Solution 5

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

- Problem 1, Bishop, Chapter 7, Problem 7.4 & 7.5, Page 357

Margin: $\frac{1}{\rho} = \|\underline{w}\|_2$

At optimum $L(\underline{w}, b, \underline{a}) = q(\underline{a})$

$$optimum\ L(\underline{w}^*, b^*, \underline{a}^*) = \frac{1}{2}\underline{w}^{*T}\underline{w}^*$$

$$optimum\ q(\underline{a}^*) = \sum_{n=1}^{N} a_n^* - \frac{1}{2}\underline{w}^{*T}\underline{w}^*$$

$$\Rightarrow \underline{w}^{*T}\underline{w}^* = \sum_{n=1}^{N} a_n^* = \frac{1}{\rho^2} = 2q(\underline{a}^*)$$

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

$$\nabla_{\underline{w}} L = \underline{0} \Rightarrow \underline{w}^* = \sum_{n=1}^{N} a_n t_n \underline{\phi}(\underline{x}_n)$$

$$\nabla_b L = 0 \Rightarrow \sum_{n=1}^{N} a_n t_n = 0$$

$$q(\underline{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\underline{w}^{*T}\underline{w}^*$$

$$s.t. \sum_{n=1}^{N} a_n t_n = 0; a_n \geq 0; n = 1, 2, .., N$$

At optimum, primal = dual $\Rightarrow \sum_{n=1}^{N} a_n^* - \frac{1}{2}\underline{w}^{*T}\underline{w}^* = \frac{1}{2}\underline{w}^{*T}\underline{w}^*$

$$\Rightarrow \underline{w}^{*T}\underline{w}^* = \sum_{n=1}^{N} a_n^* = \frac{1}{\rho^2} = 2q(\underline{a}^*)$$

Problem 2, Problem 7.19, pp. 358.

$$p(t \mid y) = \sigma(y)^t [1-\sigma(y)]^{1-t}; \sigma(y) = [1+\exp(-y)]^{-1} = \hat{t}; y(\underline{x}) = \underline{w}^T \underline{\phi}(\underline{x})$$

$$\ln p(\underline{w} \mid \underline{t}^N, \underline{X}^N) = \ln p(\underline{t}^N \mid \underline{w}, \underline{X}^N) + \ln p(\underline{w}) + const.$$

$$p(\underline{w}) = N(\underline{0}, A^{-1}); A = Diag(\alpha_i) \text{...Information Matrix needs to be learned}$$

$$p(\underline{t}^N \mid \underline{w}, \underline{X}^N) = \prod_{n=1}^{N} \sigma(y^n)^{t^n} [1-\sigma(y^n)]^{1-t^n}$$

$$\begin{aligned} &\nabla_{\underline{w}} \left[t^n \ln \sigma(y^n) + (1-t^n) \ln(1-\sigma(y^n)) \right] \\ &= \left(t^n \left[1-\sigma(y^n) \right] - (1-t^n)\sigma(y^n) \right) \underline{\phi}(\underline{x}^n) \\ &= \left(t^n - \sigma(y^n) \right) \underline{\phi}(\underline{x}^n) \end{aligned}$$

$$\ln p(\underline{w} \mid \underline{t}^N, \underline{X}^N) = \sum_{n=1}^{N} z^n \underbrace{\ln \sigma(y^n) + (1-z^n) \ln(1-\sigma(y^n))}_{\text{negative cross entropy}} - \frac{1}{2} \underline{w}^T A \underline{w} + const.$$

$$\nabla_{\underline{w}} \ln p(\underline{w} \mid \underline{t}^N, \underline{X}^N) = \Phi^T (\underline{t}^N - \hat{\underline{t}}^N) - A\underline{w} \Rightarrow \underline{w}^* = A^{-1}\Phi^T (\underline{t}^N - \hat{\underline{t}}^N) = A^{-1}\Phi^T (\underline{t}^N - \sigma(\Phi \underline{w}^*))$$

$$\nabla^2_{\underline{w}} \ln p(\underline{w} \mid \underline{t}^N, \underline{X}^N) = -(\Phi^T D_N \Phi + A) \Rightarrow H = (\Phi^T D_N \Phi + A); D_N = diag[\hat{t}^n (1-\hat{t}^n)]$$

