## 1.1 basic concept (待, 可以用来充字数)

Polynomial func: $y(x,w)=w_0+w_1x+\cdots+w_Mx^M=\sum_{j=0}^{M}w_jx^j$

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

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

Bayesian: $P(w|D)=\frac{P(D|w)\cdot P(w)}{P(D)}\longrightarrow \int P(D|w)P(w)dw$ & $\xrightarrow{\text{classify}} P(C_k|x)=\frac{P(x|C_k)\cdot P(C_k)}{P(x)}$

Gaussian: $N(x|\mu,\delta^2)=\frac{1}{(2\pi\delta^2)^{1/2}}\exp\{-\frac{1}{2\delta^2}(x-\mu)^2\}$

D-Gaussian: $N(x|\mu,\delta^2)=\frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$

## 1.2 Maximum Likelihood

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

$\therefore P(D|w)=P(t|x,w,\beta)=\prod_{n=1}^{N}N(t_n|x,w,\beta)$ $\therefore \ln P(t|x,w,\beta)=$
$-\frac{\beta}{2}\sum_{n=1}^{N}(y(x_n,w)-t_n)^2+\frac{N}{2}\ln\beta-\frac{N}{2}\ln(2\pi)$, Let $\beta^{-1}=\frac{1}{N}\sum_{n=1}^{N}\{y(x_n,w_{ML})-t_n\}^2$; $w_{MAP}=\arg\max_w P(w|x,t,\alpha,\beta)=\arg\min_w\{\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,w)-t_n\}^2+\frac{\alpha}{2}w^Tw\}$

## 2.1 basic concepts of Probability Distribution.

### 2.1.1 Bernoulli

$Bin(m|N,\mu)=\binom{N}{m}\mu^m(1-\mu)^{N-m}$

Beta Distrib: $Beta(\mu|a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$: $E(\mu)=\frac{a}{a+b}$ $Var[\mu]=\frac{ab}{(a+b)^2(a+b+1)}$

$\underbrace{P(\mu|m,l,a,b)}_{\text{posterior}}\propto Bin(m|N,\mu)\underbrace{Beta(\mu|a,b)}_{\text{prior}}$

### 2.2.1 Conditional Gaussian

$\Delta^2=(x-\mu)^T\Sigma^{-1}(x-\mu)$ $\Lambda=\Sigma^{-1}$

$\Rightarrow -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)=-\frac{1}{2}\sum_{a,b}^{\oplus}(x_a-\mu_a)^T\Lambda_{aa}(x_a-\mu_a)$

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

### 2.2.2 Marginal Gaussian Distribution

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

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

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

## 2.3 Bayesian inference for Gaussian

① unknown $\mu$: $P(x|\mu)=\prod_{n=1}^{N}P(x_n|\mu)=\frac{1}{(2\pi\delta^2)^{N/2}}\exp\{-\frac{\sum_{n=1}^{N}(x_n-\mu)^2}{2\delta^2}\}$

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

Posterior distribution: $P(\mu|x)\propto P(x|\mu)\cdot P(\mu)$

$\mu_N=\frac{\delta^2}{N\delta_0^2+\delta^2}\mu_0+\frac{N\delta_0^2}{N\delta_0^2+\delta^2}\mu_{ML}$ $\frac{1}{\delta_N^2}=\frac{1}{\delta_0^2}+\frac{N}{\delta^2}$

② unknown variance: $P(x|\lambda)=\prod_{n=1}^{N}N(x_n|\mu,\lambda^{-1})\propto\lambda^{N/2}\exp\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\}$

Conjugate prior: $Gam(\lambda|a,b)=\frac{1}{\Gamma(a)}b^a\lambda^{a-1}\exp(-b\lambda)$

Posterior distribution: $P(\lambda|x)\propto\lambda^{a_0-1}\lambda^{N/2}\exp\{-b_0\lambda-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\}$

$\Rightarrow Gam(\lambda|a_N,b_N)$ $a_N=a_0+\frac{N}{2}$ $b_N=b_0+\frac{N}{2}\delta_{ML}^2$

③ both unknown: $P(x|\mu,\lambda)=\prod_{n=1}^{N}(\frac{\lambda}{2\pi})^{1/2}\exp\{-\frac{\lambda}{2}(x_n-\mu)^2\}$

