Take Home Solution

Prof. Krishna R. Pattipati

Dept. of Electrical and Computer Engineering
University of Connecticut
Contact: krishna@engr.uconn.edu (860) 486-2890

ECE 6141
Neural Networks for Classification and Optimization

Copyright ©2018 by K. Pattipati

Problem 1:

1. $p(x \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$

$p(\mu) = \lambda \exp(-\lambda\mu); \mu \geq 0; \lambda > 0$

$p(\mu \mid \{x_n\}_{n=1}^N) \propto p(\{x_n\}_{n=1}^N \mid \mu) p(\mu) = L(\mu)$

$$\min_{\mu \geq 0} J = -\ln L(\mu) = \frac{1}{2\sigma^2}\left(\sum_{n=1}^{N}(x_n-\mu)^2\right) + \lambda\mu + cons\tan t$$

$$\frac{\partial J}{\partial \mu} = \lambda - \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu)$$

$$\Rightarrow \hat{\mu}_{MAP} = \max(\frac{\sum_{n=1}^{N} x_n - \sigma^2\lambda}{N}, 0)$$

$$\frac{\partial^2 J}{\partial \mu^2} = H = \frac{N}{\sigma^2}$$

so, Laplace approximation of posterior : $p(\mu \mid \{x_n\}_{n=1}^N) \sim N(\hat{\mu}_{MAP}, \frac{\sigma^2}{N})$

Copyright ©2018 by K. Pattipati

Problem 2

$a) \underline{x} \in R^n$

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{k \in C_i} \underline{x}_k ; i = 1,2; C_i = \text{samples from class } i; |\, C_i \,| = n_i$$

$$\hat{\Sigma} = \frac{1}{N-2} \left(\sum_{i=1}^{2} \sum_{k \in C_i} (\underline{x}_k - \hat{\mu}_i)(\underline{x}_k - \hat{\mu}_i)^T \right)$$

$$\hat{\pi}_i = \frac{n_i}{N} ; i = 1,2$$

$$g(\underline{x}) = g_2(\underline{x}) - g_1(\underline{x})$$

$$= (\underline{\hat{\mu}}_2 - \underline{\hat{\mu}}_1)^T \hat{\Sigma}^{-1} \underline{x} - [\frac{1}{2} (\underline{\hat{\mu}}_2^T \Sigma^{-1} \underline{\hat{\mu}}_2 - \underline{\hat{\mu}}_1^T \Sigma^{-1} \underline{\hat{\mu}}_1) - \ln \frac{\hat{\pi}_2}{\hat{\pi}_1}]$$

$$\Rightarrow \underline{x}^T \hat{\Sigma}^{-1} (\underline{\hat{\mu}}_2 - \underline{\hat{\mu}}_1) > \frac{1}{2} \underline{\hat{\mu}}_2^T \hat{\Sigma}^{-1} \underline{\hat{\mu}}_2 - \frac{1}{2} \underline{\hat{\mu}}_1^T \hat{\Sigma}^{-1} \underline{\hat{\mu}}_1 + \ln \frac{n_1}{n_2}$$

Copyright ©2018 by K. Pattipati

Problem 1

$$b)\ J = \sum_{i=1}^{2} \sum_{k \in C_i} (z_i - \underline{w}_a^T \underline{x}_{ak})^2 ; \underline{x}_{ak} = [1, \underline{x}_k^T]^T$$

$$\left(\sum_{i=1}^{2} \sum_{k \in C_i} \underline{x}_{ak} \underline{x}_{ak}^T \right) \hat{\underline{w}}_a = \left(\sum_{i=1}^{2} z_i \left(\sum_{k \in C_i} \underline{x}_{ak} \right) \right) = N(\hat{\underline{\mu}}_{2a} - \hat{\underline{\mu}}_{1a})$$

$$where\ \hat{\underline{\mu}}_{ia} = [1, \hat{\underline{\mu}}_i^T]^T$$

$$For\ w_0\ term: N\hat{w}_0 = -\hat{\underline{w}}^T \sum_{i=1}^{2} \sum_{k \in C_i} \underline{x}_k = -\hat{\underline{w}}^T \sum_{i=1}^{2} n_i \underline{\mu}_i \Rightarrow \hat{w}_0 = -\hat{\underline{w}}^T \underline{\mu}$$

$$For\ \underline{w}\ term: N\underline{\mu}\hat{w}_0 + \left[(N-2)\hat{\Sigma} + \sum_{i=1}^{2} n_i \underline{\mu}_i \underline{\mu}_i^T \right] \hat{\underline{w}} = N(\hat{\underline{\mu}}_2 - \hat{\underline{\mu}}_1)$$

$$\Rightarrow \left[(N-2)\hat{\Sigma} + \sum_{i=1}^{2} n_i \underline{\mu}_i \underline{\mu}_i^T - N\underline{\mu}\underline{\mu}^T \right] \hat{\underline{w}} = N(\hat{\underline{\mu}}_2 - \hat{\underline{\mu}}_1)$$

$$\left[(N-2)\hat{\Sigma} + \sum_{i=1}^{2} n_i \underline{\mu}_i \underline{\mu}_i^T - \frac{1}{N} \left(\sum_{i=1}^{2} n_i \underline{\mu}_i \right) \left(\sum_{j=1}^{2} n_j \underline{\mu}_j \right)^T \right] \hat{\underline{w}} = N(\hat{\underline{\mu}}_2 - \hat{\underline{\mu}}_1)$$

$$\left((N-2)\hat{\Sigma} + \frac{n_1 n_2}{N} \hat{\Sigma}_B \right) \underline{w} = N(\hat{\underline{\mu}}_2 - \hat{\underline{\mu}}_1)$$

where

$$\hat{\Sigma}_B = (\hat{\underline{\mu}}_2 - \hat{\underline{\mu}}_1)\ (\hat{\underline{\mu}}_2 - \hat{\underline{\mu}}_1)^T$$

