(x|\mu,\lambda,\nu) \to N(x|\mu,\lambda^{-1})$$

☆ mixture of Gaussians:



$P(x) = \sum_{k=1}^{K}\pi_k N(x|\mu_k,\Sigma_k)$, where $\sum_{k=1}^{K}\pi_k = 1$ & $\pi_k \in [0,1]$

$$= \sum_{k=1}^{K} P(k)\cdot P(x|k)$$

Let $\Lambda = \{\pi = \{\pi_1,\cdots,\pi_k\}, \mu = \{\mu_1,\cdots,\mu_k\}, \Sigma = \{\Sigma_1,\cdots,\Sigma_k\}\}$.

☆ The Expential Family (broad class of disturb)

$$P(x|\eta) = h(x)g(\eta)\exp\{\eta^T u(x)\}.$$

13.

Example 1: Bernoulli distrib

$$P(x|\mu)=\mu^{x}(1-\mu)^{1-x}=\exp\{x\ln\mu+(1-x)\cdot\ln(1-\mu)\}$$

$$=(1-\mu)\exp\{\underbrace{\ln\frac{\mu}{1-\mu}}_{\eta}\cdot\underbrace{x}_{u(x)}\}$$

$$\mu=\frac{1}{1+\exp(-\eta)}=\delta(\eta)$$

Example 2: Gaussian disturb

$$P(x|\mu,\delta^2)=\frac{1}{(2\pi\delta^2)^{\frac{1}{2}}}\exp\{-\frac{1}{2\delta^2}(x-\mu)^2\}$$

$$=\frac{1}{(2\pi\delta^2)^{\frac{1}{2}}}\exp\{-\frac{1}{2\delta^2}x^2+\frac{\mu}{\delta^2}x-\frac{1}{2\delta^2}\mu^2\}$$

$$=h(x)\cdot g(\eta)\cdot\exp\{\eta^T u(x)\}.$$

where $\eta=\begin{pmatrix}\frac{\mu}{\delta^2}\\ -1/2\delta^2\end{pmatrix}\begin{matrix}\eta_1\\ \eta_2\end{matrix}$ $u(x)=\begin{pmatrix}x\\ x^2\end{pmatrix}$ $h(x)=(2\pi)^{-\frac{1}{2}}$ $g(\eta)=(-2\eta_2)^{\frac{1}{2}}\exp(\frac{\eta_1^2}{4\eta_2})$

Conjugate Prior:

$$\underbrace{P(\eta|\chi,\nu)}_{\text{Conjugate Prior}}=\underbrace{f(\chi,\nu)}_{\text{normalization coeff}}\cdot g(\eta)^{\nu}\exp\{\nu\eta^T\cdot\chi\}$$

$$\underbrace{P(\eta|X,\chi,\nu)}_{\text{Posterior}}\propto g(\eta)^{\nu+N}\exp\{\eta^T(\sum_{n=1}^{N}u(x_n)+\nu\chi)\}$$

Day 3: Linear Models for regression.

3.1 Linear Basis Function Models: Basis function $\{\phi_j(x)\}\equiv$ feature

$P(t|x)$ is the uncertainty of t for certain x.

① $$y(x,w)=\underbrace{w_0}_{\text{bias parameter}}+\sum_{j=1}^{M-1}w_j\underbrace{\phi_j(x)}_{\text{basis function}}=\sum_{j=0}^{M-1}w_j\phi_j(x)=w^T\phi(x)\quad w=\begin{bmatrix}w_0\\ w_1\\ \vdots\\ w_{M-1}\end{bmatrix}_M$$

$\phi_j(x)=x^j$ { kernel function $k\{x_n,x_m\}=\phi^T(x_n)\cdot\phi(x_n)$

14.

~~对~~ ② $\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$ (实验效果较差)

Gaussian basis function with spatial.

③ sigmoid basis function

$\phi_j(x) = \delta\left(\frac{x-\mu_j}{s}\right)$ & $\delta(a) = \frac{1}{1+e^{-a}}$

$\tanh(a) = 2\delta(a) - 1 = \frac{1-e^{-a}}{1+e^{-a}}$

Maximum likelihood & least squares

$t = y(x, w) + \epsilon \quad \epsilon \propto N(0, \beta^{-1})$ → Gaussian Noise

→ $P(t \mid x, w, \beta) = N(t \mid y(x, w), \beta^{-1})$

→ $E[t \mid x] = \int t\, p(t \mid x) \cdot dt = y(x, w)$

Input data $X = \{x_1, x_2, \cdots, x_N\}$

$P(t \mid x, w, \beta) = \prod_{n=1}^{N} N(t_n \mid w^T \phi(x_n), \beta^{-1})$

$\ln P(t \mid X, w, \beta) = \sum_{n=1}^{N} \ln N(t_n \mid w^T \phi(x_n), \beta^{-1})$

$= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta \left\{\frac{1}{2} \sum_{n=1}^{N} \{t_n - w^T \phi(x_n)\}^2\right\}$

$\nabla_w \ln P(t \mid X, w, \beta) = \sum_{n=1}^{N} \{t_n - w^T \phi(x_n)\} \phi^T(x_n) \overset{let}{=} 0$

$\therefore \sum_{n=1}^{N} t_n \phi(x_n)^T - w^T \left(\sum_{n=1}^{N} \phi(x_n) \cdot \phi(x_n)^T\right) = 0$

$W_{ML} = (\Phi^T \Phi)^{-1} \Phi^T t$, where $\Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_{M-1}(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \cdots & \phi_{M-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix}$

The error function becomes

$E_D(w) = \frac{1}{2} \sum_{n=1}^{N} \{t_n - w_0 - \sum_{j=1}^{M-1} w_j \phi_j(x_n)\}^2$

let $\frac{\partial E_D(w_0)}{\partial w_0} = 0$, $\therefore w_0 = \frac{1}{N} \sum_{n=1}^{N} t_n - \sum_{j=1}^{M-1} w_j \left(\frac{1}{N} \sum_{n=1}^{N} \phi_j(x_n)\right)$

Error function could be seen as a maximum likelihood solution with Gaussian Noise Model.

15.

Sequential Learning:

Stochastic Gradient Descent (SGD)

$$w^{(\tau+1)} = w^{(\tau)} - \eta \nabla E_n$$
$$= w^{(\tau)} - \eta (t_n - w^{(\tau)T} \phi_n) \phi_n$$

Regularization Least Square:

$E_w(w) = \frac{1}{2} w^T w = \frac{1}{2} \|w\|^2$ (weight decay) $w_{RLS} = (\lambda I + \Phi^T \Phi)^{-1} \Phi^T t$

q (hyperparameter)

3.2 The Bias-Variance Decomposition.

The expected squared loss can by written by:

$$E[L] = \int \{y(x) - h(x)\}^2 p(x) \cdot dx + \int \{h(x) - t\}^2 \cdot p(x,t)\, dx\, dt.$$

expected loss = $(\text{bias})^2$ + variance + noise.

average prediction: $\bar{y}(x) = \frac{1}{L} \sum_{l=1}^{L} y^{(l)}(x)$ prediction function

$(\text{bias})^2 = \frac{1}{N} \sum_{n=1}^{N} \{\bar{y}(x_n) - h(x_n)\}^2$

$\text{variance} = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{L} \sum_{l=1}^{L} \{y^{(l)}(x_n) - \bar{y}(x_n)\}^2$

3.3 Bayesan Linear Regression

Given w conjugate prior

$$P(w) = N(w \mid m_0, S_0)$$

$$P(w \mid t) \propto P(t \mid w) \cdot \underbrace{P(w)}_{\text{prior}} = N(w \mid m_N, S_N)$$

$\Rightarrow m_N = S_N (S_0^{-1} m_0 + \beta \Phi^T t) = w_{MAP}$ To simplify the treatment.