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

Conjugate Prior: $P(\mu,\lambda)\propto[\lambda^{1/2}\exp(-\frac{\lambda\mu^2}{2})]^{\beta}\exp[c\lambda\mu-d\lambda]$

$=\exp\{-\frac{\beta\lambda}{2}(\mu-c/\beta)^2\}\lambda^{\beta/2}\exp\{-(d-\frac{c^2}{2\beta})\lambda\}$

$=P(\mu|\lambda)P(\lambda)$

Normal-gamma distribution: $\mu_0=c/\beta$, $a=1+\beta/2$, $b=d-c^2/2\beta$

## Mixture of Gaussians

$P(x)=\sum_{k=1}^{K}\pi_kN(x|\mu_k,\Sigma_k)=\sum_{k=1}^{K}P(k)P(x|k)$

$\Rightarrow ML$: $\ln P(x|\pi,\mu,\Sigma)=\sum_{n=1}^{N}\ln\{\sum_{k=1}^{K}\pi_kN(x_n|\mu_k,\Sigma_k)\}$

待考但很有用

## The expential family

$P(x|\eta)=h(x)g(\eta)\exp\{\eta^Tu(x)\}$

$\Rightarrow$ ① Bernoulli: $P(x|\mu)=\mu^x(1-\mu)^{1-x}=\exp\{x\ln\mu+(1-x)\ln(1-\mu)\}$
$=(1-\mu)\exp\{\ln\frac{\mu}{1-\mu}\cdot x\}=\delta(-\eta)\cdot\exp(\eta x)$

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

② Gaussian Distrib: $P(x|\mu,\delta^2)=\frac{1}{(2\pi\delta^2)^{1/2}}\exp\{-\frac{1}{2\delta^2}(x-\mu)^2\}$
$=\frac{1}{(2\pi\delta^2)^{1/2}}\exp\{-\frac{1}{2\delta^2}x^2+\frac{\mu}{\delta^2}x-\frac{1}{2\delta^2}\mu^2\}$
$=h(x)g(\eta)\exp\{\eta^Tu(x)\}\Rightarrow h(x)=(2\pi)^{-\frac{1}{2}}$ $g(\eta)=(-2\eta_2)^{\frac{1}{2}}\exp(\frac{\eta_1^2}{4\eta_2})$

$\eta=\begin{pmatrix}\mu/\delta^2\\ -1/2\delta^2\end{pmatrix}_{\eta_2}^{\eta_1}$ $u(x)=\begin{pmatrix}x\\ x^2\end{pmatrix}$

$\Rightarrow$ Conjugate Prior:

$P(\eta|X,\nu)=f(\chi,\nu)\cdot g(\eta)^{\nu}\exp\{\nu\cdot\eta^T\chi\}$

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

## 3.1 Linear Basis Function Models

### 3.1.1 Maximum likelihood & least squares

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

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

$\rightarrow E(t|x)=\int tp(t|x)dt=y(x,w)$

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

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

$\ln P(t|x,w,\beta)=\sum_{n=1}^{N}\ln N(t_n|w^T\phi(x_n),\beta^{-1})$
$=\frac{N}{2}\ln\beta-\frac{N}{2}\ln(2\pi)-\beta\{\frac{1}{2}\sum_{n=1}^{N}\{t_n-w^T\phi(x_n)\}^2\}$

$\nabla\ln P(t|x,w,\beta)=\sum_{n=1}^{N}\{t_n-w^T\phi(x_n)\}\phi^T(x_n)=0$

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

$w_{ML}=(\Phi^T\Phi)^{-1}\Phi^Tt$, where $\Phi=\begin{pmatrix}\phi_0(x_1)&\phi_1(x_1)\cdots&\phi_{M-1}(x_1)\\ \cdots&\cdots&\cdots\end{pmatrix}$

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

## 3.3 Bayesian Linear Regression

Given $w$ conjugate prior $P(w)=N(w|m_0,S_0)$

$P(w|t)\propto P(t|w)\cdot P(w)=N(w|m_N,S_N)$

$\Rightarrow m_N=S_N(S_0^{-1}m_0+\beta\Phi^Tt)=W_{MAP}$, $S_N^{-1}=S_0^{-1}+\beta\Phi^T\Phi$

