Polynomial Curve Fitting

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

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

$\|\vec{w}\|^2 \equiv \vec{w}^T\vec{w} = w_0^2 + w_1^2 + \cdots + w_n^2$

The Rules of Probability

sum rule: $p(X) = \sum_Y p(X,Y)$ product rule: $p(X,Y) = p(Y|X)p(X)$

pdf $\int_{-\infty}^{\infty} p(x)dx = 1$. CDF: $\int_{-\infty}^{z} p(x)dx = P(z)$

expectation: $E[f] = \sum_x p(x)f(x)$ or $E[f] = \int p(x)f(x)dx$

conditional expectation: $E_x[f|y] = \sum_x p(x|y)f(x)$

variance: $\mathrm{Var}[f] = E[(f(x) - E[f(x)])^2] = E(x^2) - [E(x)]^2$

covariance: $\mathrm{cov}(x,y) = E_{x,y}[\{x - E[x]\}\{y - E[y]\}] = E_{xy}[xy] - E[x]E[y]$

matrix: $\mathrm{cov}(\vec{x},\vec{y}) = E_{xy}[\vec{x}\vec{y}^T] - E[\vec{x}]E[\vec{y}^T]$

Bayesian Prob: $p(\vec{w}|D) = \frac{p(D|\vec{w})p(\vec{w})}{p(D)}$ ($p(\vec{w}|D)$: posterior, $p(\vec{w})$: prior, $p(D|\vec{w})$: likelihood)

posterior $\propto$ likelihood $\times$ prior

Normal/Gaussian $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$, $\beta = \frac{1}{\sigma^2}$ precision

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

ML: $\ln p(\vec{x}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$. $\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N}x_n$

$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2$, $E(\mu_{ML}) = \mu$, $E(\sigma^2_{ML}) = (\frac{N-1}{N})\sigma^2$

$\tilde{\sigma}^2 = \frac{N}{N-1}\cdot\sigma^2_{ML} = \frac{1}{N-1}\sum(x_n-\mu_{ML})^2$ unbiased.

Bernoulli $\mathrm{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}$, $E(x) = \mu$, $V(x) = \mu(1-\mu)$

$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N}x_n$ (sample mean)

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

$E(m) = \sum_{m=0}^{N} m\,\mathrm{Bin}(m|N,\mu) = N\mu$, $V(m) = \sum_{m=0}^{N}(m-E(m))^2\mathrm{Bin}(m|N,\mu) = N(1-\mu)\mu$

Beta $\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$, $E(\mu) = \frac{a}{a+b}$, $V(\mu) = \frac{ab}{(a+b)^2(a+b+1)}$

Multinomial Bern: $p(\vec{x}|\vec{\mu}) = \prod_{k=1}^{K}\mu_k^{x_k}$, $\sum_k p(\vec{x}|\vec{\mu}) = \sum_{k=1}^{K}\mu_k = 1$, $E[\vec{x}|\vec{\mu}] = \vec{\mu}$

log-likelihood: $p(D|\vec{\mu}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mu_k^{x_{nk}} = \prod_{k=1}^{K}\mu_k^{m_k}$ ($\sum_n x_{nk} = m_k$)

$\mu_k^{ML} = \frac{m_k}{N}$ → observation size (prop)

exp family $p(\vec{x}|\vec{\eta}) = h(\vec{x})g(\vec{\eta})\exp\{\vec{\eta}^T\vec{u}(\vec{x})\}$. $\vec{x}$ scale/vector, $\vec{\eta}$ natural paras of ∗ distribution, $\vec{u}(\vec{x})$ some function of $\vec{x}$.

cts/discrete.

Bern: $p(x|\mu) = \mathrm{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}$, $p(x|\mu) = \exp\{x\ln\mu + (1-x)\ln(1-\mu)\} = (1-\mu)\exp\{\ln(\frac{\mu}{1-\mu})x\}$

so $\eta = \ln(\frac{\mu}{1-\mu})$, $\sigma(\eta) = \mu = \frac{1}{1+\exp(-\eta)}$

$p(x|\eta) = \sigma(-\eta)\exp(\eta x)$ ⇒ $u(x) = x$, $h(x) = 1$, $g(\eta) = \sigma(-\eta)$

Multinomial: $\vec{u}(\vec{x}) = \vec{x}$, $h(\vec{x}) = 1$, $g(\vec{\eta}) = 1$.