Problem 2, Problem 7.19, pp. 358.

$$p(\underline{t}^N \mid \underline{X}^N, A) = \int_{\underline{w}} p(\underline{t}^N \mid \underline{w}, \underline{X}^N)\, p(\underline{w} \mid \underline{X}^N, A)\, d\underline{w}$$

$$\approx p(\underline{t}^N \mid \underline{w}^*, \underline{X}^N)\, p(\underline{w}^* \mid A, \underline{X}^N)(2\pi)^{M/2}(\Phi^T D_N^* \Phi + A)^{-1/2}$$

$$\ln p(\underline{t}^N \mid \underline{X}^N, A) = \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|(\Phi^T D_N^* \Phi + A)| - \frac{1}{2}\underline{w}^{*T} A \underline{w}^* + \frac{1}{2}\sum_{i=1}^{M}\ln\alpha_i$$

$$\frac{\partial \ln p(\underline{t}^N \mid \underline{X}^N)}{\partial \alpha_i} = \frac{1}{2\alpha_i} - \frac{w_i^{*2}}{2} - \frac{1}{2}\underline{e}_i^T \Sigma(\underline{\alpha})\underline{e}_i = 0; \Sigma = (\Phi^T D_N^* \Phi + A)^{-1}$$

$$\Rightarrow 1 - \alpha_i w_i^{*2} - \alpha_i \Sigma_{ii}(\underline{\alpha}) = 0 \Rightarrow \alpha_i^{new} = \frac{\gamma_i}{w_i^{*2}}; \gamma_i = 1 - \alpha_i \Sigma_{ii}(\underline{\alpha})$$

- Problem 3, Bishop, Problem 6.22, pp. 322

$$t_n = y(\underline{x}_n) + \varepsilon_n = y_n + \varepsilon_n$$

$$p(t_n \mid y_n) = N(y_n, \sigma_\varepsilon^2)$$

$$y(\underline{x}) = \underline{w}^T \phi(\underline{x}) = \phi^T(\underline{x})\underline{w};\ p(\underline{w}) = N(\underline{0}, \sigma_w^2 I)$$

$$\underline{y}_N = \Phi_N \underline{w} \Rightarrow \underline{y}_N = N(\underline{0}, \sigma_w^2 \Phi_N \Phi_N{}^T)$$

$$\underline{t}_N = \underline{y}_N + \underline{\varepsilon}_N \Rightarrow \underline{t} = N(\underline{0}, \sigma_w^2 \Phi_N \Phi_N{}^T + \sigma_\varepsilon^2 I_N) = N(\underline{0}, C_N)$$

$$\underline{t}_{N+L} = \begin{bmatrix} \underline{t}_N \\ \underline{t}_L \end{bmatrix} = N(\underline{0}, C_{N+L})$$

$$C_{N+L} = \begin{bmatrix} C_N & K \\ K^T & L \end{bmatrix}; K = \sigma_w^2 \Phi_N \Phi_L^T; L = \sigma_w^2 \Phi_L \Phi_L{}^T + \sigma_\varepsilon^2 I_L$$

$$\underline{\hat{t}}_L = K^T C_N^{-1} \underline{t}_N; \Sigma_L = L - K^T C_N^{-1} K$$

For a component j, $\hat{t}_{Lj} = \underline{k}_j^T C_N^{-1} \underline{t}_N; \Sigma_{Ljj} = L_{jj} - \underline{k}_j^T C_N^{-1} \underline{k}_j$

Problem 4, Bishop, Problem 12.2, page 599

$$\tilde{J} = tr(\hat{U}^T S\hat{U}) + tr(H(I_{D-M} - \hat{U}^T\hat{U})$$

$\hat{U}$ is a D x $(D\text{-}M)$ matrix=$[\underline{u}_1, \underline{u}_2, ..., \underline{u}_{D-M}]$; $H = H^T$ is $(D\text{-}M)$x$(D-M)$ symmetric matrix

$$\nabla_{\hat{U}}\tilde{J} = 2S\hat{U} - 2\hat{U}H = 0 \Rightarrow S\underline{u}_i = \sum_{j=1}^{D-M} h_{ji}\underline{u}_j \forall i = 1, 2, .., (D-M)$$

$h_{ii} = \lambda_i$ where λ_i is an eigen value satisfies the necessary condition.