$S_N^{-1} = S_0^{-1} + \beta \Phi^T \Phi.$

we consider a zero-mean isotropic Gaussian as a prior

$$P(w \mid \alpha) = N(w \mid 0, \alpha^{-1} I)$$

$\Rightarrow m_N = \beta S_N \Phi^T t$ & $\ln P(w \mid t) = -\frac{\beta}{2} \sum_{n=1}^{N} \{t_n - w^T \phi(x_n)\}^2 - \frac{\alpha}{2} w^T w$

$S_N^{-1} = \alpha I + \beta \Phi^T \Phi.$

$\therefore P(w|\alpha) = \left[\frac{q}{2}\left(\frac{\alpha}{2}\right)^{1/q}\frac{1}{\Gamma(1/q)}\right]^M \exp\left(-\frac{\alpha}{2}\sum_{j=1}^{M}|w_j|^q\right)$.

★: **Prediction distribution**

$$p(t|\mathbf{t},\alpha,\beta)=\int p(t|w,\mathbf{t},\beta)\,p(w|\mathbf{t},\alpha,\beta)\,dw$$

↑ Gaussian $N(t|y(x,w),\beta^{-1})$

★ $p(t|x,\mathbf{t},\alpha,\beta)=N(t|m_N^T\phi(x),\ \sigma_N^2(x))$

where $\sigma_N^2(x)=\frac{1}{\beta}+\phi(x)^T S_N\,\phi(x)$.

noise data　　uncertainty due to w

Equavalent Kernel:　　new data　　former data.

★ $y(x,m_N)=m_N^T\phi(x)=\beta\phi^T(x)\cdot S_N\Phi^T(\mathbf{t})$

$=\sum_{n=1}^{N}\beta\phi^T(x)\cdot S_N\,\phi(x_n,t_n)=\sum_{n=1}^{N}k(x,x_n)\cdot t_n$

where $k(x,x')=\beta\cdot\phi^T(x)S_N\,\phi(x')$

① prediction.

★ **3.4 Bayesian Model Comparison**: ② set a criterion to select model.

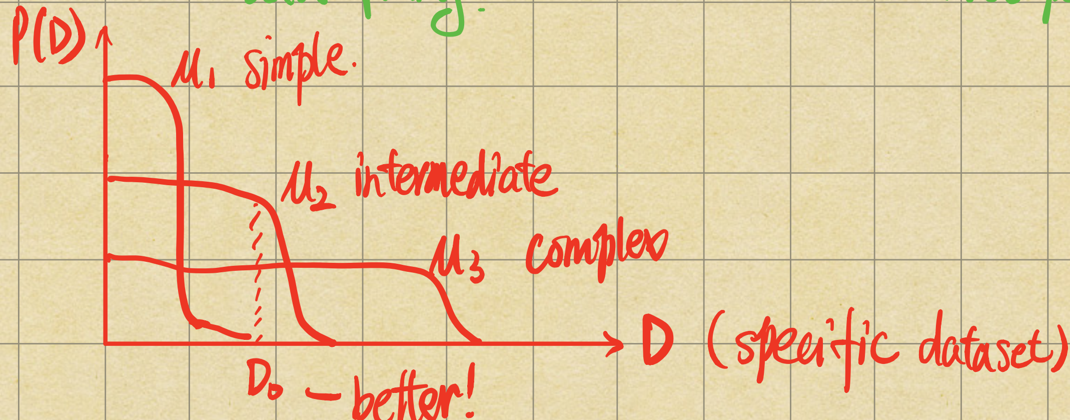
$p(M_i|D)\propto p(M_i)\cdot p(D|M_i)$

$p(D|M_i)=\int p(D|w,M_i)\,p(w|M_i)\,dw$

★ $p(D)=\int \underbrace{p(D|w)}\cdot\underbrace{p(w)}_{\frac{1}{\Delta w_{prior}}}\cdot \underbrace{dw}_{\Delta w_{posterior}} \simeq p(D|w_{MAP})\cdot\frac{\Delta w_{posterior}}{\Delta w_{prior}}$

assume sharp.

$\Rightarrow$ $\ln p(D)\simeq \underbrace{\ln p(D|w_{MAP})}_{\text{data fitting}}+\ln\left(\frac{\Delta w_{posterior}}{\Delta w_{prior}}\right)$ ← model penalty.



17.

3.5 The Evidence Approximation. (

cross validation: for deciding hyperparameter.

Predictive distribution:

$$P(t|\mathbf{t}) = \iiint P(t|w,\beta)\cdot P(w|\mathbf{t},\alpha,\beta)\cdot P(\alpha,\beta|\mathbf{t})\, dw\, d\alpha\, d\beta$$

$$\star\ P(t|\mathbf{t}) \simeq P(t|\mathbf{t},\hat{\alpha},\hat{\beta}) = \int P(t|w,\hat{\beta})\cdot P(w|\mathbf{t},\hat{\alpha},\hat{\beta})\cdot dw$$

$$(\hat{\alpha},\hat{\beta}) = \underset{(\alpha,\beta)}{\operatorname{argmax}}\, P(\alpha,\beta|\mathbf{t}) = \underset{(\alpha,\beta)}{\operatorname{argmax}}\, P(\mathbf{t}|\alpha,\beta)\cdot P(\alpha,\beta)$$

☆☆☆ Evaluation of evidence function → predictive distribution.

$$P(\mathbf{t}|\alpha,\beta) = \int P(\mathbf{t}|w,\beta)\cdot P(w|\alpha)\cdot dw$$

$$= \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\{-E(w)\}\cdot dw.$$

where $E(w) = \beta E_D(w) + \alpha E_W(w)$

$$= \frac{\beta}{2}\|\mathbf{t}-\Phi w\|^2 + \frac{\alpha}{2} w^T\cdot w.$$

☆ $\int$ is for evidence!!!

$$= \boxed{E(m_N)} + \boxed{\frac{1}{2}(w-m_N)^T A\cdot(w-m_N)}$$

$$\boxed{\underbrace{\frac{\beta}{2}\|\mathbf{t}-\Phi m_N\|^2}_{\beta E_D(w)} + \underbrace{\frac{\alpha}{2} m_N^T m_N}_{\alpha E_W(w)}} \quad \boxed{\alpha I + \beta\cdot\Phi^T\Phi}$$

$$\int \exp\{-E(w)\}dw = \exp\{-E(m_N)\}\int \exp\left\{-\frac{1}{2}(w-m_N)^T A(w-m_N)\right\}\cdot dw$$

$$= \exp\{-E(m_N)\}(2\pi)^{M/2}|A|^{-\frac{1}{2}}$$

$$\ln P(\mathbf{t}|\alpha,\beta) = \frac{N}{2}\ln\beta + \frac{M}{2}\ln\alpha - \frac{N}{2}\ln(2\pi) - E\{m_N\} - \frac{1}{2}\ln|A|$$

Maximizing the evidence function

① $\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}}\, P(\mathbf{t}|\alpha,\beta)$

Given the eigenvector equation $(\beta\Phi^T\Phi)u_i = \lambda_i u_i$

A has eigenvalues $\{\alpha+\lambda_i\}$

18.