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

$\vec{\eta} = (\mu/\sigma^2, -\frac{1}{2\sigma^2})^T$, $\vec{u}(x) = (x, x^2)^T$, $h(\vec{x}) = (2\pi)^{-\frac{1}{2}}$, $g(\vec{\eta}) = (-2\eta_2)^{\frac{1}{2}}\exp(\frac{\eta_1^2}{4\eta_2})$

ML & SS(exp family): gradient (**)

$\nabla g(\vec{\eta})\int h(\vec{x})\exp\{\vec{\eta}^T\vec{u}(\vec{x})\}d\vec{x} + g(\vec{\eta})\int h(\vec{x})\exp\{\vec{\eta}^T\vec{u}(\vec{x})\}\vec{u}(\vec{x})d\vec{x} = 0$

use (**): $-\frac{1}{g(\vec{\eta})}\nabla g(\vec{\eta}) = g(\vec{\eta})\int h(\vec{x})\exp\{\vec{\eta}^T\vec{u}(\vec{x})\}\vec{u}(\vec{x})d\vec{x} = E(\vec{u}(\vec{x}))$

$-\nabla\ln g(\vec{\eta}) = E(\vec{u}(\vec{x}))$

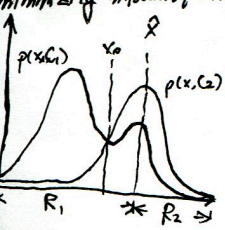
* **Likelihood function**: $p(\vec{X}|\vec{\eta}) = (\prod_{n=1}^{N}h(\vec{x}_n))g(\vec{\eta})^N\exp\{\vec{\eta}^T\sum_{n=1}^{N}\vec{u}(\vec{x}_n)\}$

$\vec{X} = \{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_n\}$. set gradient of $\ln p(\vec{X}|\vec{\eta}) = 0$.

$-\nabla\ln g(\vec{\eta}_{ML}) = \frac{1}{N}\sum_{n=1}^{N}\vec{u}(\vec{x}_n)$ ⇒ $\vec{\eta}_{ML} = ?$ depends only on $\sum_n \vec{u}(\vec{x}_n)$ → sufficient statistics

Decision theory

minimizing misclassification rate: $p(\text{mistake}) = p(\vec{x}\in R_1, C_2) + p(\vec{x}\in R_2, C_1) = \int_{R_1}p(\vec{x},C_2)d\vec{x} + \int_{R_2}p(\vec{x},C_1)d\vec{x}$



$p(\text{correct}) = \sum_{k=1}^{K}p(\vec{x}\in R_k, C_k) = \sum_{k=1}^{K}\int_{R_k}p(\vec{x},C_k)d\vec{x}$

minimizing expected loss: $E[L] = \sum_k\sum_j\int_{R_j}L_{kj}p(\vec{x},C_k)d\vec{x}$

loss function for reg: $E[L] = \iint L(t,y(\vec{x}))p(\vec{x},t)d\vec{x}dt$

$L(t,y(\vec{x})) = \{y(\vec{x}) - t\}^2$ common.

$y(\vec{x}) = \int t\,p(t|\vec{x})dt = E(t|\vec{x})$

· normalized: $\int \text{density} = 1$

linear basis func reg model: $y(x,\vec{w}) = \sum_{j=0}^{M-1}w_j\phi_j(x) = \vec{w}^T\vec{\phi}(x)$

Two bias parameter, $\phi_j(x)$: basis function

so. defined $\phi_0(x) = 1$ ⇒ $y(x,w) = \sum_{j=0}^{M-1}w_j\phi_j(x) = w^T\phi(x)$

maximum likelihood & least square

$\ln p(\vec{t}|\vec{w},\beta) = \sum_{n=1}^{N}\ln\mathcal{N}(t_n|\vec{w}^T\vec{\phi}(\vec{x}_n),\beta^{-1})$

$= \frac{1}{2}N\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\vec{w})$

where sum of squares error function is $E_D(\vec{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \vec{w}^T\vec{\phi}(\vec{x}_n)\}^2$

$\vec{w}_{ML} = (\vec{\Phi}^T\vec{\Phi})^{-1}\vec{\Phi}^T\vec{t}$

regularized least square: introduce to control over-fitting in the form:

$E_D(\vec{w}) + \lambda E_W(\vec{w})$

· $\frac{1}{2}\sum_{n=1}^{N}\{t_n - \vec{w}^T\vec{\phi}(\vec{x}_n)\}^2 + \frac{\lambda}{2}\vec{w}^T\vec{w}$

λ: regularization coefficient.