Copyright ©2018 by K. Pattipati

Problem 1

$c)\ \left(\hat{\Sigma}_B\right)\hat{\underline{w}} = (\hat{\underline{\mu}}_2 - \hat{\underline{\mu}}_1)\,(\hat{\underline{\mu}}_2 - \hat{\underline{\mu}}_1)^T\,\hat{\underline{w}} \propto\ (\hat{\underline{\mu}}_2 - \hat{\underline{\mu}}_1)$

$so,\ \hat{\underline{w}} \propto\ \ \hat{\Sigma}^{-1}(\hat{\underline{\mu}}_2 - \hat{\underline{\mu}}_1)$

$d)\ for\ any\ distinct\ coding,\ z_1 \neq z_2$

$$RHS: \left(\sum_{i=1}^{2} z_i \left(\sum_{k\in C_i} \underline{x}_{ak}\right)\right) = \sum_{i=1}^{2} z_i n_i \mu_{ai} = \begin{bmatrix} \sum_{i=1}^{2} z_i n_i \\ \sum_{i=1}^{2} z_i n_i \underline{\mu}_i \end{bmatrix}$$

$$For\ w_0\ term: N\hat{w}_0 = -\hat{\underline{w}}^T \sum_{i=1}^{2}\sum_{k\in C_i} \underline{x}_k + \sum_{i=1}^{2} z_i n_i = -\hat{\underline{w}}^T \sum_{i=1}^{2} n_i \underline{\mu}_i + \frac{1}{N}\sum_{i=1}^{2} z_i n_i \Rightarrow \hat{w}_0 = \sum_{i=1}^{2} \hat{\pi}_i (z_i - \hat{\underline{w}}^T \underline{\mu}_i) = \sum_{i=1}^{2} \hat{\pi}_i z_i - \hat{\underline{w}}^T \underline{\mu}$$

$$For\ \underline{w}\ term: N\underline{\mu}\hat{w}_0 + \left[(N-2)\hat{\Sigma} + \sum_{i=1}^{2} n_i \underline{\mu}_i \underline{\mu}_i^T\right]\hat{\underline{w}} = \sum_{i=1}^{2} z_i n_i \underline{\mu}_i$$

$$\Rightarrow \left[(N-2)\hat{\Sigma} + \sum_{i=1}^{2} n_i \underline{\mu}_i \underline{\mu}_i^T - N\underline{\mu}\underline{\mu}^T\right]\hat{\underline{w}} = \sum_{i=1}^{2} z_i n_i \underline{\mu}_i - N\left(\sum_{i=1}^{2} \hat{\pi}_i z_i\right)\underline{\mu} = \sum_{i=1}^{2} n_i \underline{\mu}_i z_i - \frac{1}{N}\left(\sum_{i=1}^{2} n_i z_i\right)\left(\sum_{j=1}^{2} n_j \underline{\mu}_j\right)$$

$$\left[(N-2)\hat{\Sigma} + \sum_{i=1}^{2} n_i \underline{\mu}_i \underline{\mu}_i^T - \frac{1}{N}\left(\sum_{i=1}^{2} n_i \underline{\mu}_i\right)\left(\sum_{j=1}^{2} n_j \underline{\mu}_j\right)^T\right]\hat{\underline{w}} = \frac{n_1 n_2}{N}(z_2 - z_1)(\hat{\underline{\mu}}_2 - \hat{\underline{\mu}}_1)$$

$$\left((N-2)\hat{\Sigma} + \frac{n_1 n_2}{N}\hat{\Sigma}_B\right)\underline{w} == \frac{n_1 n_2}{N}(z_2 - z_1)(\hat{\underline{\mu}}_2 - \hat{\underline{\mu}}_1)$$

where

$$\hat{\Sigma}_B = \ (\hat{\underline{\mu}}_2 - \hat{\underline{\mu}}_1)\,(\hat{\underline{\mu}}_2 - \hat{\underline{\mu}}_1)^T$$

Evidently, when $z_2 = \frac{N}{n_2}$ and $z_1 = -\frac{N}{n_1}$, $\frac{n_1 n_2}{N}(z_2 - z_1) = N$

$e)$ Decision rule: $\sum_{i=1}^{2} \hat{\pi}_i z_i + \hat{\underline{w}}^T(\underline{x}_i - \underline{\mu}) \underset{class\ 1}{\overset{class\ 2}{>}} 0 \Rightarrow Need\ \sum_{i=1}^{2} \hat{\pi}_i z_i = 0$

$\Rightarrow \sum_{i=1}^{2} n_i z_i = 0 \Rightarrow possible\ codings: z_1 = -\frac{N}{n_1}, z_2 = \frac{N}{n_2}$.... no need for equal number of samples

$z_1 = -1, z_2 = 1\ \&\ n_1 = n_2\ etc.$

Copyright ©2018 by K. Pattipati

Problem 3

$z \sim N(z; \mu, \sigma^2)$

$z = \mu + \sigma t$ where $t \sim N(0,1)$

$$E\left[\, z \mid z \geq c\right] = E\left[\, \mu + \sigma t \mid t \geq \frac{c-\mu}{\sigma}\right]$$

$$\text{Recall } \phi(t \mid t \geq \frac{c-\mu}{\sigma}) = \frac{\phi(t)}{1-\Phi(\frac{c-\mu}{\sigma})}$$

$$so, E\left[\, \mu + \sigma t \mid t \geq \frac{c-\mu}{\sigma}\right] = \mu + \frac{1}{1-\Phi(\frac{c-\mu}{\sigma})} \frac{\sigma}{\sqrt{2\pi}} \int_{\frac{c-\mu}{\sigma}}^{\infty} t e^{-t^2/2} dt$$