$P(w|\alpha)=N(w|0,\alpha^{-1}I)\Rightarrow m_N=\beta S_N\Phi^Tt$,
$S_N^{-1}=\alpha I+\beta\Phi^T\Phi$ $\therefore P(w|\alpha)=[\frac{E}{2}(\frac{\alpha}{2})^{1/2}\frac{1}{\Gamma(1/q)}]^M\exp$
$(-\frac{\alpha}{2}\sum_{j=1}^{M}|w_j|^q)$

### 3.3.1 Prediction distribution

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

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

where $\delta_N^2(x)=\frac{1}{\beta}+\phi(x)^TS_N\phi(x)$

### 3.3.2 Equavalent kernel

$y(x,m_N)=m_N^T\phi(x)=\beta\phi^T(x)\cdot S_N\Phi^T(t)=\sum_{n=1}^{N}\beta\phi^T(x)S_N\phi(x_n,t_n)$
$=\sum_{n=1}^{N}k(x,x_n)\cdot t_n$, where $k(x,x')=\beta\cdot\phi^T(x)\cdot S_N\phi(x')$

## 3.4 Baysian Model Comparsion

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

$P(D)=\int P(D|w)P(w)dw\propto P(D|W_{MAP})\cdot\frac{\Delta W_{posterior}}{\Delta W_{prior}}$

$\Rightarrow\ln P(D)=\ln P(D|W_{MAP})+\ln(\frac{\Delta W_{posterior}}{\Delta W_{prior}})$

## 3.5 The Evidence Approximation

### 3.5.1: Predictive distribution

$P(t,\mathbf{t})=\iint P(t|w,\beta)P(w|\mathbf{t},\alpha,\beta)$ • $P(\alpha,\beta|\mathbf{t})dw\,d\alpha\,d\beta$; $P(t,\mathbf{t})\simeq P(t|\mathbf{t},\hat{\alpha},\hat{\beta})=$
$\int P(t|w,\hat{\beta})P(w|\mathbf{t},\hat{\alpha},\hat{\beta})dw$; $(\hat{\alpha},\hat{\beta})=\arg\max_{(\alpha,\beta)}P(\alpha,\beta|\mathbf{t})=$
$\arg\max_{(\alpha,\beta)}P(\mathbf{t}|\alpha,\beta)\cdot P(\alpha,\beta)$

### 3.5.2 Evaluation of evidence function:

$P(\mathbf{t}|\alpha,\beta)=\int P(\mathbf{t}|w,\beta)\cdot P(w|\alpha)dw$
$=(\frac{\beta}{2\pi})^{N/2}(\frac{\alpha}{2\pi})^{M/2}\int\exp\{-E(w)\}dw$

where $E(w)=\beta E_D(w)+\alpha\cdot E_W(w)=\frac{\beta}{2}\|t-\Phi w\|^2+\frac{\alpha}{2}w^Tw$
$=E(m_N)+\frac{1}{2}(w-m_N)^TA(w-m_N)$; $\int\exp\{-E(w)\}dw=$
$\exp\{-E(m_N)\}\cdot\int\exp\{-\frac{1}{2}(w-m_N)^TA(w-m_N)^T\}dw$
$=\exp\{-E(m_N)\}(2\pi)^{M/2}|A|^{-\frac{1}{2}}$

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

### 3.5.3 Maximizing the evidence function

① $\hat{\alpha}=\arg\max P(\mathbf{t}|\alpha,\beta)$; Given the eigenvector equation $(\beta\Phi^T\Phi)u_i=\lambda_iu_i$. And $A$ has eigenvalues $\{\alpha+\lambda_i\}$, $\frac{d}{d\alpha}|A|=\frac{d}{d\alpha}\ln\prod_i(\lambda_i+\alpha)$
$=\frac{d}{d\alpha}\sum_i\ln(\lambda_i+\alpha)=\sum_i\frac{1}{\lambda_i+\alpha}$ ※ $\therefore\frac{d}{d\alpha}\ln P(\mathbf{t}|\alpha,\beta)=\frac{M}{2\alpha}-\frac{1}{2}m_N^Tm_N-$
$\frac{1}{2}\sum_i\frac{1}{\lambda_i+\alpha}=0$; $\alpha m_N^Tm_N=M-\alpha\sum_{i=1}^{M}\frac{1}{\lambda_i+\alpha}\overset{\text{let}}{=}r=\sum_i\frac{\lambda_i}{\alpha+\lambda_i}$ $\therefore\alpha=\frac{r}{m_N^Tm_N}$ ※