· setting gradient of ← to be zero, wrt $\vec{w}$, get $\vec{w} = (\lambda I + \vec{\Phi}^T\vec{\Phi})^{-1}\vec{\Phi}^T\vec{t}$ (ridge regression)

bias-var decomposition

expected loss = (bias)^2 + variance + noise

$(\text{bias})^2 = \int\{E_D[y(\vec{x};D)] - h(\vec{x})\}^2p(\vec{x})d\vec{x}$

$\text{variance} = \int E_D[\{y(\vec{x};D) - E_D[y(\vec{x};D)]\}^2]p(\vec{x})d\vec{x}$

$\text{noise} = \int\{h(\vec{x}) - t\}^2p(\vec{x},t)d\vec{x}dt$

"trade-off" between bias & var.

bayesian inference:

log of posterior := Σ(log likelihood + log of prior)

$\ln p(\vec{w}|\vec{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \vec{w}^T\vec{\phi}(\vec{x}_n)\}^2 - \frac{\alpha}{2}\vec{w}^T\vec{w} + \text{const.}$

predictive distribution: ⇒(integrating over posterior) $p(t|\vec{x},\vec{t},\alpha,\beta) = \mathcal{N}(t|\vec{m}_N^T\vec{\phi}(\vec{x}), \sigma_N^2(\vec{x}))$

where $\sigma_N^2(\vec{x}) = \frac{1}{\beta} + \vec{\phi}(\vec{x})^T\vec{S}_N\vec{\phi}(\vec{x})$

bayesian model comparison

$\ln p(D) \approx \ln p(D|\vec{w}_{MAP}) + M\ln(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}})$

marginal likelihood function

$p(\vec{t}|\alpha,\beta) = \int p(\vec{t}|\vec{w},\beta)p(\vec{w}|\alpha)d\vec{w}$

weight parameter $\vec{w}$.

discriminant function: $y(\vec{x}) = \vec{w}^T\vec{x} + w_0$

$\vec{w}$ weight vector, w_0 bias. C_1 if $y(\vec{x}) \geq 0$, C_2 o.w. $y(\vec{x}) = 0$ decision surface.

$\frac{\vec{w}^T\vec{x}}{\|\vec{w}\|} = -\frac{w_0}{\|\vec{w}\|}$ normal distance from origin to surface.

Simple solution for K-class discriminant

$y_k(x) = x^Tw_k + w_{k0}$, $k = 1,\cdots K$.

assign x to C_k if $y_k(x) > y_j(x)$ $\forall j \neq k$.

A, B in R_k: $y_k(x_A) > y_j(x_A)$, $y_k(x_B) > y_j(x_B)$

⇒ $\alpha > 0$, $y_k(\alpha x_A + (1-\alpha)x_B) > y_j(\alpha x_A + (1-\alpha)x_B)$

Fisher's linear discriminant

Gaussian example (prob. generative models)

$p(\vec{x}|C_k) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\vec{\Sigma}|^{\frac{1}{2}}}\exp(-\frac{1}{2}(\vec{x}-\vec{\mu}_k)^T\vec{\Sigma}^{-1}(\vec{x}-\vec{\mu}_k))$

2 classes, posterior is logistic func:

$p(C_k|\vec{x}) = \sigma(\vec{w}^T\vec{x} + w_0)$

where $\vec{w} = \vec{\Sigma}^{-1}(\vec{\mu}_1 - \vec{\mu}_2)$

$w_0 = -\frac{1}{2}\vec{\mu}_1^T\vec{\Sigma}^{-1}\vec{\mu}_1 + \frac{1}{2}\vec{\mu}_2^T\vec{\Sigma}^{-1}\vec{\mu}_2 + \ln\frac{p(C_1)}{p(C_2)}$

reg model with square loss func.

$E[L] = \iint(y(\vec{x}) - t)^2p(\vec{x},t)dxdt$

Show $y(x)$ for which $E[L]\downarrow$ by conditional expectation

$E(t|x) = \int t\,p(t|x)dt$.

$E[L] = \iint y(x)^2p(x,t)dtdx + \iint t^2p(x,t)dxdt - 2\iint y(x)tp(x,t)dxdt$

$= \int y(x)^2p(x)dx + E[t^2] - 2\int y(x)p(x)E[t|x]dx$

$= E_t(t^2) + E_x[y(x)^2 - 2y(x)E(t|x)]$ ✓

linear reg model: $p(t|x,w,\beta) = \prod_{n=1}^{N}\mathcal{N}(t_n|w^T\phi(x_n),\beta^{-1})$

· $\log p(w|x,t) = \log p(t|x,w) + \log p(w|\alpha) + \text{constant}$

$= -\frac{1}{2}w^T\alpha Iw + \sum_{n=1}^{N}[t_n\ln y_n + (1-t_n)\ln(1-y_n)] + \text{cons}$

⇒ $p(w|x,t) \propto \exp\{-\frac{1}{2}\alpha w^Tw\}\{\prod_{n=1}^{N}y_n^{t_n}(1-y_n)^{(1-t_n)}\}$

where $y_n = \sigma(w^Tx_n) = \frac{1}{1+\exp(-w^Tx_n)}$

· $-\log p(w|x,t) \Leftrightarrow \downarrow -\Sigma(\ ) + \frac{\alpha}{2}w^Tw = \sum_{n=1}^{N}E_n + \frac{\alpha}{2}w^Tw$

$\lambda = \alpha$, $E_n = \begin{cases} -\ln y_n & \text{if } t_n = 1 \\ -\ln(1-y_n) & \text{if } t_n = 0 \end{cases}$

ΣE_n is cross-entropy error in total.