From $H = H\hat{U}^T\hat{U} = \hat{U}^T S\hat{U}$

$Let\ S = V\Lambda V^T \Rightarrow H = \hat{U}^T V\Lambda V^T\hat{U}$

$\underline{u}_i = \underline{v}_i$ for i=1,2,..,$D\text{-}M$ and the orthogonality of $\hat{U}$ and V will make $H = \Lambda_{D-M}$

See lecture notes for a more general proof.

Problem 5, Murphy, Problem 12.6, page 417

This problem comes from the following idea of Fisher's LDA

$$\mu_{yi} = \underline{w}^T \underline{\mu}_i; \sigma_i^2 = \underline{w}^T \Sigma_i \underline{w}; i = 1, 2$$

$$J(\underline{w}) = \frac{\underline{w}^T (\underline{\mu}_1 - \underline{\mu}_2)(\underline{\mu}_1 - \underline{\mu}_2)^T \underline{w}}{\underline{w}^T (\pi_1 \Sigma_1 + \pi_2 \Sigma_2) \underline{w}} = \frac{\underline{w}^T S_B \underline{w}}{\underline{w}^T S_W \underline{w}}$$

$$\nabla_{\underline{w}} J = \frac{2\underline{w}^T S_W \underline{w}(S_B \underline{w}) - 2\underline{w}^T S_B \underline{w}(S_W \underline{w})}{(\underline{w}^T S_W \underline{w})^2} = 0$$

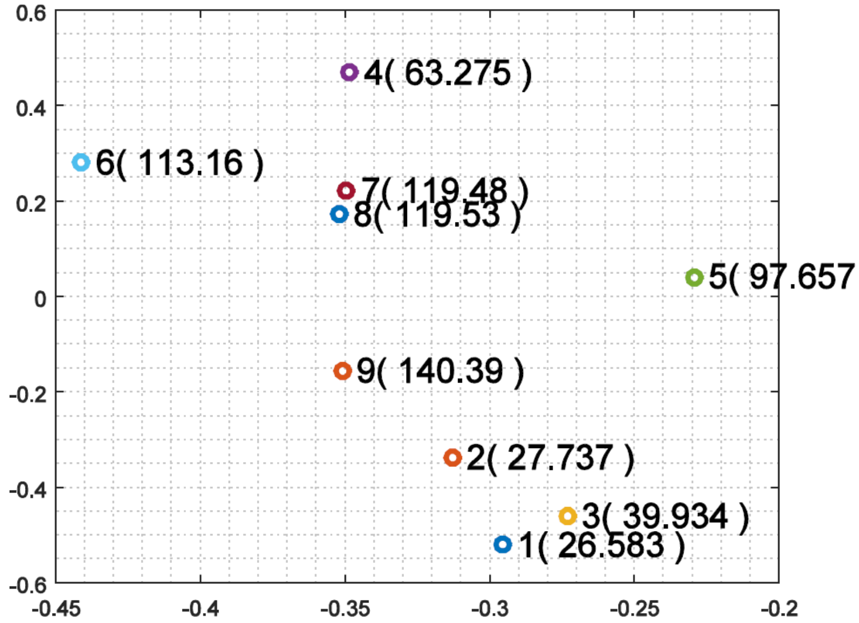
$$\Rightarrow S_B \underline{w} = \frac{\underline{w}^T S_B \underline{w}}{\underline{w}^T S_W \underline{w}} S_W \underline{w}$$

$$\Rightarrow S_W^{-1} S_B \underline{w} = \frac{\underline{w}^T S_B \underline{w}}{\underline{w}^T S_W \underline{w}} \underline{w} = \lambda \underline{w}$$

$\underline{w}$ *is an eigen vector of* $S_W^{-1} S_B$ *with maximum eigen value* $\frac{\underline{w}^T S_B \underline{w}}{\underline{w}^T S_W \underline{w}}$.