$$\frac{d}{d\alpha}\ln|A| = \frac{d}{d\alpha}\ln\prod_i(\lambda_i+\alpha) = \frac{d}{d\alpha}\sum_i \ln(\lambda_i+\alpha) = \sum_i \frac{1}{\lambda_i+\alpha} \ \text{※}$$

$$\therefore \frac{d}{d\alpha}\ln p(t|\alpha,\beta) = \frac{M}{2\alpha} - \frac{1}{2}m_N^T\cdot m_N - \frac{1}{2}\sum_i\frac{1}{\lambda_i+\alpha} = 0$$

$$\alpha m_N^T\cdot m_N = M - \alpha\sum_{i=1}^{M}\frac{1}{\lambda_i+\alpha} = r = \sum\frac{\lambda_i}{\alpha+\lambda_i}$$

$$☆\ \alpha = \frac{r}{m_N^T\cdot m_N} \ \text{※}\ \alpha^{(0)} \to m_N^{(0)} \to r^{(0)} \to \alpha^{(1)} \to$$

② $\hat{\beta} = \text{argmax}\, p(t|\alpha,\beta)$

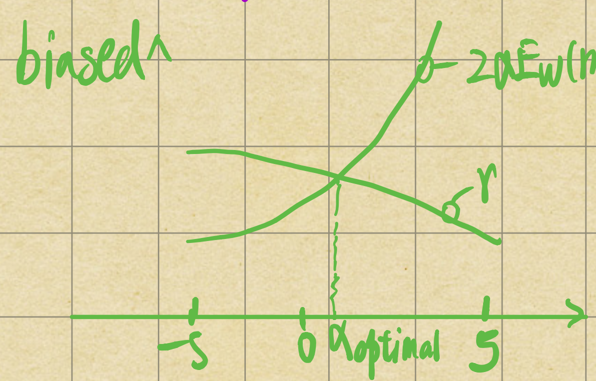
$$\frac{d}{d\beta}\ln|A| = \frac{d}{d\beta}\sum_i\ln(\lambda_i+\alpha) = \frac{1}{\beta}\sum_i\frac{\lambda_i}{\lambda_i+\alpha} = \frac{r}{\beta}$$

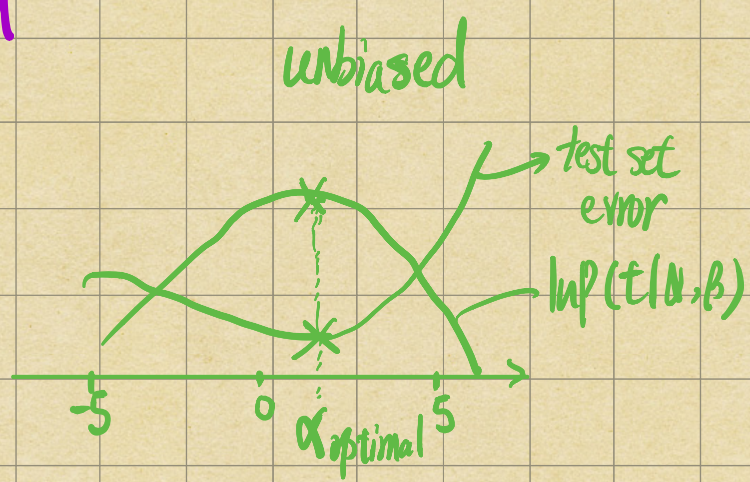
$$\frac{d}{d\beta}\ln p(t|\alpha,\beta) = \frac{N}{2\beta} - \frac{1}{2}\sum_{n=1}^{N}\{t_n - m_N^T\phi(x_n)\} - \frac{r}{2\beta} = 0$$

$$☆\ \beta^{-1} = \frac{1}{N-r}\sum_{n=1}^{N}\{t_n - m_N^T\phi(x_n)\}^2 \ \text{※}\ \beta^{(0)} \to m_N^{(0)} \to r^{(0)} \to \beta^{(1)} \to \cdots$$

effective number of parameters

$0 \le \frac{\lambda_i}{\lambda_i+\alpha} \le 1 \quad 0 \le r \le M$





↗activate function

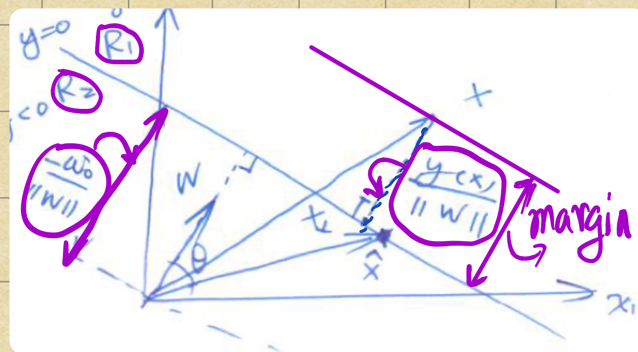
19. # Day 4: Linear Models for Classification $y(x)=f(w^Tx+w_0)$

intro: 1. Generative Model → $P(x|C_i)$

2. Discriminative Model → $P(C_i|x)$

3. Linear Discriminant Function

4.1 Discriminant Function:



two classes: $y(x)=w^Tx+w_0$ $C_1 \geq 0$ $C_2 \leq 0$

$\frac{w}{\|w\|}=x_\perp$ $x=x_\perp+r\frac{w}{\|w\|}$

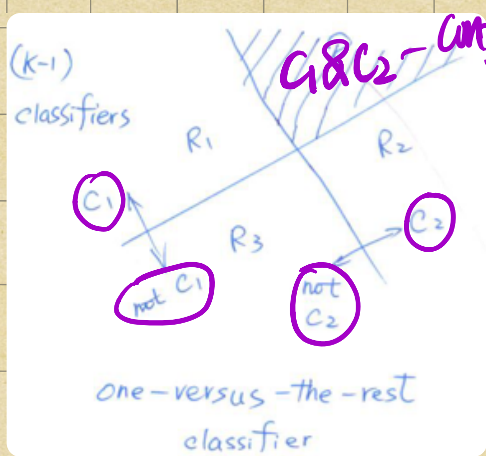
$\Rightarrow w^Tx=w^Tx_\perp+r\cdot\frac{w^T\cdot w}{\|w\|}$

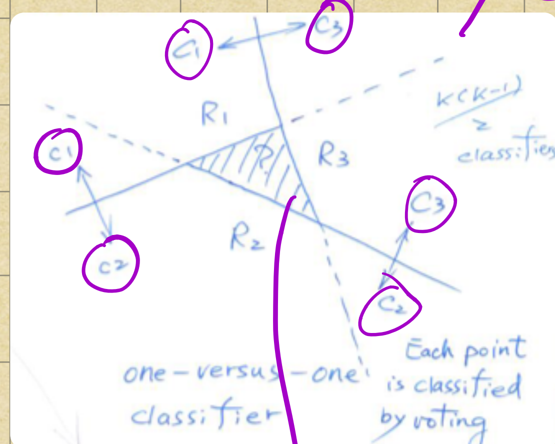
with $w^T\cdot x_\perp+w_0=0$ $\therefore w^Tx=-w_0+r\frac{w^Tw}{\|w\|}$

$\Rightarrow r=\frac{y(x)}{\|w\|}$ ※

Multiple classes $(K>2)$

voting!





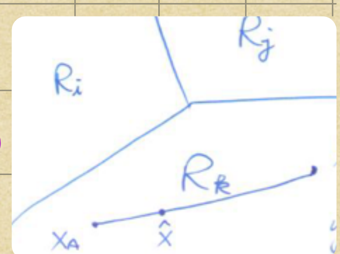
confuse

we can avoid these difficulties by:

$y_k(x)=w_k^Tx+w_{k0}$

Decision boundary between C_k & C_j is given by

$y_k(x)=y_j(x)\Rightarrow(w_k-w_j)^Tx+(w_{k0}-w_{j0})=0$



20.