$$= \mu - \frac{\sigma}{1-\Phi(\frac{c-\mu}{\sigma})} \int_{\frac{c-\mu}{\sigma}}^{\infty} \phi'(t) dt = \mu - \frac{\sigma}{1-\Phi(\frac{c-\mu}{\sigma})} .\phi(t) \Big|_{\frac{c-\mu}{\sigma}}^{\infty} = \mu + \frac{\sigma\phi(\frac{c-\mu}{\sigma})}{1-\Phi(\frac{c-\mu}{\sigma})} = \mu + \sigma H\left(\frac{c-\mu}{\sigma}\right)$$

$$where\ H(u) = \frac{\phi(u)}{1-\Phi(u)}$$

$$E\left[\, z^2 \mid z \geq c\right] = E\left[\left(\mu + \sigma t\right)^2 \mid t \geq \frac{c-\mu}{\sigma}\right] = \mu^2 + 2\sigma\mu H\left(\frac{c-\mu}{\sigma}\right) + \sigma^2 E[t^2 \mid t \geq \frac{c-\mu}{\sigma}]$$

$$E[t^2 \mid t \geq \frac{c-\mu}{\sigma}] = \left[\frac{1}{1-\Phi(\frac{c-\mu}{\sigma})}\right]\left[\int_{\frac{c-\mu}{\sigma}}^{\infty} t^2 \phi(t) dt\right]$$

$$\int_{\frac{c-\mu}{\sigma}}^{\infty} t^2 \phi(t) dt = -\int_{\frac{c-\mu}{\sigma}}^{\infty} t\phi'(t) dt = \left(\frac{c-\mu}{\sigma}\right)\phi(\frac{c-\mu}{\sigma}) + \int_{\frac{c-\mu}{\sigma}}^{\infty} \phi(t) dt = \left(\frac{c-\mu}{\sigma}\right)\phi(\frac{c-\mu}{\sigma}) + 1 - \Phi(\frac{c-\mu}{\sigma})$$

$$so, E\left[\, z^2 \mid z \geq c\right] = \mu^2 + 2\sigma\mu H\left(\frac{c-\mu}{\sigma}\right) + \sigma(c-\mu) H\left(\frac{c-\mu}{\sigma}\right) + \sigma^2$$

$$= \mu^2 + \sigma^2 + \sigma(c+\mu) H\left(\frac{c-\mu}{\sigma}\right)$$

Copyright ©2018 by K. Pattipati

Problem 4:

$$\underline{w}^{(n+1)} = \underline{w}^{(n)} + \eta(z^n - \underline{w}^{(n)T}\underline{x}^n)\underline{x}^n = \underline{w}^{(n)} + \eta(\underbrace{z^n - \underline{w}^{*T}\underline{x}^n}_{e^{*n}} - (\underline{w}^{(n)} - \underline{w}^*)^T\underline{x}^n)\underline{x}^n$$

$$\underline{v}^{(n+1)} = \left[I - \eta\underline{x}^n\underline{x}^{nT}\right]\underline{v}^{(n)} + \eta e^{*n}\underline{x}^n; \underline{v}^{(n)} = \underline{w}^{(n)} - \underline{w}^*$$

Let $\Sigma_n = E\{\underline{v}^{(n)}\underline{v}^{(n)T}\}; R_x = E[\underline{x}^n\,\underline{x}^{nT}]$ **~ Correlation matrix of data;** $E[\left(e^{*n}\right)^2] = \sigma_e^2$

$$\Sigma_{n+1} = E\{\underline{v}^{(n+1)}\underline{v}^{(n+1)T}\} = E\{\left[I - \eta\underline{x}^n\underline{x}^{nT}\right]\underline{v}^{(n)}\underline{v}^{(n)T}\left[I - \eta\underline{x}^n\underline{x}^{nT}\right]^T\} + \eta E\{e^{*n}\left[I - \eta\underline{x}^n\underline{x}^{nT}\right]\underline{v}^{(n)}\underline{x}^{nT}\} +$$

$$+\eta E\{e^{*n}\underline{x}^n\underline{v}^{(n)T}\left[I - \eta\underline{x}^n\underline{x}^{nT}\right]^T\} + \eta^2 E\{e^{*n2}\underline{x}^n\underline{x}^{nT}\}$$

Using LMS assumption and the orthogonality of error and the weight estimate, we have

$$\Sigma_{n+1} = \Sigma_n - \eta R_x\Sigma_n - \eta\Sigma_n R_x + \eta^2 E\{\underline{x}^n\underline{x}^{nT}\underline{v}^{(n)}\underline{v}^{(n)T}\underline{x}^n\underline{x}^{nT}\} + \eta^2 E\{\left(e^{*n}\right)^2\underline{x}^n\underline{x}^{nT}\}$$

$$= \Sigma_n - \eta R_x\Sigma_n - \eta\Sigma_n R_x + 2\eta^2 R_x\Sigma_n R_x + \eta^2 R_x tr\{\Sigma_n R_x\} + \eta^2\sigma_e^2 R_x$$

Note : *Use* $E\{x_1x_2x_3x_4\} = E\{x_1x_2\}E\{x_3x_4\} + E\{x_1x_3\}E\{x_2x_4\} + E\{x_1x_4\}E\{x_2x_3\}$

Vector case : $E\{\underline{x}^n\underline{x}^{nT}\underline{v}^{(n)}\underline{v}^{(n)T}\underline{x}^n\underline{x}^{nT}\} = E\{\underline{x}^n\underline{x}^{nT}\}E\{\underline{v}^{(n)}\underline{v}^{(n)T}\underline{x}^n\underline{x}^{nT}\} + E\{\underline{x}^n\underline{x}^{nT}\underline{v}^{(n)}\underline{v}^{(n)T}\}E\{\underline{x}^n\underline{x}^{nT}\}$

$$+E\{\underline{x}^{nT}\underline{v}^{(n)}\underline{v}^{(n)T}\underline{x}^n\}E\{\underline{x}^n\underline{x}^{nT}\}$$