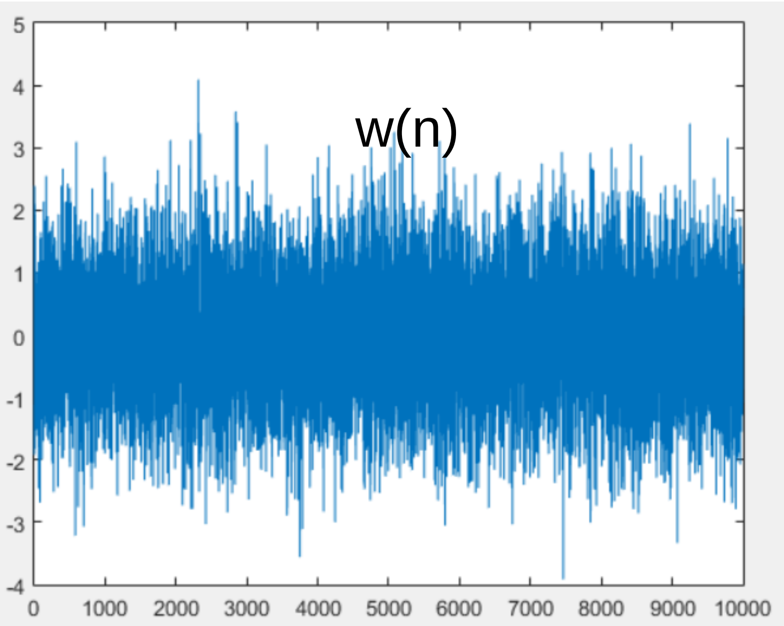
Why maximum? because any other eigen value will reduce J.

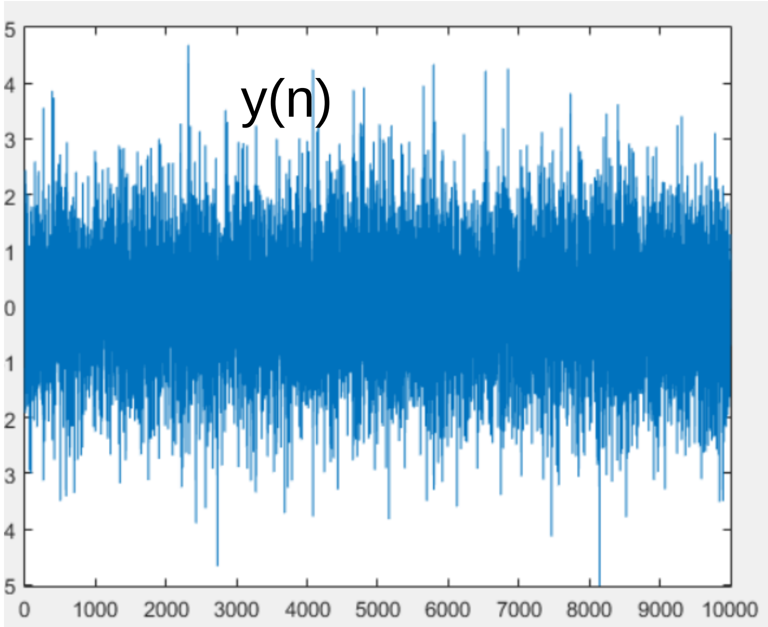
Problem 6, Murphy, Problem 12.8, page 418



- The angles are [26.5830 39.9344 27.7366 140.3943 97.6569 119.5270 119.4842 113.1596 63.2745], so the top 3 closest matches are documents 1,3,2.