· derivative of E_n (cross entropy)

$E[L] = -\sum_{n=1}^{N}\{t_n\ln y_n + (1-t_n)\ln(1-y_n)\} + \frac{\alpha}{2}w^Tw$

$= \sum_{n=1}^{N}E_n + \frac{\alpha}{2}w^Tw$

$\frac{dE_n}{dy_n} = \frac{y_n - t_n}{y_n(1-y_n)}$, $\frac{dy_n}{dw} = y_n(1-y_n)x_n$

since $\frac{dy_n}{dw_n} = \frac{d\sigma(w^Tx_n)}{dw} = y_n(1-y_n)x_n$

⇒ $\frac{dE_n}{dw} = (y_n - t_n)x_n$ ⇒ $\nabla_w E[L] = \sum(y_n - t_n)x_n + \alpha w$

Bayesian linear regression:

likelihood: $p(t|X,w,\beta) = \prod_{n=1}^{N}\mathcal{N}(t_n|w^T\phi(x_n),\beta^{-1})$

conjugate prior: $p(w) = \mathcal{N}(w|m_0,S_0)$

⇒ $p(w|t,X,w,\beta) \propto (\ )(\ )$

posterior takes $S_N(S_0^{-1}m_0 + \beta\Phi^Tt)$. $S_N^{-1} = S_0^{-1} + \beta\Phi^T\Phi$

$w_{ML} = (\Phi^T\Phi)^{-1}\Phi^Tt$ ⇒ $m_N = S_N(S_0^{-1}m_0 + \beta\Phi^T\Phi w_{ML})$.

Predictive distribution: integrating over posterior

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

$= \mathcal{N}(t|m_N^T\phi(x), \sigma_N^2(x))$

where $\sigma_N^2(x) = \frac{1}{\beta} + \phi(x)^TS_N\phi(x)$. $m_N = \beta S_N\Phi^Tt$.

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

evidence approximation: fully Bayesian predictive distribution: $p(t^*|x^*,D) = \iiint p(t^*|x^*,w,\beta)p(w|D,\alpha,\beta)P(\alpha,\beta|D)dwd\alpha d\beta$

LS for classification $y_k(x) = x^Tw_k + w_{k0}$, $k = 1,\cdots k$.

or $y(\tilde{x}) = \tilde{w}^T\tilde{x}$ ($\tilde{x} \in (x_1,\cdots x_{D+1})^T$, $\tilde{w}_k = (w_{k0}, w_k^T)^T$)

$\tilde{w} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^TT$

Prob Generative Models

posterior: $p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1) + p(x|C_2)p(C_2)} = \frac{1}{1+\exp(-a)} = \sigma(a)$

where $a = \ln\frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)} = \ln\frac{p(C_1|x)}{1-p(C_1|x)}$

$\sigma(-a) = 1 - \sigma(a)$

$\frac{d}{da}\sigma(a) = \sigma(a)(1-\sigma(a))$ Sigmoid

A2Q1 minimize weighted SSE

$\frac{\partial E_D(\vec{w})}{\partial w} = -\sum_{n=1}^{N}r_n(t_n - w^T\phi(x_n))\phi(x_n)^T$ set $= 0$

⇒ $\sum_{n=1}^{N}r_nt_n\phi(x_n)^T - w^T\sum_{n=1}^{N}r_n\phi(x_n)\phi(x_n)^T = 0$

let $t = [t_1,\cdots,t_N]^T$, $\Phi = \begin{pmatrix}\phi_0(x_1) & \cdots & \phi_{m-1}(x_1)\\ \vdots & & \vdots \\ \phi_0(x_N) & \cdots & \phi_{m-1}(x_N)\end{pmatrix}$

$R = \begin{pmatrix} r_1 & & 0 \\ & \ddots & \\ 0 & & r_n\end{pmatrix}$ ⇒ $t^TR\Phi = w^T(\Phi^TR\Phi^T)$