Let $R_x = Q\Lambda_x Q^T$ **and let** $\hat{\Sigma}_{n+1} = Q^T\Sigma_{n+1}Q$

$$Q^T\Sigma_{n+1}Q = Q^T\Sigma_n Q - \eta Q^T R_x QQ^T\Sigma_n Q - \eta Q^T\Sigma_n QQ^T R_x Q + 2\eta^2 Q^T R_x QQ^T\Sigma_n QQ^T R_x Q$$

$$+\eta^2 R_x tr\{Q^T\Sigma_n QQ^T R_x Q\} + \eta^2\sigma_e^2 Q^T R_x Q$$

$$\Rightarrow \hat{\Sigma}_{n+1} = \hat{\Sigma}_n - \eta\Lambda_x\hat{\Sigma}_n - \eta\hat{\Sigma}_n\Lambda_x + 2\eta^2\Lambda_x\hat{\Sigma}_n\Lambda_x + \eta^2\Lambda_x tr\{\hat{\Sigma}_n\Lambda_x\} + \eta^2\sigma_e^2\Lambda_x$$

Copyright ©2018 by K. Pattipati

Problem 4:

$$\hat{\Sigma}_{n+1} = \hat{\Sigma}_n - \eta\Lambda_x\hat{\Sigma}_n - \eta\hat{\Sigma}_n\Lambda_x + 2\eta^2\Lambda_x\hat{\Sigma}_n\Lambda_x + \eta^2\Lambda_x tr\{\hat{\Sigma}_n\Lambda_x\} + \eta^2\sigma_e^2\Lambda_x$$

Now consider the diagonal elements of $\hat{\Sigma}_{n+1}$ and represent them as a vector $\underline{s}_{n+1}$

$$\underline{s}_{n+1} = (I_{p+1} - 2\eta\Lambda_x + 2\eta^2\Lambda_x^2 + \eta^2\underline{\lambda}\underline{\lambda}^T)\underline{s}_n + \eta^2\sigma_e^2\underline{\lambda}$$

where $\underline{\lambda} = \begin{bmatrix}\lambda_1 & \lambda_2 & . & \lambda_{p+1}\end{bmatrix}^T$

Need eigen values of $(I_{p+1} - 2\eta\Lambda_x + 2\eta^2\Lambda_x^2 + \eta^2\underline{\lambda}\underline{\lambda}^T)$ *inside unit circle.*

$$\Rightarrow |\lambda_i\{\underbrace{(I_{p+1} - \eta\Lambda_x)^2 + \eta^2(\Lambda_x^2 + \underline{\lambda}\underline{\lambda}^T)}_{\Phi}\}| < 1 \forall i = 1,2,..,p+1$$

$$\Phi_{ij} = \begin{cases}(1-\eta\lambda_i)^2 + 2\eta^2\lambda_i^2; i = j \\ \eta^2\lambda_i\lambda_j; i \neq j\end{cases}$$

$$\| \Phi \|_\infty < 1 \Rightarrow \max_i \left\{(1-\eta\lambda_i)^2 + \eta^2\lambda_i\sum_{j=1}^{p+1}\lambda_j\right\} < 1$$

$$\Rightarrow -2\eta\lambda_i + 2\eta^2\lambda_i^2 + \eta^2\lambda_i Tr(R_x) < 0 \ \forall i$$

$$\Rightarrow 2\eta\lambda_i + \eta Tr(R_x) < 2 \ \& \ \eta > 0$$

$$\Rightarrow 0 < \eta < \min_i \frac{2}{tr(R_x) + 2\lambda_i} < \frac{2}{tr(R_x) + 2\lambda_{\max}} < \frac{2}{tr(R_x)}$$ **(in practice:** $0 < \eta < \frac{2}{tr(R_x) + 2\lambda_{\max}}$**)**

Let $f(\lambda) = -2\eta\lambda + 2\eta^2\lambda^2 + \eta^2\lambda Tr(R_x)$

Max of f *at*: $\lambda^* = \frac{2 - \eta Tr(R_x)}{4\eta}$

$$f(\lambda^*) = -\frac{\eta}{2}(1 - \frac{\eta}{2}Tr(R_x))^2 \Rightarrow 0 < \eta < \frac{2}{Tr(R_x)}$$

Sum min g *over all* i

$$2\eta Tr(R_x) + (p+1)\eta Tr(R_x) < 2(p+1) \ \& \ \eta > 0 \Rightarrow 0 < \eta < \frac{2}{Tr(R_x)[1 + \frac{2}{p+1}]} < \frac{2}{tr(R_x)}$$

Copyright ©2018 by K. Pattipati

Problem 5

Consider a general regularized least squares regression problem.

$$J = \frac{1}{N} \| \underline{z} - X \underline{w} \|_2^2 + \frac{\lambda}{N} \underline{w}^T \Gamma^T \Gamma \underline{w}; \underline{z} \in R^N; X \in R^{N \mathbf{x} (p+1)}$$

where $\underline{z} = X \underline{w} + \underline{v}; v_n \sim N(0, \sigma^2) \; \forall n = 1, 2, .., N$

Let $\hat{\underline{w}}(0, \Gamma) = (X^T X)^{-1} X^T \underline{z}$, *least squares solution when* $\lambda = 0$.