① Least squares for classification

$y_k(x) = w_k^T x + w_{k0}$ $k=1,2,\dots,K$ (class) [linear discriminant function]

Error function can be written by:

$$E_D(\tilde{w}) = \frac{1}{2} Tr\{(\tilde{X}\tilde{w} - T)^T(\tilde{X}\tilde{w} - T)\}$$

Let $\nabla_{\tilde{w}} E_D(\tilde{w}) = 0 \Rightarrow \tilde{w} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T T = \tilde{X}^{+} T$

$$y(x) = \tilde{w}^T \tilde{x} = T^T(\tilde{X}^{+})^T \tilde{x}$$

② Linear Discriminant (aka: LDA ↔ PCA Linear Discriminant Analysis)

Dimensionality Reduction: → projection.

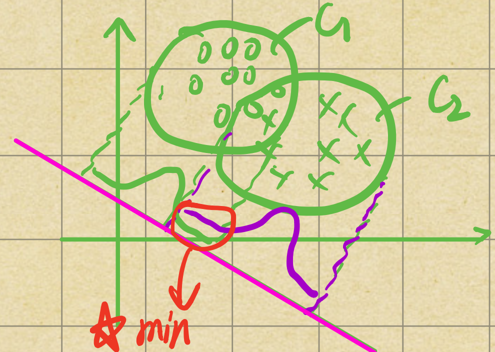
$$y = w^T x \underset{C_2}{\overset{C_1}{\gtrless}} w_0$$

Fisher's Criterion: $J(w) = \frac{(m_2 - m_1)^2}{S_1^2 + S_2^2} = \frac{w^T S_B w}{w^T S_w \cdot w}$

Max? mean var

$m_1 = \frac{1}{N_1}\sum_{n\in C_1} x_n$, $m_2 = \frac{1}{N_2}\sum_{n\in C_2} x_n$ $m_2 - m_1 = w^T(m_2 - m_1)$

where $S_B = (m_2 - m_1)(m_2 - m_1)^T$ $S_w = \sum_{n\in C_1}(x_n - m_1)(x_n - m_1)^T + \sum_{n\in C_2}(x_n - m_2)(x_n - m_2)^T$



① less variance
② more distance

Let $\nabla_w J(w) = 0 \Rightarrow (w^T S_B w) S_w \cdot w = (w^T S_w \cdot w) S_B \cdot w$.

where $S_B w = (m_2 - m_1)(m_2 - m_1)^T \cdot w$

Finally, $w \propto S_w^{-1}(m_2 - m_1)$

③ perceptron algorithm

21.

4.2 Probabilistic Generative Models

$$P(C_1|X) = \frac{P(X|C_1)\cdot P(C_1)}{P(X|C_1)\cdot P(C_1) + P(X|C_2)\cdot P(C_2)} = \frac{1}{1+\exp(-a)} = \delta(a)$$

divide $P(X|C_1)\cdot P(C_1)$

where $a = \ln \frac{P(X|C_1)\, P(C_1)}{P(X|C_2)\cdot P(C_2)}$

logistic sigmoid function

For $k>2$ $\quad P(C_k|X) = \frac{P(X|C_k)\cdot P(C_k)}{\sum_j P(X|C_j)\cdot P(C_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$

Generative Model: $P(X|C_i)$ [Indirect Model]

where $a_k = \ln P(X|C_k)\, P(C_k)$

Continuous Inputs

$$P(X|C_k) = \frac{1}{(2\pi)^{D/2}} \cdot \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(X-\mu_k)^T \cdot \Sigma^{-1}(X-\mu_k)\right\}$$

$P(C_1|X) = \delta(W^T X + W_0)$ $\{\mu_1, \mu_2, \Sigma, X = P(C_i)\}$

$W = \Sigma^{-1}(\mu_1 - \mu_2) \quad W_0 = -\frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + \ln\frac{P(C_1)}{P(C_2)}$

Maximum likelihood solution:

Gaussian class-conditional density.

$P(C_1) = \pi \quad P(C_2) = 1-\pi$ $\quad t_n = 0/1$

$P(X_n, C_1) = P(C_1)\cdot P(X_n|C_1) = \pi N(X_n|\mu_1, \Sigma)$ $\quad t_n = 1$

$P(X_n, C_2) = P(C_2)\cdot P(X_n|C_2) = (1-\pi)\cdot N(X_n|\mu_2, \Sigma)$ $\quad t_n = 0.$

likelihood function:

$$P(t|\pi, \mu_1, \mu_2, \Sigma) = \prod_{n=1}^{N} [\pi N(X_n|\mu_1, \Sigma)]^{t_n} [(1-\pi) N(X_n|\mu_2, \Sigma)]^{1-t_n}$$

After lots of compute:

1. $\pi = \frac{1}{N}\sum_{n=1}^{N} t_n = \frac{N_1}{N} = \frac{N_1}{N_1+N_2}$
2. $\mu_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n \cdot X_n$
3. $\mu_2 = \frac{1}{N_2}\sum_{n=1}^{N}(1-t_n)X_n$
4. $\Sigma = -\frac{N}{2}\ln|\Sigma| - \frac{N}{2}\operatorname{Tr}\{\Sigma^{-1}S\}$

22.

5. $S_1 = \frac{1}{N_1}\sum_{n\in C_1}(x_n-\mu_1)(x_n-\mu_1)^T \Rightarrow S = \frac{N_1}{N}S_1 + \frac{N_2}{N}S_2$

$S_2 = \frac{1}{N_2}\sum_{n\in C_2}(x_n-\mu_2)(x_n-\mu_2)^T \quad \Sigma = S$ ※

4.3 Probabilistic Discriminative Models (Direct Model) $P(C_k|X)$

必考!

Intro: $\begin{cases} \text{Indirect: } P(X|C_k) \to ML \to P(C_k|X) \to P(C_k) \\ \text{Direct: } P(C_k|X) \to \end{cases}$

$\begin{cases} P(C_1|\phi) = y(\phi) = \sigma(w^T\phi) \\ P(C_2|\phi) = 1 - P(C_1|\phi) \end{cases}$ $\frac{d\sigma}{da} = \sigma(1-\sigma)$

$D = \{\phi_n, t_n\}_{n=1}^N,\ t_n = \{0,1\}$

$P\{t|w\} = \prod_{n=1}^N y_n^{t_n}\{1-y_n\}^{1-t_n}$ & Cross-Entropy

$E(w) = -\ln P(t|w)$

$= -\sum_{n=1}^N\{t_n\ln y_n + (1-t_n)\ln(1-y_n)\}$ ※

Iterative Reweighted Least Squares (IRLS)

1. $w^{(new)} = w^{(old)} - \eta\nabla E(w)$ steepest descent alg.

2. Newton-Raphson alg

$w^{(new)} = w^{(old)} - H^{-1}\nabla E(w)$, where $H = \nabla\nabla E(w)$

$\nabla E(w) = \sum_{n=1}^N (y_n - t_n)\phi_n = \Phi^T(y-t)$ data matrix

$H = \nabla\nabla E(w) = \sum_{n=1}^N y_n(1-y_n)\phi_n\phi_n^T = \Phi^T R\Phi$, weighting matrix $R_{N\times N} = [R_{nn}]$

$\Rightarrow w^{(new)} = w^{(old)} - (\Phi^T R\Phi)^{-1}\Phi^T(y-t)$

$= (\Phi^T R\Phi)^{-1}\{\Phi^T R\Phi w^{(old)} - \Phi^T(y-t)\}$

$= (\Phi^T R\Phi)^{-1}\Phi^T R z$, where $z = \Phi w^{(old)} - R^{-1}(y-t)$

Core idea: $w \to R \to w \to R$

4.4 Laplace Approximation

classification $\to \int\sigma \to$ 不可积, $\Rightarrow$ approximation. (Taylor & ln)

$P(z) = \frac{f(z)}{\int f(z)dz}$ via Taylor series to quadratic equation.