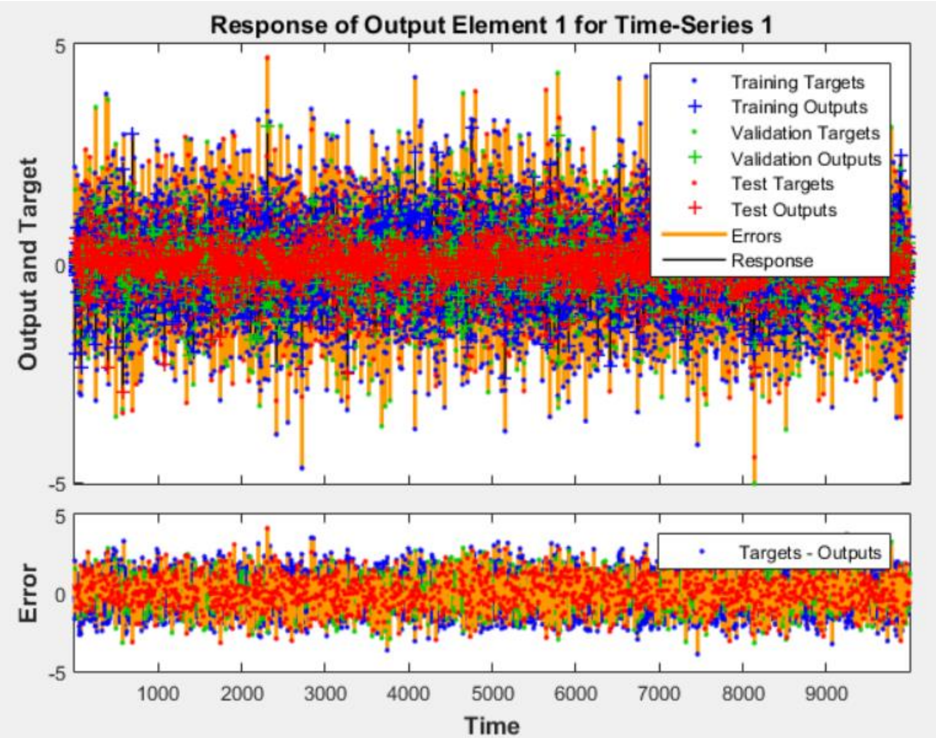
Problem 7


w(n)
5
4
3
2
1
0
-1
-2
-3
-4
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000


y(n)
5
4
3
2
1
0
1
2
3
4
5
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000

Problem 7

MLP

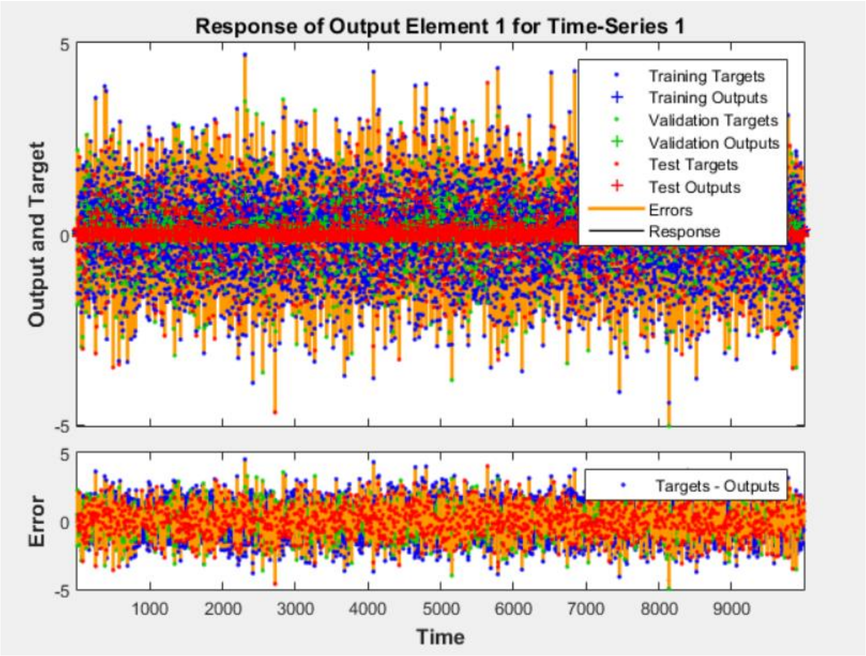


Results

	Target Values	MSE	R
Training:	7000	9.77812e-1	4.52310e-1
Validation:	1500	9.80923e-1	4.14389e-1
Testing:	1500	9.63141e-1	4.41403e-1

Problem 7

LMS



Results

	Target Values	MSE	R
Training:	7000	1.17477e-0	1.50068e-1
Validation:	1500	1.24034e-0	1.51571e-1
Testing:	1500	1.21404e-0	1.34426e-1