② $\hat{\beta}=\arg\max P(\mathbf{t}|\alpha,\beta)$: $\frac{d}{d\beta}\ln|A|=\frac{d}{d\beta}\sum_i\ln(\lambda_i+\alpha)$
$=\frac{1}{\beta}\sum_i\frac{\lambda_i}{\lambda_i+\alpha}=\frac{r}{\beta}$; $\frac{d}{d\beta}\ln P(\mathbf{t}|\alpha,\beta)=\frac{N}{2\beta}-\frac{1}{2}\sum_{n=1}^{N}\{t_n-m_N^T\phi(x_n)\}-\frac{r}{2\beta}=0$

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

## 4.3 Probabilistic Discriminative Model

必考

### logistic regression

Let $P(C_1|\phi)=y(\phi)=\delta(w^T\phi)$
$P(C_2|\phi)=1-P(C_1|\phi)$, $D=\{\phi_n,t_n\}_{n=1}^{N}$, $t_n=\{0,1\}$,
$\phi_n=\phi(x_n)$ likelihood function is: $P(t|w)=\prod_{i=1}^{N}y_n^{t_n}\{1-y\}^{(1-t_n)}$

$E(w)=-\ln P(t|w)=-\sum_{n=1}^{N}\{t_n\ln y_n+(1-t_n)\ln(1-y_n)\}$; $\nabla_wE(w)=\sum_{n=1}^{N}(y_n-t_n)\phi_n$ ※

### 4.3.1 Iterative Reweighted Least Squares (IRLS)

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

### 4.3.2 Laplace Approximation

$P(z)=\frac{f(z)}{\int f(z)dz}$ (Taylor series)

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

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

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

$P(z)$ is updated by $Q(z)$, where $Q(z)=\frac{|A|^{1/2}}{(2\pi)^{M/2}}\exp\{-\frac{1}{2}(z-z_0)^TA(z-z_0)\}$

### 4.4 Model comparison and BIC

We have mode evidence:
$P(D)=\int P(D|\theta)P(\theta)\cdot d\theta$; $\frac{P(D|\theta)\cdot P(\theta)}{P(D)}=P(\theta|D)=\frac{f(\theta)}{Z=\int f(\theta)d\theta}$

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

## Supplement: Chapter 1. (decision theroy & information theory)

1.3 Decision Theory (Case of Cancer): $\hat{C}_{MAP}=\arg\max P(C_k|x)=\arg\max\frac{P(C_k)\cdot P(x|C_k)}{P(x)}$

1.3.1 Minimizing the misclassification Rate: $P(\text{mistake})=P(x\in R_1,C_2)+P(x\in R_2,C_1)=\int_{R_1}P(x,C_2)dx+\int_{R_2}P(x,C_1)dx$ 1.3.2 Minimizing the expected loss

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

1.3.3 loss function for regression: $E[L]=\iint L(t,y(x))\cdot P(x,t)dxdt=\int\{y(x)-t\}^2P(x,t)dt$; $\frac{\partial E[L]}{\partial y(x)}=2\int\{y(x)-t\}\cdot P(x,t)\cdot dt=0$; $y^*(x)=\frac{\int tP(x,t)}{P(x)}=E_t[t|x]$

## 6 kernal function:

### 6.1 Dual Representation:

Given $J(w)=\frac{1}{2}\sum_{n=1}^{N}\{w^T\phi(x_n)-t_n\}^2+\frac{\lambda}{2}w^Tw$ ; $\nabla_w J(w)=0 \rightarrow w_{opt}=-\frac{1}{\lambda}\sum_{n=1}^{N}\{w^T\phi(x_n)-t_n\}\phi(x_n)=\sum_{n=1}^{N}a_n\phi(x_n)=\Phi^T a$ ;

where $a_n=-\frac{1}{\lambda}\{w^T\phi(x_n)-t_n\}$, and $\Phi^T=[\phi(x_1),\phi(x_2),\cdots,\phi(x_n)]$, $a=[a_1,\cdots,a_n]^T \Rightarrow J(a)=\frac{1}{2}a^T\Phi\Phi^T\Phi\Phi^Ta-a^T\Phi\Phi^Tt+\frac{1}{2}t^Tt+\frac{\lambda}{2}a^T\Phi\Phi^Ta$ ; let $K=\Phi\Phi^T$,