⇒ $w = (\Phi^TR\Phi)^TRt$

$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_N) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_N) \end{bmatrix}, \begin{bmatrix} k(x_1,x_1) & \cdots & k(x_1,x_N) \\ \vdots & \ddots & \vdots \\ k(x_N,x_1) & \cdots & k(x_N,x_N) \end{bmatrix} \right)$

mean, covariance kernel

commonly use $k(x_n,x_m) = E[f(x_n)f(x_m)] = \exp(-\frac{\theta}{2}\|x_n - x_m\|^2)$

GP Regression noise

$t_n = f_n + \varepsilon_n$ (Gaussian) Gaussian noise $p(t_n|f_n) = \mathcal{N}(t_n|f_n, \beta^{-1})$

so given $X=\{x_1,\cdots,x_N\}$, $t=[t_1,\cdots,t_N]$. the conditional is

$p(t|f) = \mathcal{N}(t|f, \beta^{-1}I_N)$.

Marginal distribution GP $p(f) = \mathcal{N}(f|0,K)$

$p(t) = \int p(t|f)p(f)df = \mathcal{N}(t|0,C)$, covariance matrix $C(x_n,x_m)$

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

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

Prediction predict t_{N+1} for new input x_{N+1}

joint dist over t & t_{N+1} is $p(\begin{bmatrix} t \\ t_{N+1} \end{bmatrix}) = \mathcal{N}(0, \begin{bmatrix} C_N & k \\ k^T & c \end{bmatrix})$

where C_N is $N\times N$ matrix with element

$C_N(x_n,x_m) = k(x_n,x_m) + \beta^{-1}\delta_{nm}$. c is scalar $= k(x_{N+1},x_{N+1}) + \beta^{-1}$

k is N by 1 vector with elements $k(x_n,x_{N+1})$

conditional distribution is Gaussian $p(t_{N+1}|t) = \mathcal{N}(m(x_{N+1}), \sigma^2(x_{N+1}))$

with $m(x_{N+1}) = k^T C_N^{-1} t$, $\sigma^2(x_{N+1}) = c - k^T C_N^{-1} k$

key results of GP!

hyperparameters (θ like θ!)

max log of marginal likelihood wrt θ: $\ln p(t|\theta) = -\frac{1}{2}\ln|C_N| - \frac{1}{2}t^T C_N^{-1} t - \frac{N}{2}\ln(2\pi)$

$\Rightarrow$ gradient $\frac{\partial}{\partial\theta_i}\ln p(t|\theta) = -\frac{1}{2}\mathrm{Tr}(C_N^{-1}\frac{\partial C_N}{\partial\theta_i}) + \frac{1}{2}t^T C_N^{-1}\frac{\partial C_N}{\partial\theta_i}C_N^{-1}t$, $\ln p(t|\theta) = -\frac{1}{2}\ln|C_N| - \frac{1}{2}t^T C_N^{-1} t - \frac{N}{2}\ln(2\pi)$

Automatic Relevance Determination

$k(x_n,x_m) = \theta_0 \exp(-\frac{1}{2}\sum_{i=1}^D \eta_i(x_{ni}-x_{mi})^2) + \theta_2 + \theta_3 x_n^T x_m$, $\ln p(t|\theta) = -\frac{1}{2}\ln|C_N| - \frac{1}{2}t^T C_N^{-1} t - \frac{N}{2}\ln(2\pi)$

Mixture models

objective $J = \sum_{n=1}^N \sum_{k=1}^K r_{nk}\|x_n - \mu_k\|^2$, each data x_n

as binary vector r_n of length K indicates which K cluster

x_n is in. μ_k is prototype. try to minimize J. (K-Means)

(clustering) given μ_k, (Estep) $r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|x_n - \mu_j\|^2 \\ 0 & \text{o.w.} \end{cases}$

(M-) minimize J: i.e. assign x_n to its closest center. (Estep)

(Mstep) given r_{nk}, minimize J wrt μ_k. $\mu_k = \frac{\sum_n r_{nk}x_n}{\sum_n r_{nk}}$

i.e. set μ_k equal to the mean of all data assigned to cluster k.

Mixture Gaussian

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

z K-dim binary r.v. $z = [z_1,\cdots,z_K]$, $z_k = \{0,1\}$, $\sum_k z_k = 1$

$p(z_k=1) = \pi_k$, $0 \le \pi_k \le 1$, $\sum_k \pi_k = 1$. $p(z) = \prod_{k=1}^K \pi_k^{z_k}$

conditional distri. $p(x|z_k=1) = \mathcal{N}(x|\mu_k,\Sigma_k)$

$\Rightarrow$ joint dist. $p(x,z) = p(x|z)p(z)$

marginal dist. $p(x) = \sum_z p(z)p(x|z) = \sum_{k=1}^K \pi_k \mathcal{N}(x|\mu_k,\Sigma_k)$

corresponding posterior $p(z_k=1|x) = \frac{p(z_k=1)\,p(x|z_k=1)}{\sum_{j=1}^K p(z_j=1)p(x|z_j=1)} = \frac{\pi_k\mathcal{N}(x|\mu_k,\Sigma_k)}{\sum_{j=1}^K \pi_j\mathcal{N}(x|\mu_j,\Sigma_j)}$

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

ML for Gaussian mixture.