$$(a) \nabla_{\underline{w}} J = \frac{2}{N} [-X^T (\underline{z} - X \underline{w}) + \lambda \Gamma^T \Gamma \underline{w}] = \underline{0}$$

$$\Rightarrow \hat{\underline{w}}(\lambda, \Gamma) = (X^T X + \lambda \Gamma^T \Gamma)^{-1} X^T \underline{z} = (X^T X + \lambda \Gamma^T \Gamma)^{-1} X^T (X \underline{w} + \underline{v})$$

$$= (X^T X + \lambda \Gamma^T \Gamma)^{-1} [X^T X \underline{w} + X^T \underline{v})$$

$$= (X^T X + \lambda \Gamma^T \Gamma)^{-1} [(X^T X + \lambda \Gamma^T \Gamma) \underline{w} - \lambda \Gamma^T \Gamma \underline{w} + X^T \underline{v})$$

$$= \underline{w} - \lambda (X^T X + \lambda \Gamma^T \Gamma)^{-1} \Gamma^T \Gamma \underline{w} + (X^T X + \lambda \Gamma^T \Gamma)^{-1} X^T \underline{v}$$

$$\Gamma = I_{p+1} \Rightarrow \hat{\underline{w}}(\lambda, I_{p+1}) = \underline{w} - \lambda (X^T X + \lambda I_{p+1})^{-1} \underline{w} + (X^T X + \lambda I_{p+1})^{-1} X^T \underline{v}$$

$$\Gamma = X \Rightarrow \hat{\underline{w}}(\lambda, X) = \frac{\underline{w}}{1+\lambda} + \frac{1}{1+\lambda} (X^T X)^{-1} X^T \underline{v} = \frac{\hat{\underline{w}}(0, X)}{1+\lambda} = \frac{\hat{\underline{w}}(0, \Gamma)}{1+\lambda} = \frac{\hat{\underline{w}}(0)}{1+\lambda}$$

It scales the least squares solution by $1/(1+\lambda)$

Copyright ©2018 by K. Pattipati

Problem 5

(b)

$$\underline{b} = \underline{w} - E\{\underline{\hat{w}}(\lambda,\Gamma)\} = \underline{w} - \underline{w} + \lambda(X^T X + \lambda\Gamma^T\Gamma)^{-1}\Gamma^T\Gamma\,\underline{w} - E\{(X^T X + \lambda\Gamma^T\Gamma)^{-1} X^T \underline{v}\}$$

$$= \lambda(X^T X + \lambda\Gamma^T\Gamma)^{-1}\Gamma^T\Gamma\,\underline{w}$$

$$\Gamma = I_{p+1} \Rightarrow \underline{b} = \lambda(X^T X + \lambda I_{p+1})^{-1}\underline{w}$$

$$\Gamma = X \Rightarrow \underline{b} = \frac{\lambda}{1+\lambda}\underline{w}$$

(c)

$$r = z - \hat{z} = \underline{x}^T\underline{w} + v - \underline{x}^T\underline{\hat{w}}(\lambda,\Gamma)$$

$$= \lambda\underline{x}^T(X^T X + \lambda\Gamma^T\Gamma)^{-1}\Gamma^T\Gamma\,\underline{w} + v - \underline{x}^T(X^T X + \lambda\Gamma^T\Gamma)^{-1} X^T\underline{v}$$

$$\Gamma = I_{p+1} \Rightarrow \underline{r} = \lambda\underline{x}^T(X^T X + \lambda I_{p+1})^{-1}\underline{w} + v - \underline{x}^T(X^T X + \lambda I_{p+1})^{-1} X^T\underline{v}$$

$$\Gamma = X \Rightarrow \underline{r} = \frac{\lambda}{1+\lambda}(\underline{x}^T\underline{w} + v) + \frac{1}{1+\lambda}[v - \underline{x}^T(X^T X)^{-1} X^T\underline{v}]$$

Copyright ©2018 by K. Pattipati

Problem 5

(d)

$$r = \lambda \underline{x}^T (X^T X + \lambda \Gamma^T \Gamma)^{-1} \Gamma^T \Gamma \underline{w} + v - \underline{x}^T (X^T X + \lambda \Gamma^T \Gamma)^{-1} X^T \underline{v}$$

$$bias^2(\lambda, \Gamma) = E(r)^2 \approx \lambda^2 \underline{w}^T \Gamma^T \Gamma (N\Sigma_x + \lambda \Gamma^T \Gamma)^{-1} \Sigma_x (N\Sigma_x + \lambda \Gamma^T \Gamma)^{-1} \Gamma^T \Gamma \underline{w}$$

$$where\ \Sigma_x = E_{\underline{x}}(\underline{x}\underline{x}^T) \approx \frac{X^T X}{N}$$

$$\Gamma = I_{p+1} \Rightarrow bias^2(\lambda, I_{p+1}) = \lambda^2 \underline{w}^T (N\Sigma_x + \lambda I_{p+1})^{-1} \Sigma_x (N\Sigma_x + \lambda I_{p+1})^{-1} \underline{w}$$

$$If\ \Sigma_x = I_{p+1}\ also, bias^2(\lambda, I_{p+1}) = \left(\frac{\lambda}{N+\lambda}\right)^2 \underline{w}^T \underline{w}$$

$$\Gamma = X \Rightarrow bias^2(\lambda, X) = \left(\frac{\lambda}{\lambda+1}\right)^2 \underline{w}^T \Sigma_x \underline{w}$$

$$If\ \Sigma_x = I_{p+1}\ also, bias^2(\lambda, X) = \left(\frac{\lambda}{\lambda+1}\right)^2 \underline{w}^T \underline{w}$$

Copyright ©2018 by K. Pattipati

Problem 5

(e)

$$\text{var}(\lambda,\Gamma) = E\{[r - E(r)]^2\} =$$

$$\sigma^2 + [E_{\underline{x},\underline{v}}\{\underline{x}^T (X^T X + \lambda\Gamma^T\Gamma)^{-1} X^T \underline{v}\underline{v}^T X (X^T X + \lambda\Gamma^T\Gamma)^{-1}\underline{x}]$$

$$\approx \sigma^2(1 + N.tr([\Sigma_x(N\Sigma_x + \lambda\Gamma\Gamma^T)^{-1}]^2)$$

$$where\ \Sigma_x = E_{\underline{x}}(\underline{x}\underline{x}^T) \approx \frac{X^T X}{N}$$

$$\Gamma = I_{p+1} \Rightarrow \text{var}(\lambda, I_{p+1}) = \sigma^2(1 + N.tr([\Sigma_x(N\Sigma_x + \lambda I_{p+1})^{-1}]^2)$$

$$If\ \Sigma_x = I_{p+1}\ also, \text{var}(\lambda, X) = \sigma^2[1 + \frac{N(p+1)}{(N+\lambda)^2}]$$

$$\Gamma = X \Rightarrow \text{var}(\lambda, X) = \sigma^2[1 + \frac{(p+1)}{N(1+\lambda)^2}] ....independent\ of\ \Sigma_x$$