$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2}(z-z_0)^T A(z-z_0)$, where $A = -\nabla\nabla\ln f(z)|_{z=z_0}$

$\Rightarrow f(z) \approx f(z_0)\exp\{-\frac{1}{2}(z-z_0)^T A(z-z_0)\}$

23.

$\therefore \int f(z)\,dz \simeq f(z_0)\frac{(2\pi)^{M/2}}{|A|^{1/2}}$ *

$P(z) \longleftarrow q(z)$

non-Gaussian. Gaussian

$q(z)=\frac{|A|^{1/2}}{(2\pi)^{M/2}}\exp\left\{-\frac{1}{2}(z-z_0)^T A(z-z_0)\right\}$ *

Model comparison and BIC:

We have mode evidence: $P(D)=\int P(D|\theta)\cdot P(\theta)\,d\theta$

$\frac{P(D|\theta)\cdot P(\theta)}{P(D)}=P(\theta|D)=\frac{f(\theta)}{[Z=\int f(\theta)d\theta]}$

$\ln P(D)\simeq \ln P(D|\theta_{MAP})+\ln P(\theta_{MAP})+\frac{M}{2}\ln(2\pi)-\frac{1}{2}\ln|A|$

Day5: kernal function: Memory based Methods

$K(x,x')=\phi(x)^T\cdot\phi(x')$ * similarity between x & x'

Gaussian Process (可能考)

6.1: Dual Representation

$J(w)=\frac{1}{2}\sum_{n=1}^{N}\{w^T\phi(x_n)-t_n\}^2+\frac{\lambda}{2}w^Tw$

$\nabla_w J(w)=0 \rightarrow w_{opt}=-\frac{1}{\lambda}\sum_{n=1}^{N}\{w^T\phi(x_n)-t_n\}\phi(x_n)=\sum_{n=1}^{N}a_n\phi(x_n)=\Phi^T a$

$a_n=-\frac{1}{\lambda}\{w^T\phi(x_n)-t_n\}$, where $\Phi^T=[\phi(x_1),\cdots,\phi(x_n)]$, $a=[a_1,\cdots,a_n]^T$

$\Rightarrow J(a)=\frac{1}{2}a^T\Phi\Phi^T\Phi\Phi^T a-a^T\Phi\Phi^T t+\frac{1}{2}t^Tt+\frac{\lambda}{2}a^T\Phi\Phi^T\cdot a$

let $K=\Phi\cdot\Phi^T$, we obtain $K_{nm}=\phi(x_n)^T\phi(x_m)=K(x_n,x_m)$

$J(a)=\frac{1}{2}a^TKKa-a^TKt+\frac{1}{2}t^Tt+\frac{\lambda}{2}a^TKa$

$\nabla_a J(a)=0 \quad a_{opt}=(K+\lambda I_N)^{-1}t$

$y(x)=w^T\phi(x)=a^T\Phi\phi(x)=K(x)^T(K+\lambda I_N)^{-1}t$

6.4: Gaussian Process

Linear regression revisited:

$y(x) = w^T \phi(x)$ & $p(w) = \mathcal{N}(w \mid 0, \alpha^{-1} I)$

Given $\{x_1, \cdots, x_n\}$. we have $y = \{y(x_1), \cdots, y(x_n)\}^T$

$\Rightarrow y = \Phi w \therefore E[y] = \Phi E(w) = 0$

$cov[y] = E[yy^T] = \Phi E[w \cdot w^T] \Phi^T = \frac{1}{\alpha} \Phi \Phi^T = K.$

Gaussian Process for regression:

可能考!

$t_n = y_n + \underbrace{\epsilon_n}_{\text{noise}}$

$P(t_n \mid y_n) = \mathcal{N}(t_n \mid y_n, \beta^{-1})$

$P(t \mid y) = \mathcal{N}(t \mid y, \beta^{-1} I_N)$ or $\mathcal{N}(t - y \mid 0, \beta^{-1} I_N)$

$P(y) = \mathcal{N}(y \mid 0, k)$

$P(t) = \int p(t \mid y) \cdot P(y) dy = \mathcal{N}(\underset{(N \times 1)}{t} \mid 0, C_{N \times N})$

where $C(x_n, x_m) = k(x_n, x_m) + \beta^{-1} \delta_{nm}$

One widely used kernel function:

$k(x_n, x_m) = \theta_0 \exp\{-\frac{\theta_1}{2} \|x_n - x_m\|^2\} + \theta_2 + \theta_3 x_n^T x_m$

$(\theta_1, \theta_2, \theta_3, \theta_4)$ & $P(t)$ hyperparameter.

$P(t_{N+1}) = P(t_1, \cdots, t_{N+1}) = \mathcal{N}(t_{N+1} \mid 0, C_{N+1})$, where $C_{N+1} = \begin{pmatrix} C_N & k \\ k^T & c \end{pmatrix}$

$c = k(x_{N+1}, x_{n+1}) + \beta^{-1}$

$P(t_{n+1} \mid t) = \mathcal{N}(t_{n+1} \mid m(x_{n+1}), \delta^2(x_{n+1})) = K^T C_N^{-1} t = c - K^T C_N^{-1} K$

Learning parameters:

$\hat{\theta} = \underset{\theta}{\text{argmax}} \log P(t \mid \theta) \rightarrow \ln P(t \mid \theta) = -\frac{1}{2} \ln |C_N| - \frac{1}{2} t^T C_N^{-1} t - \frac{N}{2} \ln(2\pi)$

$\frac{\partial}{\partial \theta_i} \ln P(t \mid \theta) = -\frac{1}{2} Tr(C_N^{-1} \frac{\partial C_N}{\partial \theta_i}) + \frac{1}{2} t^T C_N^{-1} \frac{\partial C_N}{\partial \theta_i} \cdot C_N^{-1} \cdot t$

25. 不考 太复杂 6.5 Gaussian processes for classification

$P(t|a)=\sigma(a)^{t_N}(1-\sigma(a))^{t_N}$

Laplace approximation:

$P(a_{N+1}|t_N)=\int P(a_{N+1},a_N|t_N)\,da_N$ $\quad P(a_{N+1}|a_N)P(a_N)\cdot P(t_N|a_N)$

$=\int P(a_{N+1}|a_N)\,P(a_N|t_N)\,da_N$

......

$q(a_N)=N(a_N|a_N^{*},H^{-1})$

$\rightarrow P(a_{N+1}|a_N)\cong\int P(a_{N+1}|a_N)\,q(a_N)\,da_N$

then $P(t_{N+1}=1|t_N)=\int P(t_{N+1}=1|a_{N+1})P(a_{N+1}|t_N)\,da_{N+1}$

Day 6. 7: Sparse Kernal Machines (linear model)

Intro: 1. Support Vector Machines (SVM) $\Rightarrow$ (select important vector to predict)

2. Relevance Vector Machines (RVM)

Support Vector Machines.