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

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

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

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

## 6.4 Gaussian Process

### 6.4.1 Linear regression revisited:

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

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

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

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

### 6.4.2 Gaussian Process for regression:

$t_n=y_n+\epsilon_n$ noise $\quad P(t_n|y_n)=N(t_n|y_n,\beta^{-1})$

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

$P(y)=N(y|0,k)$; $P(t)=\int P(t|y)P(y)\cdot dy$

$=N(t|0,C_{(N)})$, where $C(x_n,x_m)=k(x_n,x_m)+\beta^{-1}\delta_{nm}$

$P(t_{N+1})=P(t_1,\cdots,t_{N+1})=N(t_{N+1}|0,C_{N+1})$

where $C_{N+1}=\begin{pmatrix} C_N & k \\ k^T & c\end{pmatrix}$ $\quad c=k(x_{N+1},x_{N+1})+\beta^{-1}$

$P(t_{N+1}|t)=N(t_{N+1}|m(x_{N+1}),\delta^2(x_{N+1}))$

Learning parameters: $\hat{\theta}=\arg\max_{\theta}\log P(t|\theta)$;

$\ln P(t|\theta)=-\frac{1}{2}\ln|C_N|-\frac{1}{2}t^TC_N^{-1}t-\frac{N}{2}\ln(2\pi)$

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

## 7. sparse kenal Machines

### 7.1 two-class classification problem

$y=w^T\phi(x)+b$, we obtain Distance: $r=\frac{y(x)}{||w||}$

then add class label we will get Margin:

Margin: $\frac{t_ny(x_n)}{||w||}=\frac{t_n(w^T\phi(x_n)+b)}{||w||}$

Maximum margin solution is:

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

with $t_n(w^T\phi(x_n)+b)\geq 1$, we could obtain:

$(\hat{w},\hat{b})=\arg\min_{(w,b)}\frac{1}{2}||w||^2$ ※

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

let $\frac{\partial L}{\partial w}=0$, $\frac{\partial L}{\partial b}=0$ $\therefore w=\sum_{n=1}^{N}a_nt_n\phi(x_n)$ $\quad 0=\sum_{n=1}^{N}a_nt_n$

Let's see Dual representation of the maximum margin problem: $\max_a\{\tilde{L}(a)=\sum_{n=1}^{N}a_n-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}a_na_mt_nt_mK(x_m,x_n)\}$, where $\sum_{n=1}^{N}a_nt_n=0$; $a_n\geq 0$; $t_ny(x_n-1)\geq 0$ ※

### 7.2 In classification problem:

$y(x)=\sum_{n=1}^{N}a_nt_nk(x,x_n)+b$, with the KKT conditions: $a_n\geq 0$, $t_ny(x_n)-1=0$, $a_n\{t_ny(x_n)-1\}=0$

$\therefore a_n=0$ or $t_ny(x_n)=1$ $\quad$ the parameter $b$ is:

$t_n(\sum_{m\in S}a_mt_mk(x_n,x_m)+b)=1 \Rightarrow$

$b=\frac{1}{N_S}\sum_{n\in S}(t_n-\sum_{m\in S}a_mt_mk(x_n,x_m))$; then the error function is: $\sum_{n=1}^{N}E_\infty(y(x_n)t_n-1)+\lambda||w||^2$

#### 7.2.1 Overlapping class distribution:

Slack varibles are introduced to measure for misclassified point: $\xi_n\geq 0$, $n=1,\cdots,N$

And classification constrains are replaced by: $t_ny(x_n)\geq 1-\xi_N$. Therefore, $\min_w(C\sum_{n=1}^{N}\xi_n+\frac{1}{2}||w||^2)$