Copyright ©2018 by K. Pattipati

Problem 5

(f)

$\Gamma = I_{p+1}; \Sigma_x = I_{p+1}$

$$MSE = bias^2(\lambda, I_{p+1}) + \text{var}(\lambda, I_{p+1}) = \left(\frac{\lambda}{N+\lambda}\right)^2 \underline{w}^T \underline{w} + \sigma^2[1 + \frac{N(p+1)}{(N+\lambda)^2}]$$

$$\frac{dMSE}{d\lambda} = \left(\frac{2\lambda}{(N+\lambda)^2} - \frac{2\lambda^2}{(N+\lambda)^3}\right)\underline{w}^T \underline{w} - \frac{2\sigma^2 N(p+1)}{(N+\lambda)^3} = 0$$

$$\Rightarrow \frac{2N\lambda}{(N+\lambda)^3}\underline{w}^T \underline{w} - \frac{2\sigma^2 N(p+1)}{(N+\lambda)^3} = 0 \Rightarrow \lambda^* = \frac{\sigma^2(p+1)}{\underline{w}^T \underline{w}}$$

$$Note: \frac{dMSE}{d\lambda} = 2N\underline{w}^T \underline{w}\left(\frac{\lambda - \lambda^*}{(N+\lambda)^3}\right)$$

$$\Rightarrow \frac{d^2 MSE}{d\lambda^2} = 2N\underline{w}^T \underline{w}\left(\frac{1}{(N+\lambda)^3} - \frac{3(\lambda - \lambda^*)}{(N+\lambda)^4}\right) = \frac{2N\underline{w}^T \underline{w}}{(N+\lambda^*)^3} > 0 \; at \; \lambda = \lambda^*$$

Copyright ©2018 by K. Pattipati

Problem 5

(f)

$\Gamma = X; \Sigma_x = I_{p+1}$

$$MSE = bias^2(\lambda, X) + \text{var}(\lambda, X) = \left(\frac{\lambda}{\lambda+1}\right)^2 \underline{w}^T \underline{w} + \sigma^2[1 + \frac{(p+1)}{N(1+\lambda)^2}]$$

$$\frac{dMSE}{d\lambda} = \left(\frac{2\lambda}{(\lambda+1)^2} - \frac{2\lambda^2}{(\lambda+1)^3}\right)\underline{w}^T \underline{w} - \frac{2\sigma^2(p+1)}{N(1+\lambda)^3} = 0$$

$$\Rightarrow \frac{2\lambda}{(1+\lambda)^3}\underline{w}^T \underline{w} - \frac{2\sigma^2(p+1)}{N(1+\lambda)^3} = 0 \Rightarrow \lambda^* = \frac{\sigma^2(p+1)}{N.\underline{w}^T \underline{w}}$$

$$Note: \frac{dMSE}{d\lambda} = 2\underline{w}^T \underline{w}\left(\frac{\lambda - \lambda^*}{(1+\lambda)^3}\right)$$

$$\Rightarrow \frac{d^2MSE}{d\lambda^2} = 2\underline{w}^T \underline{w}\left(\frac{1}{(1+\lambda)^3} - \frac{3(\lambda-\lambda^*)}{(1+\lambda)^4}\right) = \frac{2\underline{w}^T \underline{w}}{(1+\lambda^*)^3} > 0 \; at \; \lambda = \lambda^*$$

Copyright ©2018 by K. Pattipati

Problem 6:

$$C_1 : \left\{ \underline{x}^1 = \begin{bmatrix} 1 \\ 5 \end{bmatrix}; \underline{x}^2 = \begin{bmatrix} -2 \\ -4 \end{bmatrix} \right\}$$

$$C_2 : \left\{ \underline{x}^3 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}; \underline{x}^4 = \begin{bmatrix} -1 \\ 5 \end{bmatrix} \right\}$$

$$\underline{\Phi}(\underline{x}) = [1 \sqrt{2}x_1 \sqrt{2}x_2 \sqrt{2}x_1x_2 \; x_1^2 \; x_2^2 \;]^T$$

$$\underline{\Phi}_1 = [1 \sqrt{2} \; \sqrt{2}\text{x}5 \sqrt{2}\text{x}5 \; 1 \; 25]^T ; \underline{\Phi}_2 = [1 -\sqrt{2} \; \text{x}2 -\sqrt{2}\text{x}4 \sqrt{2}\text{x}8 \; 4 \; 16]^T ; z^1 = z^2 = -1$$

$$\underline{\Phi}_3 = [1 \sqrt{2} \; \text{x}2 \sqrt{2}\text{x}3 \sqrt{2}\text{x}6 \; 4 \; 9]^T ; \underline{\Phi}_2 = [1 -\sqrt{2} \; \sqrt{2}\text{x}5 \; -\sqrt{2}\text{x}5 \; 1 \; 25]^T ; z^3 = z^4 = 1$$

$$Dual : \max_{\underline{\lambda}} q(\underline{\lambda}) = \sum_{i=1}^{4} \lambda_i - \frac{1}{2} \sum_{i=1}^{4} \sum_{j=1}^{4} \lambda_i \lambda_j z^i z^j \underline{\Phi}(\underline{x}^i)^T \underline{\Phi}(\underline{x}^j)$$

$$subject\ to : \sum_{i=1}^{N} \lambda_i z^i = 0 \text{ and } 0 \le \lambda_i \Rightarrow \lambda_1 + \lambda_2 = \lambda_3 + \lambda_4$$

$$\underline{\lambda} = \begin{bmatrix} 0.0154 & 0.0067 & 0.0126 & 0.0095 \end{bmatrix}^T$$

$$\underline{w}^T = [0.0000 \quad 0.0192 \quad 0.0494 \quad -0.1453 \quad 0.0178 \quad -0.1415]$$

$$w_0 = -3.1713$$

$$g(\underline{x}) = \underline{w}^T \underline{\Phi}(\underline{x}) - w_0 = 0.0272x_1 + 0.0699x_2 - 0.2037x_1x_2 + 0.0178x_1^2 - 0.1415x_2^2 + 3.1713 = 0$$