$\ln p(X|\pi,\mu,\Sigma) = \sum_{n=1}^N \ln\sum_{k=1}^K \pi_k\mathcal{N}(x|\mu_k,\Sigma_k)$

$\Rightarrow$ diff $\mu_k = \frac{1}{N_k}\sum_n \gamma(z_{nk})x_n$, $N_k = \sum_n \gamma(z_{nk})$

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

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

EM: update $\gamma(z_{nk})$ $\sigma^2 \to 0$ posterior covariance $\to 0$.

M: update μ_k, π_k, Σ_k & evaluate log-like. don't forget check convergence

General EM: given joint $p(Z,X|\theta)$, max $p(X|\theta)$ wrt θ.

E: compute posterior over latent $p(Z|X,\theta^{old})$

M: find new est of θ^{new}. $\theta^{new} = \arg\max_\theta Q(\theta,\theta^{old})$

where $Q(\cdot) = \sum_Z p(Z|X,\theta^{old})\ln p(X,Z|\theta)$

Gaussian Mixture Revisit

$p(X,Z|\pi,\mu,\Sigma) = \prod_{n=1}^N\prod_{k=1}^K [\pi_k\mathcal{N}(x|\mu_k,\Sigma_k)]^{z_{nk}}$

$\ln p(X,Z|\pi,\mu,\Sigma) = \sum_{n=1}^N\sum_{k=1}^K z_{nk}[\ln\pi_k + \ln\mathcal{N}(x|\mu_k,\Sigma_k)]$.

max $\Rightarrow \pi_k = \frac{1}{N}\sum_n z_{nk}$, similarly for mean & cov.

expected complete log-like: $\to \ln p(X,Z|\pi,\mu,\Sigma) = \sum_n\sum_k z_{nk}(\ln\pi_k + \ln\mathcal{N}(x_n|\mu_k,\Sigma_k))$

$E[z_{nk}] = \frac{\pi_k\mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^K \pi_j\mathcal{N}(x_n|\mu_j,\Sigma_j)} = \gamma(z_{nk})$

$\Rightarrow E_Z[\ln p(X,Z|\cdots)] = \sum_n\sum_k \gamma(z_{nk})[\ln\pi_k + \ln\mathcal{N}(\cdots)]$

maximizing $\Rightarrow$ same result.

PCA: goal find M eigenvectors of SC cov matrix) -- projection:

select M eigenvectors of SC cov matrix) -- projection:

$\begin{bmatrix} u_1^T \\ \vdots \\ u_M^T \end{bmatrix}[x_1,\cdots x_N] = [z_1,\cdots,z_N]$ { max var / min errors (Let $M=1$)

Max var: $\frac{1}{N}\sum_{n=1}^N \{u_1^T x_n - u_1^T\bar{x}\}^2 = u_1^T S u_1$

where $\bar{x} = \frac{1}{N}\sum_{n=1}^N x_n$, $S = \frac{1}{N}\sum_{n=1}^N (x_n-\bar{x})(x_n-\bar{x})^T$

constrain $\|u_1\|=1$ (use Lagrange multiplier) to max

$u_1^T S u_1 + \lambda(1 - u_1^T S u_1)$, $Su_1 = \lambda_1 u_1 \Rightarrow \lambda_1 = u_1^T S u_1$

so u_1 is opt. component (eigenvector with maximal e.v.)

min error

$u_i^T u_j = \delta_{ij}$, $x_n = \sum_{i=1}^D \alpha_{ni}u_i$, $\alpha_{ni} = x_n^T u_i$

$\tilde{x}_n = \sum_{i=1}^M z_{ni}u_i + \sum_{i=M+1}^D b_i u_i$. Minimize $J = \frac{1}{N}\sum_{n=1}^N \|x_n - \tilde{x}_n\|^2$

wrt. $z_{nj} = x_n^T u_j$; $b_j = \bar{x}^T u_j$

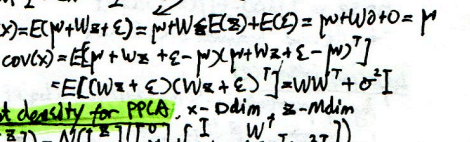
so J reduces to $= \frac{1}{N}\sum_{n=1}^N\sum_{i=M+1}^D (x_n^T u_i - \bar{x}^T u_i)^2 = \sum_{i=M+1}^D u_i^T S u_i$ Backward

$\Rightarrow J = \sum_{i=M+1}^D \lambda_i$ (same results)...