Classification:

(一): Nonoverlapping classes

(二): Overlapping Classes

Regression: (support vector regression)

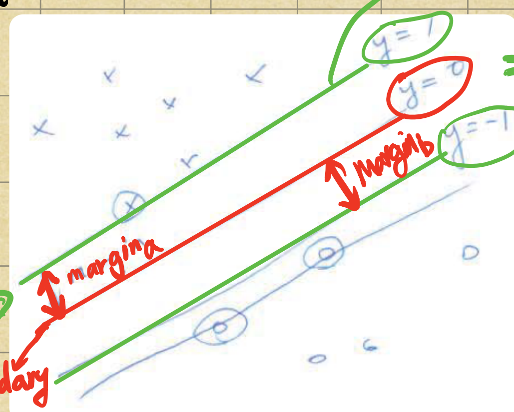
Two-class classification problem

$y=w^T\phi(x)+b$

found a bound maximum the "Margin" among dataset.

Margin: $margin_a=margin_b$

Support Vector

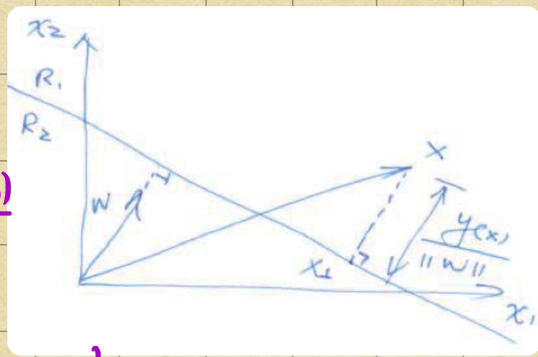


$\Rightarrow$ Nonoverlapping

Geometric Interpretation.

26.

Distance: $r = \frac{y(x)}{\|w\|}$



add class label we obtain:

$\Rightarrow$ margin: $\frac{t_n y(x_n)}{\|w\|} = \frac{t_n(w^T\phi(x_n)+b)}{\|w\|}$

☆ **Maximum margin solution is:**

$(\hat{w},\hat{b}) = \underset{(w,b)}{\arg\max}\left\{\frac{1}{\|w\|}\min_n[t_n(w^T\phi(x_n)+b)]\right\}$

with $t_n(w^T\phi(x_n)+b) \geq 1$ (as constrain) we obtain: $(\hat{w},\hat{b}) = \underset{(w,b)}{\arg\min}\frac{1}{2}\|w\|^2$ ※

$\Rightarrow L(w,b,a) = \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} a_n\{t_n(w^T\phi(x_n)+b)-1\}$

Lagrarge multiplier

let $\frac{\partial L}{\partial w}=0$, $\frac{\partial L}{\partial b}=0$

$\therefore w = \sum_{n=1}^{N} a_n t_n \phi(x_n)$ $\quad 0 = \sum_{n=1}^{N} a_n t_n$

Dual representation of the maximum margin problem.

$\rightarrow \max_a\left\{\tilde{L}(a) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m K(x_n, x_m)\right\}$, where $\begin{cases} \sum_{n=1}^{N} a_n t_n = 0 \\ a_n \geq 0 \text{ ※} \\ t_n y(x_n - 1) \geq 0 \end{cases}$

Solution: Sequential minimal optimization. (SMO)

In classification problem:

$y(x) = \sum_{n=1}^{N} a_n t_n k(x, x_n) + b$

test data — training data

the KKT conditions: $a_n \geq 0$, $t_n y(x_n) - 1 = 0$ $\quad a_n\{t_n y(x_n) - 1\} = 0$

$\therefore a_n = 0$ or $t_n y(x_n) = 1$

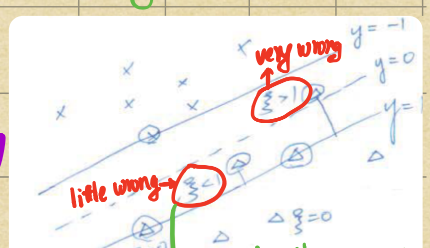
the parameter b:

$t_n\left(\sum_{m \in S} a_m t_m k(x_n, x_m) + b\right) = 1$

$\Rightarrow b = \frac{1}{N_S}\sum_{n \in S}\left(t_n - \sum_{m \in S} a_m t_m k(x_n, x_m)\right)$

error function: $\sum_{n=1}^{N} E_\infty(y(x_n)t_n - 1) + \lambda\|w\|^2$ → quadratic regularizer.

Overlapping class distribution:



Slack varibles are introduced: $\xi_n \geq 0$, $n=1,\cdots,N$

↳ to measure to misclassified point.

inside the margin but is correct.

classification constraints are replaced by: $t_n y(x_n) \geq 1 - \xi_n$

27.

Therefore, $\min_{w}\left(C\sum_{n=1}^{N}\xi_n+\frac{1}{2}\|w\|^2\right)$

$C \downarrow$ hyperparameter for trading-off

KKT condition is given by:

$a_n\geq 0,\ t_n y(x_n)-1+\xi_n\geq 0\quad a_n(t_n y(x_n)-1+\xi_n)=0$

$\mu_n\geq 0,\ \xi_n\geq 0,\ \mu_n\xi_n=0$

Lagrangian is written by:

$L(w,b,a)=\frac{1}{2}\|w\|^2+C\sum_{n=1}^{N}\xi_n-\sum_{n=1}^{N}a_n\{t_n y(x_n)-1+\xi_n\}-\sum_{n=1}^{N}\mu_n\xi_n$

where. $\mu_n\geq 0,\ a_n\geq 0$

then $\partial L/\partial w=0$ $\partial L/\partial b=0$, $\partial L/\partial \xi=0$, we obtain $\begin{cases} w=\sum_{n=1}^{N}a_n t_n\phi(x_n) \\ \sum_{n=1}^{N}a_n t_n=0 \end{cases}$ $a_n=C-\mu_n$

and dual Lagrange is: $\min_{\{a_n\}}\left\{\tilde{L}(a)=\sum_{n=1}^{N}a_n-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}a_n a_m t_n t_m k(x_n,x_m)\right\}$

subject to $\begin{cases} 0\leq a_n\leq C \\ \sum_{n=1}^{N}a_n t_n=0 \end{cases}$ → control the quantity of support vectors.

Solution Interpretation:

① $a_n=0\Rightarrow$ nonsupport vector

② $0<a_n<C$, then $\mu_n>0$, then $\xi_n=0$, this point lies on the margin.

③ if $a_n=C$, then $\mu_n=0$, $\xi_n\leq 1$. or $\xi_n>1$
(classified) (misclassified)

To determin b, support vectors a_n satisfy $0<a_n<C$, $\xi=0$, $t_n y(x_n)=1$.

then we have $t_n\left(\sum_{m\in S}a_m t_m k(x_n,x_m)+b\right)=1$

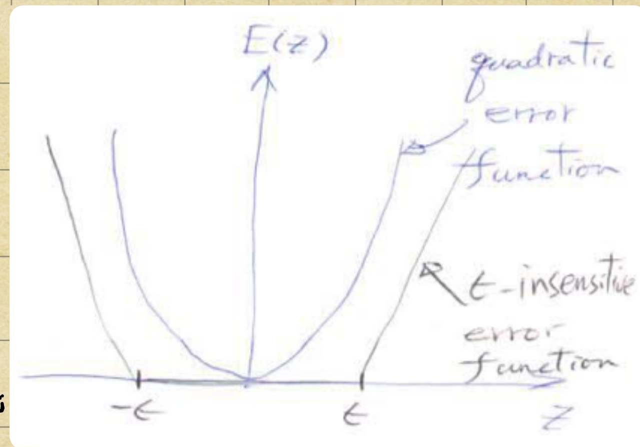
$\Rightarrow b=\frac{1}{N_M}\sum_{n\in M}\left(t_n-\sum_{m\in S}a_m t_m k(x_n,x_m)\right)$

↳ M is a set with data points having $0<a_n<C$.