Support vectors: All four are support vectors!

$$\frac{1}{\rho} = \|\text{w}\|_2 = \frac{1}{\text{margin}} = \sqrt{\sum_{i=1}^{N} \lambda_i} = 0.2104$$

$$m \arg in = \rho = 4.7529$$

Copyright ©2018 by K. Pattipati

- Problem 7:

$$N = 11; p = 2$$

$$\underline{\hat{w}}_{ML} = (X^T X)^{-1} X^T \underline{z} = \begin{bmatrix} -3.2564 \\ 0.0427 \end{bmatrix}$$

$$\hat{\sigma}^2 = \frac{1}{N-p} \| \underline{z} - X \underline{\hat{w}} \|^2 = 0.0170$$

$$\Sigma_{ML} = \hat{\sigma}^2 (X^T X)^{-1} = \begin{bmatrix} 0.1323 & -0.0012 \\ -0.0012 & 0.00001142 \end{bmatrix}$$

$$\underline{w}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}; \Sigma_0 = \begin{bmatrix} 10^{10} & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Sigma_{MAP} = \left[\Sigma_{ML}^{-1} + \Sigma_0^{-1} \right]^{-1} = \begin{bmatrix} 0.132323933270331 & -0.001222250333231 \\ -0.001222250333231 & 0.000011422900311 \end{bmatrix}$$

$$\underline{\hat{w}}_{MAP} = \Sigma_{MAP} \left[\Sigma_{ML}^{-1} \underline{\hat{w}}_{ML} + \Sigma_0^{-1} \underline{w}_0 \right] = \begin{bmatrix} -3.256376353285603 \\ 0.042650925986531 \end{bmatrix}$$

$$w_1 \sim N(w_1; 0.042650925986531, 0.000011422900311)$$

When the prior is non-informative, ML and MAP give virtually the same estimates

Copyright ©2018 by K. Pattipati

Problem 8

$$J = -\ln p(\theta_1, \theta_2 \mid D) = N\theta_2 + \frac{e^{-2\theta_2}}{2}\left[Ns^2 + N(\bar{z} - \theta_1)^2 \right]$$

where $\bar{z}$ is the sample mean and s^2 is the sample variance

$$\nabla_{\underline{\theta}} J = \begin{bmatrix} \frac{\partial J}{\partial \theta_1} \\ \frac{\partial J}{\partial \theta_2} \end{bmatrix} = \begin{bmatrix} -Ne^{-2\theta_2}(\bar{z} - \theta_1) \\ N - \left[Ns^2 + N(\bar{z} - \theta_1)^2 \right] e^{-2\theta_2} \end{bmatrix}$$

$$\nabla_{\underline{\theta}} J \mid_{\underline{\hat{\theta}}_{MAP}} = \underline{0} \Rightarrow \hat{\theta}_{1,MAP} = \bar{z}; \hat{\theta}_{2,MAP} = \ln s$$

$$\nabla^2_{\underline{\theta}} J = \begin{bmatrix} \frac{\partial^2 J}{\partial \theta_1^2} & \frac{\partial^2 J}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 J}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 J}{\partial \theta_2^2} \end{bmatrix} = \begin{bmatrix} Ne^{-2\theta_2} & 2Ne^{-2\theta_2}(\bar{z} - \theta_1) \\ 2Ne^{-2\theta_2}(\bar{z} - \theta_1) & 2Ns^2 e^{-2\theta_2} \end{bmatrix}$$

$$\nabla^2_{\underline{\theta}} J \mid_{\underline{\hat{\theta}}_{MAP}} = H = \begin{bmatrix} \frac{N}{s^2} & 0 \\ 0 & 2N \end{bmatrix} \Rightarrow \Sigma = \begin{bmatrix} \frac{s^2}{N} & 0 \\ 0 & \frac{1}{2N} \end{bmatrix}$$

$$p(\theta_1, \theta_2 \mid D) = N(\theta_1; \bar{z}, \frac{s^2}{N}) N(\theta_2; \ln s, \frac{1}{2N}) = \frac{1}{\sqrt{2}\pi} \frac{N}{s} e^{-N\left[\frac{(\theta_1 - \bar{z})^2}{2s^2} + (\theta_2 - \ln s)^2 \right]}$$

Copyright ©2018 by K. Pattipati

Problem 9:

$$P(\underline{v},\underline{h}) = \frac{1}{Z}\exp(\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}h_i v_j + \sum_{i=1}^{m} b_i h_i + \sum_{j=1}^{n} c_j v_j)$$

$$where\ Z = \sum_{\underline{v}}\sum_{\underline{h}} \exp(\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}h_i v_j + \sum_{i=1}^{m} b_i h_i + \sum_{j=1}^{n} c_j v_j)$$

$$P(\underline{h}\,|\,\underline{v}) = \frac{P(\underline{h},\underline{v})}{P(\underline{v})} = \frac{P(\underline{h},\underline{v})}{\sum_{\underline{h}'} P(\underline{h}',\underline{v})} = \frac{\exp(\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}h_i v_j + \sum_{i=1}^{m} b_i h_i + \sum_{j=1}^{n} c_j v_j)}{\sum_{\underline{h}'} \exp(\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}h_i v_j + \sum_{i=1}^{m} b_i h_i + \sum_{j=1}^{n} c_j v_j)}$$

$$= \frac{\exp(\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}h_i v_j + \sum_{i=1}^{m} b_i h_i)}{\sum_{\underline{h}'} \exp(\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}h_i v_j + \sum_{i=1}^{m} b_i h_i)} = \prod_{i=1}^{m} \frac{\exp(\sum_{j=1}^{n} d_{ij}v_j + b_i)h_i}{\left[1+\exp(\sum_{j=1}^{n} d_{ij}v_j + b_i)\right]}; h_i \in \{0,1\}$$

and consequently

$$P(h_i = 1\,|\,\underline{v}) = \frac{\exp(\sum_{j=1}^{n} d_{ij}v_j + b_i)}{\left[1+\exp(\sum_{j=1}^{n} d_{ij}v_j + b_i)\right]} = g(\sum_{j=1}^{n} d_{ij}v_j + b_i)...sigmoid\ \ function$$