PCA for high dim ($N<D$): $\frac{1}{N}XX^T(Xu_i) = \lambda_i(Xu_i)$

$X^T v_i$ is an eigenvector of S with eigenvalue λ_i.

Probabilistic PCA generative view



Marginal data density (predictive distribution)

$p(x) = \int_z p(z)p(x|z)dz = \mathcal{N}(x|\mu, WW^T + \sigma^2 I)$

$E(x) = E[\mu + Wz + \varepsilon] = \mu + W\cancel{E(z)} + E(\varepsilon) = \mu + W\cdot 0 = \mu$

$C = \mathrm{cov}(x) = E[(\mu + Wz + \varepsilon - \mu)(\mu + Wz + \varepsilon - \mu)^T]$

$= E[(Wz + \varepsilon)(Wz + \varepsilon)^T] = WW^T + \sigma^2 I$

Joint density for PPCA. x- Ddim z-Mdim

$p(\begin{bmatrix} z \\ x \end{bmatrix}) = \mathcal{N}(\begin{bmatrix} z \\ x \end{bmatrix}|\begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} I & W^T \\ W & WW^T + \sigma^2 I \end{bmatrix})$

$\mathrm{cov}(z,x) = E[(z-0)(x-\mu)^T] = E[z(\mu + Wz + \varepsilon - \mu)^T]$

$= E[z(Wz+\varepsilon)^T] = W^T$

- matrix inverse lemma (used): $C^{-1} = \sigma^{-1}I - \sigma^{-2}W(W^TW + \sigma^2 I)^{-1}W^T$

posterior over latent variable:

$p(z|x) = \mathcal{N}(z|m,V)$, $m = M^{-1}W^T(x-\mu)$, $V = \sigma^2 M^{-1}$, $M = W^TW + \sigma^2 I$

- PPCA is a constrained Gaussian.

ML for PCA: $L(\theta;X) = \ln p(X|\theta) = \sum_n \log p(x_n|\theta) = -\frac{N}{2}\log|C| - \frac{1}{2}\sum_n (x_n-\mu)C^{-1}(x_n-\mu)^T$

$= -\frac{N}{2}\log|C| - \frac{1}{2}\mathrm{Tr}[C^{-1}\sum_n (x_n-\mu)(x_n-\mu)^T] + \text{const}$

$\mu_{ML} = \bar{x}$, $\log p(X|\theta) = -\frac{N}{2}\log|C| - \frac{1}{2}\mathrm{Tr}[C^{-1}S] + \text{const}$.

C is model cov, S is sample data covariance

sufficient stat: mean $\bar{x} = \frac{1}{N}\sum_n x_n$, sample cov. S.

max wrt W: $W_{ML} = U_M(L_M - \sigma^2 I)^{\frac{1}{2}}R$. where U_M is $D\times M$

whose cols are given by M principal eigenvectors of S.

L_M is $M\times M$ diagonal containing M largest eigenvalues.

R is arbitrary $M\times M$ orthgon matrix.

EM for PPCA: $\log p(X,Z|\mu,W,\sigma^2) = \sum_n[\log p(x_n|z_n) + \log p(z_n)]$

zero noise limit: derive standard PCA as a limit of PPCA $\sigma^2 \to 0$

ML paras same. Posterior mean reduce to

$\lim_{\sigma^2\to 0}(W^TW + \sigma^2 I)^{-1}W^T(x-\mu) = (W^TW)^{-1}W^T(x-\mu) \to$ standard PCA

Factor Analysis (FA), like PPCA. joint & marginal are also Gaussian.

marginal: $p(x) = \int_z p(z)p(x|z)dz = \mathcal{N}(x|\mu, WW^T + \Psi)$, Ψ $D\times D$ diag

joint: $p(\begin{bmatrix} z \\ x \end{bmatrix}) = \mathcal{N}(\begin{bmatrix} z \\ x \end{bmatrix}|\begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} I & W^T \\ W & WW^T + \Psi \end{bmatrix})$ W $D\times M$ factor loading

autoencoder: input x, reconstruction: $y_k(x,w) = \sum_{j=1}^M w_{kj}^{(2)}\sigma(\sum_{i=1}^D w_{ji}^{(1)}x_i)$,

$k = 1,\cdots,D$, can determine network parameter w by min reconstruction error

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

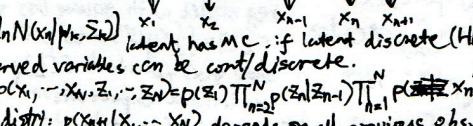
Markov models: future predictions indpdt of all but the most recents observations.