SVM for regression: (Support Vector Regression)

We define simple error function:

$$\Rightarrow \frac{1}{2}\sum_{n=1}^{N}\{y_n-t_n\}^2+\frac{\lambda}{2}\|w\|^2$$



To obtain sparse solution as:

$$E_\epsilon(y(x)-t)=\begin{cases}0 & \text{if } |y(x)-t|<\epsilon \\ |y(x)-t|-\epsilon & \end{cases}$$

A new regularized error function:

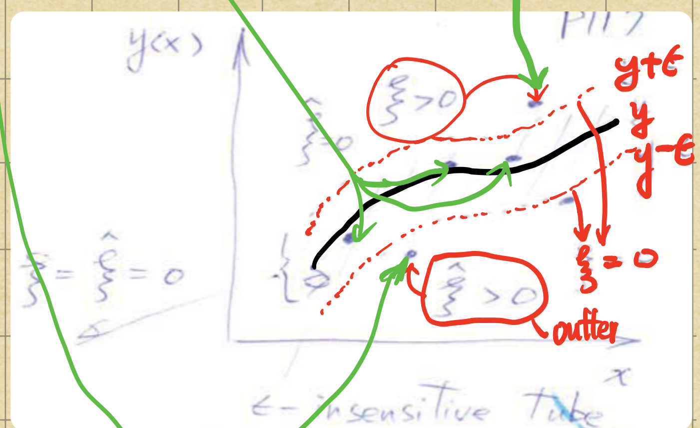
$$C\sum_{n=1}^{N}E_\epsilon(y(x_n)-t_n)+\frac{1}{2}\|w\|^2$$

By introduce two slack varibles:

$$\xi_n\geq 0 \equiv t_n>y(x_n)+\epsilon$$

$$\hat{\xi}_n\geq 0 \equiv t_n<y(x_n)-\epsilon$$

For $y_n-\epsilon\leq t_n\leq y_n+\epsilon \Rightarrow \xi_n=\hat{\xi}_n=0$



Error function of SVR:

$$C\sum_{n=1}^{N}(\xi_n+\hat{\xi}_n)+\frac{1}{2}\|w\|^2$$

Constrains:

$\xi_n\geq 0$ & $\hat{\xi}\geq 0$ &

$t_n\leq y(x_n)+\epsilon+\xi_n$ &

$t_n\geq y(x_n)-\epsilon-\hat{\xi}_n$

Lagrange optimization:

$$L=C\sum_{n=1}^{N}(\xi_n+\hat{\xi}_n)+\frac{1}{2}\|w\|^2-\sum_{n=1}^{N}(\mu_n\xi_n+\hat{\mu}_n\hat{\xi}_n)$$
$$-\sum_{n=1}^{N}a_n(\epsilon+\xi_n+y_n-t_n)-\sum_{n=1}^{N}\hat{a}_n(\epsilon+\hat{\xi}_n-y_n+t_n)$$

$$\Rightarrow\begin{cases}\partial L/\partial w=0\Rightarrow w=\sum_{n=1}^{N}(a_n-\hat{a}_n)\phi(x_n)\\ \partial L/\partial b=0\Rightarrow \sum_{n=1}^{N}(a_n-\hat{a}_n)=0\\ \partial L/\partial\hat{\xi}_n=0\Rightarrow \hat{a}_n+\hat{\mu}_n=0\\ \partial L/\partial\xi_n=0\Rightarrow a_n+\mu_n=0\end{cases}$$

29.

Dual presentation:

$$\tilde{L}(a,\hat{a}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n-\hat{a}_n)(a_m-\hat{a}_m)k(x_n,x_m) - \epsilon\sum_{n=1}^{N}(a_n+\hat{a}_n)+\sum_{n=1}^{N}(a_n-\hat{a}_n)t_n$$

From the result.

$$y(x)=\sum_{n=1}^{N}(a_n-\hat{a}_n)k(x,x_n)+b$$

The KKT conditions are given by:

$$a_n(\epsilon+\xi_n+y_n-t_n)=0, \quad \hat{a}_n(\epsilon+\hat{\xi}_n-y_n+t_n)=0$$

$$(C-a_n)\xi_n=0 \quad (C-\hat{a}_n)\hat{\xi}_n=0$$

The parameter "b" can be found by:

$$b=t_n-\epsilon-w^T\phi(x_n)$$
$$=t_n-\epsilon-\sum_{m=1}^{N}(a_m-\hat{a}_m)k(x_n,x_m)$$

7.2 Relence Vector Machine (RVM) [7.1 is all about SVM above & RVM is Sparse Kenel Machine)

Core idea:

① SVM is without probability, how about add Probability to it ?

② SVM is two-class classification, how about more ?

③ C or ν should be found from held-out data.

The SVM-like form in RVM is: ③ the number of w is much larger than others

② $y(x)=\sum_{n=1}^{N}w_n k(x_i,x_n)+b$, then likelihood function is

$P(t|X,w,\beta)=\prod_{n=1}^{N}P(t_n|x_n,w,\beta^{-1})$. And the prior distrub

is ① $P(w|\alpha)=\prod_{i=1}^{N}N(w_i|0,\alpha_i^{-1})$ hyperparameter.

we could obtain posterior distrib: $P(w|t,X,\alpha,\beta)=N(w|m,\Sigma)$

where $m=\beta\Sigma\Phi^T t$ (w_{MAP}) $\Sigma=(A+\beta\Phi^T\Phi)^{-1}$

1. Estimate α, β (training)

α, β are determined by "evidence approximation":

$P(t|X,\alpha,\beta) = \int p(t|X,w,\beta)\, p(w|\alpha)\, dw$, then

$\ln P(t|X,\alpha,\beta) = \ln N(t|0,C) = -\frac{1}{2}\{N\ln(2\pi) + \ln|C| + t^T C^{-1} t\}$.

where $C = \beta^{-1} I + \Phi A^{-1} \Phi^T$. Let $\nabla \ln p(t|X,\alpha,\beta) = 0$

After amount of calculations we can obtain:

$\alpha_i^{new} = r_i / m_i^2$, $(\beta^{new})^{-1} = \frac{\|t - \Phi m_N\|^2}{N - \sum_i r_i}$

$P(t|x,X,\alpha^*,\beta^*) = \int p(t|x,w,\beta^*) \cdot P(w|X,t,\alpha^*,\beta^*)\, dw$

$= N(t|m^T\phi(x), \sigma^2(x))$

where $\sigma^2(x) = (\beta^*)^{-1} + \phi(x)^T \Sigma \phi(x)$.

2. Predictive Distrib (test)

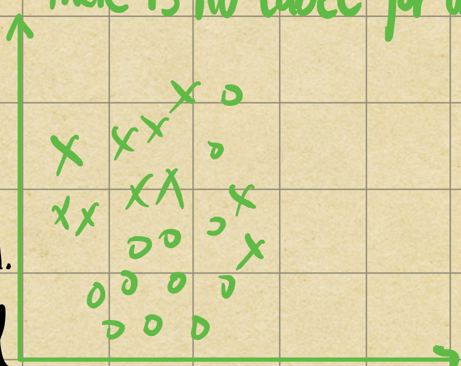
Day 7: Unsupervised Learning

$\downarrow$

clustering problem

{ with probability: mixture of Gaussian.
  without probability: geometry method

there is no label for answers.