Copyright ©2018 by K. Pattipati

Problem 9:

$$So, P(\underline{v}\,|\underline{h}) = \frac{\exp(\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}h_i v_j + \sum_{j=1}^{n} c_j v_j)}{\sum_{\underline{v}'} \exp(\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}h_i v_j + \sum_{j=1}^{n} c_j v_j)} = \prod_{j=1}^{n} \frac{\exp(\sum_{i=1}^{m} d_{ij}h_i + c_j)v_j}{\left[1+\exp(\sum_{i=1}^{m} d_{ij}h_i + c_j)\right]}; v_j \in \{0,1\}$$

and consequently

$$P(v_j = 1 \,|\, \underline{h}) = \frac{\exp(\sum_{i=1}^{m} d_{ij}h_i + c_j)}{\left[1+\exp(\sum_{i=1}^{m} d_{ij}h_i + c_j)\right]} = g(\sum_{i=1}^{m} d_{ij}h_i + c_j)$$

Copyright ©2018 by K. Pattipati

- Problem 10:

Complete Data Likelihood

$$D_c = \left\{ (x^1, \underline{z}^1),\ (x^2, \underline{z}^2)\ (x^3, \underline{z}^3) \right\};\ z \in \{[1\,0],[0\,1]\};\ x^1 = 1;\ x^2 = 10;\ x^3 = 20$$

$$\Rightarrow -\ln p(D_c \mid \underline{\theta}) = \sum_{n=1}^{3}\sum_{j=1}^{2} z_j^n \{-\ln P_j + \frac{1}{2}\ln 2\pi + \ln \sigma_j + \frac{1}{2\sigma_j^2}(x^n - \mu_j)^2\};\ \underline{\theta} = \{P_1, P_2, \sigma_1, \sigma_2, \mu_1, \mu_2\}$$

EMminimizes the expected value of the negative complete data log likelihood:

$$E_{\underline{Z}}\{-\ln p(D_c \mid \underline{\theta})\} = \sum_{n=1}^{3}\sum_{j=1}^{2} \gamma_j^n \{-\ln P_j + \frac{1}{2}\ln 2\pi + \ln \sigma_j + \frac{1}{2\sigma_j^2}(x^n - \mu_j)^2\}$$

$$\Gamma = \begin{bmatrix} 1 & 0 \\ 0.4 & 0.6 \\ 0 & 1 \end{bmatrix};\ \text{Re}\,call\ N_j = \sum_{n=1}^{N} \gamma_j^n;\ j = 1, 2, .., M \Rightarrow N_1 = 1.4;\ N_2 = 1.6$$

M-step: Re-estimate the parameters using the current responsibilities

$$\mu_j^{new} = \frac{1}{N_j}\sum_{n=1}^{N} \gamma_j^n x^n \Rightarrow \mu_1^{new} = \frac{1 + 0.4 * 10}{1.4} = 3.5714;\ \mu_2^{new} = \frac{0.6 * 10 + 20}{1.6} = 16.25$$

Copyright ©2018 by K. Pattipati

- Problem 10:

M-step: Re-estimate the parameters using the current responsibilities

$$\mu_j^{new} = \frac{1}{N_j}\sum_{n=1}^{N}\gamma_j^n x^n \Rightarrow \mu_1^{new} = \frac{1+0.4*10}{1.4} = 3.5714; \mu_2^{new} = \frac{0.6*10+20}{1.6} = 16.25$$

$$\sigma_j^{2,new} = \frac{1}{N_j}\sum_{n=1}^{N}\gamma_j^n (x^n - \mu_j^{new})^2 \Rightarrow \sigma_1^{2,new} = 16.5306; \sigma_2^{2,new} = 23.4375$$

$$P_j^{new} = \frac{N_j}{N} \Rightarrow P_1^{new} = 0.4667; P_2^{new} = 0.5333$$

E-step:

$$\gamma_j^n = \frac{P_j N(\underline{x}^n; \underline{\mu}_j, \Sigma_j)}{\sum_{k=1}^{M} P_k N(\underline{x}^n; \underline{\mu}_k, \Sigma_k)}; j = 1,2,..,M; n = 1,2,..,N$$

$$\Gamma = \begin{bmatrix} 0.9919 & 0.0081 \\ 0.4072 & 0.5928 \\ 0.0004 & 0.9996 \end{bmatrix} \textit{etc.}$$ After 5 iterations: $$\Gamma = \begin{bmatrix} 0.5373 & 0.4627 \\ 0.5373 & 0.4627 \\ 0.5373 & 0.4627 \end{bmatrix}$$;NLL=11.7250

All converge to same mean of 10.333 and variance of 60.222.

Copyright ©2018 by K. Pattipati

- Problem 10:

With $\Gamma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$, it converges to means 1 and 15 and variances of 0 and 25. NLL= -9.6366 setting var=max(0,eps)

With $\Gamma = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$, it converges to means 5.5 and 20 and variances of 20.25 and 0. NLL= -9.8473 setting var=max(0,eps)

Different initializations will give different answers!!!

Copyright ©2018 by K. Pattipati