1 order MC $p(x_1,\cdots,x_N) = p(x_1)\prod_{n=2}^N p(x_n|x_{n-1})$ conditional: $p(x_n|x_1,\cdots,x_{n-1}) = p(x_n,x_{n-1})$

2 order MC $p(\cdots) = p(x_1)p(x_2|x_1)\prod_{n=3}^N p(x_n|x_{n-1},x_{n-2})$

kth order MC $p(x_1,\cdots,x_{n+1}|x_1,\cdots,x_n) = \prod_{n=k+1}^N p(x_n|x_{n-1},x_{n-2},\cdots,x_{n-k})$

State Space models (additional latent var introduced)



latent, has MC. if latent discrete (HMMs)

observed variables can be cont/discrete.

joint $p(x_1,\cdots,x_N,z_1,\cdots,z_N) = p(z_1)\prod_{n=2}^N p(z_n|z_{n-1})\prod_{n=1}^N p(x_n|z_n)$

predictive distri: $p(x_{n+1}|x_1,\cdots,x_N)$ depends on all previous obs.

HMM:

1 order MC generates hidden state seq. (transition prob)

$p(z_n=k|z_{n-1}=j) = A_{jk}$, $p(z_1=k) = \pi_k$, $\sum_k A_{jk} = 1$

conditionals: $p(z_n|z_{n-1},A) = \prod_{k=1}^K\prod_{j=1}^K A_{jk}^{z_{n-1,j}z_{nk}}$, $p(z_1|\{k\}) = \prod_{k=1}^K \pi_k^{z_{1k}}$

emission prob. $p(x_n|z_n,\phi) = \prod_{k=1}^K p(x_n|\phi_k)^{z_{nk}} \overset{\text{(e.g.)}}{=} \prod_{k=1}^K \mathcal{N}(x_n|\mu_k,\Sigma_k)^{z_{nk}}$

(continuous) (mixture Gauss)

Joint of HMM: $p(Z,X|\theta) = p(z_1|\pi)\prod_{n=2}^N p(z_n|z_{n-1},A)\prod_{n=1}^N p(x_n|z_n,\phi)$.

$\theta = \{\pi, A, \phi\}$ state trans

EM: E posterior over latent $p(Z|X,\theta^{old})$ $\prod_{n=1}^N p(x_n|z_n,\phi)$

M: max expected complete data log-like $Q(\theta,\theta^{old}) = \sum_Z p(Z|X,\theta^{old})\cdot$

inference of hidden state:

$\gamma(z_n) = p(z_n|X) = \frac{p(X|z_n)p(z_n)}{p(X)}$, $p(z_n|X) = \frac{p(x_1,\cdots,x_n|z_n)p(x_{n+1},\cdots,x_N|z_n)}{p(X)}$ $\cdot P(z_n)$

$= \frac{p(x_1,\cdots,x_n,z_n)p(x_{n+1},\cdots,x_N|z_n)}{p(X)}$

$= \alpha(z_n)\beta(z_n)/p(X)$

$\equiv$ joint prob of obs all data up to time n & z_n

$\equiv$ conditional prob of all future data $n+1$ to N.

Forward recursion: $\alpha(z_n) = p(x_n|z_n)\sum_{z_{n-1}} p(x_1,\cdots,x_{n-1},z_{n-1},z_n)$

$= p(x_n|z_n)\sum_{z_{n-1}} \alpha(z_{n-1})p(z_n|z_{n-1})$

Backward recursion: $\beta(z_n) = \sum_{z_{n+1}} p(x_{n+1},\cdots,x_N,z_{n+1}|z_n)$

$= \sum_{z_{n+1}} p(x_{n+2},\cdots,x_N|z_{n+1})p(x_{n+1}|z_{n+1})p(z_{n+1}|z_n)$

$= \beta\sum_{z_{n+1}} \beta(z_{n+1})p(x_{n+1}|z_{n+1})p(z_{n+1}|z_n)$ $\cdot p(x_{n+1}|z_{n+1})$ $p(z_{n+1}|z_n)$

$\cdot \sum_{z_n}\gamma(z_n) = \sum_{z_n} p(z_n|X) = 1$ $\quad c = \sum_{z_n}\alpha(z_n)$

Likelihood: $p(X|\theta) = \sum_{z_n}\alpha(z_n)\beta(z_n) = \sum_{z_n}\alpha(z_n)$

EM