Chapter 9: Mixture Models and EM

E: expectation M: Maximization.

9.1 K-means Clustering (non-probability & mult-varibles)

太简单，不考.

Define $J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2$ ※ $r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|x_n - \mu_j\|^2 \\ 0 & \text{otherwise} \end{cases}$

↑ membership indicator

hard

soft

(the number of data is N & class is k)

then $\partial J / \partial \mu_k = 2\sum_{n=1}^{N} r_{nk}(x_n-\mu_k)=0$

9.2 Mixture of Gaussians: (with probability)

要考就考它!

Given $p(x)=\sum_{k=1}^{K}\pi_k N(x|\mu_k,\Sigma_k)$

mixture weight.

z is a latent varible represent one of the K latent states

$\therefore P(z_k=1)=\pi_k$. then $P(x)=\sum_{z} P(z)\cdot P(x|z)=\sum_{k=1}^{K}\pi_k N(x|\mu_k,\Sigma_k)$

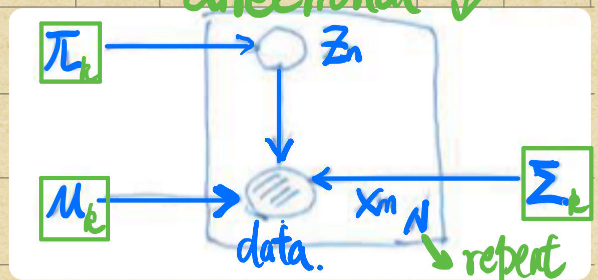
9.2.1 Maximum likelihood

Let $X=\{x_1,\cdots,x_n\}$ & $Z=\{z_1,\cdots,z_N\}$

likelihood function is:

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

directional



repeat N times.

Graphical Model

9.2.2: EM for Gaussian mixtures.

$P(z_k=1)$ $P(x_n|z_k=1)$

$$\frac{\partial}{\partial\mu_k}\ln P(X|\pi,\mu,\Sigma)=-\sum_{n=1}^{N}\frac{\pi_k N(x_n|\mu_k,\Sigma_k)}{\sum_j \pi_j N(x_n|\mu_j,\Sigma_j)}\Sigma_k^{-1}(x_n-\mu_k)=0$$

$P(x_n)$

$\Rightarrow \mu_k=\frac{1}{N_k}\sum_{n=1}^{N} r(z_{nk})x_n$, where $N_k=\sum_{n=1}^{N} r(z_{nk})$

Let $\frac{\partial}{\partial\Sigma_k}\ln P(X|\pi,\mu,\Sigma)=0 \Rightarrow \Sigma_k=\frac{1}{N_k}\sum_{n=1}^{N} r(z_{nk})(x_n-\mu_k)\cdot(x_n-\mu_k)^T$

$$\frac{\partial}{\partial\pi_k}\left(\ln P(X|\pi,\mu,\Sigma)+\lambda\left(\sum_{k=1}^{K}\pi_k-1\right)\right)=\sum_{n=1}^{N}\frac{N(x_n|\mu_k,\Sigma_k)}{\sum_j \pi_j N(x_n|\mu_j,\Sigma_j)}+\lambda$$

Lagrange multiplier

$\lambda=-N \longrightarrow \pi_k=N_k/N$

9.2.3: EM for Gaussian mixtures: (Algorithm)

① Initialize μ_k,Σ_k, evaluate log likelihood.

② E-step: $r(z_{nk})=\frac{\pi_k N(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N(x_n|\mu_j,\Sigma_j)}$

③ M-step: $\mu_k^{new}=\frac{1}{N_k}\sum_{n=1}^{N} r(z_{nk})\cdot x_n$

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

9.3 An alternative View of EM:

$\ln P(X|\theta) = \ln\left\{\sum_Z P(X, Z|\theta)\right\}$

incomplete. and $\{X, Z\}$ is complete data.

General EM Algorithm:

E-step: $Q(\theta, \theta^{old}) = E_Z[\ln P(X, Z|\theta) \mid X, \theta^{old})$

M-step: $Q^{new} = \underset{\theta}{\arg\max}\, Q(\theta, \theta^{old})$

Gaussian mixture revisited:

$\ln P(X, Z|\mu, \Sigma, \pi) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\{\ln \pi_k + \ln N(x_n|\mu_k, \Sigma_k)\}$

$\because P(X, Z) = P(Z)P(X|Z) = \prod_k \pi_k^{z_k} \prod_k N(X|\mu_k, \Sigma_k)^{z_k}$

$E_Z[\ln P(X, Z|\mu, \Sigma, \pi)] = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})\{\ln \pi_k + \ln N(x_n|\mu_k, \Sigma_k)\}$

$$\because E[z_{nk}] = \frac{\sum_{z_{nk}} z_{nk}[\pi_k N(x_n|\mu_k, \Sigma_k)]^{z_{nk}}}{\sum_{z_{nj}}[\pi_j N(x_n|\mu_j \Sigma_j)]^{z_{nj}}} = \gamma(z_{nk})$$

9.4: The EM Algorithm in general

$P(X|\theta) = \sum_Z P(X, Z|\theta)$,

$\ln P(X|\theta) = \ln \sum_Z \frac{P(X, Z|\theta)}{q(Z)} \cdot q(Z) = \ln E_{q(Z)}\left[\frac{P(X, Z|\theta)}{q(Z)}\right] \geq E_{q(Z)}\left[\ln \frac{P(X, Z|\theta)}{q(Z)}\right]$

$= \mathcal{L}(q, \theta) = \sum_Z q(Z) \ln\left\{\frac{P(X, Z)}{q(Z)}\right\}$.

$\ln P(X|\theta) = \mathcal{L}(q, \theta) + KL(q\|p)$

where $KL(q\|p) = -\sum_Z q(Z) \ln\left\{\frac{P(Z|X, \theta)}{q(Z)}\right\}$

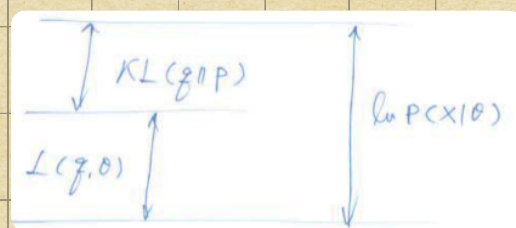


Illustration for EM algorithm:

① θ^{old}: $\hat{q} = \underset{q}{\arg\min}\, KL(q\|p) = 0$, $\hat{q}(Z) = P(Z|X, \theta^{old})$

② $\hat{\theta}^{new}$: $\underset{\theta}{\arg\max}\, \mathcal{L}(\hat{q}, \theta)$.

$\mathcal{L}(q, \theta) = \sum_Z q(Z) \ln\left\{\frac{P(X, Z|\theta)}{q(Z)}\right\} = \sum_Z P(Z|X, \theta^{old}) \ln P(X, Z|\theta) -$

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

$= Q(\theta, \theta^{old}) + const$

1.4 Information Theroy : we define $h(x,y)=h(x)+h(y)$ & $h(x)=-\log_2 P(x)$

$H[X]=-\sum_x P(x)\cdot\log_2 P(x)$, $H[p]=-\sum_i P(x_i)\ln P(x_i)=E[-\ln P(x)]$

Maximum Entropy (ME) : $\tilde{H}=-\sum_i P(x_i)\ln P(x_i)+\lambda(\sum_i P(x_i)-1)$: Given x is continuous random varibles $\Rightarrow H[X]=-\int P(x)\ln P(x)\cdot dx$ (differential entropy)