KKT condition is given by: $a_n\geq 0$, $t_ny(x_n)-1+\xi_n\geq 0$

$a_n(t_ny(x_n)-1+\xi_n)=0$, $\mu_n\geq 0$, $\xi_n\geq 0$, $\mu_n\xi_n=0$

Lagrangian is written by: $L(w,b,a)=\frac{1}{2}||w||^2+C\sum_{n=1}^{N}\xi_n-\sum_{n=1}^{N}a_n\{t_ny(x_n)-1+\xi_n\}-\sum_{n=1}^{N}\mu_n\xi_n$,

where $\partial L/\partial w=0$, $\partial L/\partial b=0$; we obtain: $w=\sum_{n=1}^{N}a_nt_n\phi(x_n)$

$\sum_{n=1}^{N}a_nt_n=0$, $a_n=C-\mu_N$ and dual Lagrange is:

$\min_{\{a_n\}}\{\tilde{L}(a)=\sum_{n=1}^{N}a_n-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}a_na_mt_nt_mk(x_n,x_m)\}$

and subject to $0\leq a_n\leq C$; $\sum_{n=1}^{N}a_nt_n=0$. Finally,

Solution interpretation is: ① $a_n=0\Rightarrow$ nonsupport vector ② $0<a_n<C$, then $\mu_n>0$ then $\xi_n=0$, this point is on the margin.

③ if $a_n=C$, then $\mu_n=0$, $\xi_n\leq 1$ or $\xi_n>1$

To determine $b$, support vector $a_n$ satisfy $0<a_n<C$, $\xi=0$, $t_ny(x_n)=1$, then we have $t_n(\sum_{n\in S}a_mt_mk(x_n,x_m)+b)=1$

$\Rightarrow b=\frac{1}{N_M}\sum_{n\in M}(t_n-\sum_{m\in S}a_mt_mk(x_n,x_m))$

### 7.3 SVM for regression (SVR):

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:

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

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:

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

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

For $y_n-\epsilon\leq t_0\leq y_n+\epsilon\Rightarrow\xi_n=\hat{\xi}_n=0$

Error function of SVR:

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

Constrains: $\xi_n\geq 0$ & $\hat{\xi}_n\geq 0$ & $t_n\leq y(x_n)+\epsilon+\xi_n$ & $t_n\geq y(x_n)-\epsilon-\hat{\xi}_n$

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

$\Rightarrow\{\frac{\partial L}{\partial w}=0;\frac{\partial L}{\partial b}=0;\frac{\partial L}{\partial\xi_n}=0;\frac{\partial L}{\partial\hat{\xi}_n}=0\}$

we obtain: $w=\sum_{n=1}^{N}(a_n-\hat{a}_n)\phi(x_n)$, $\sum_{n=1}^{N}(a_n-\hat{a}_n)=0$;

$\hat{a}_n+\hat{\mu}_N=0$; $a_n+\mu_n=0$

Dual presentation can be defined as:

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

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

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:

$b=t_n-\epsilon-w^T\phi(x_n)$

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

## 9.2: Mixture of Gaussians:

Given $P(x)=\sum_{k=1}^{K}\pi_kN(x|\mu_k,\Sigma_k)$, where $\pi_k$ is mixture weight. $\therefore P(z_k=1)=\pi_k$, then $P(x)=\sum_z P(z)\cdot P(x|z)=\sum_{k=1}^{K}\pi_k(x|\mu_k,\Sigma_k)$.

### 9.2.1 Maximum likelihood:

Let $X=\{x_1,\cdots,x_n\}$ & $Z=\{z_1,\cdots,z_n\}$, then the likelihood function is: $\ln P(X|\pi,\mu,\Sigma)=\sum_{n=1}^{N}\ln\{\sum_{k=1}^{K}\pi_kN(x_n|\mu_k,\Sigma_k)\}$

### 9.2.2: EM for Gaussian mixtures.

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

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

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

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

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

### 9.2.3: EM algorithm for Gaussian mixtures:

① Initialize $\mu_k,\Sigma_k$, evaluate log likelihood.

② E-step: $r(z_{nk})=\frac{\pi_kN(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}\pi_jN(x_n|\mu_j,\Sigma_j)}$

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

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