4.4. Show that maximization of the class separation criterion given by (4.23) with respect to w, using a Lagrange multiplier to enforce the constraint $w^Tw=1$, leads to the result that $w \propto (m_2-m_1)$.

Sol. from the question, we can get the ~~max~~ optimization problem.

$$\begin{cases} \max\ w^T(m_2-m_1) \\ \text{s.t. } w^Tw=1 \end{cases}$$

$$L(w,\lambda) = w^T(m_2-m_1) + \lambda(w^Tw-1)$$

$$\frac{\partial L(w,\lambda)}{\partial w} = (m_2-m_1) + 2\lambda w = 0 \qquad \therefore\ w = -\frac{1}{2\lambda}(m_2-m_1)$$

$$\therefore\ w \propto m_2-m_1$$

4.5. By making use of (4.20), (4.23), and (4.24), show that Fisher criterion (4.25) can be written in the form (4.26).

Sol. $$J(w) = \frac{(m_2-m_1)^2}{s_1^2+s_2^2}$$

$$(m_2-m_1)^2 = (w^Tm_2 - w^Tm_1)^2 = w^Tm_2m_2^Tw - 2w^Tm_2m_1w + w^Tm_1m_1^Tw$$

$$= w^T(m_2m_2^T - 2m_2m_1^T + m_1m_1^T)w = w^T(m_2-m_1)(m_2-m_1)^Tw$$

$$s_1^2+s_2^2 = \sum_{n\in C_1}(w^Tx_n - w^Tm_1)^2 + \sum_{n\in C_2}(w^Tx_n - w^Tm_2)^2$$

$$= \sum_{n\in C_1} w^T(x_n-m_1)(x_n-m_1)^Tw + \sum_{n\in C_2} w^T(x_n-m_2)(x_n-m_2)^Tw$$

$$= w^T\left[\sum_{n\in C_1}(x_n-m_1)(x_n-m_1)^T + \sum_{n\in C_2}(x_n-m_2)(x_n-m_2)^T\right]w$$

$$\therefore\ J(w) = \frac{(m_2-m_1)^2}{s_1^2+s_2^2} = \frac{w^TS_Bw}{w^TS_Ww}$$

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

4.6. Using the ~~defintion~~ definitions of the between-class and within-~~cl~~class covariance matrices given by (4.27) and (4.28), ~~resp~~ respectively, together ~~also~~ with (4.34) and 4.36 and the choice of target values described in Section 4.1.5, show that the expression 4.33 that minimizes the sum-of-squares error function can be written in the form (4.37).

Sol.

$$\sum_{n=1}^{N}(w^T x_n + w_0 - t_n)\, x_n = 0$$

$w_0 = -w^T m$ $\quad\therefore\quad \sum_{n=1}^{N}(w^T x_n + w_0 - t_n)x_n = \sum_{n=1}^{N}(w^T x_n - w^T m - t_n)x_n$

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

$$\sum_{n=1}^{N} t_n x_n = \sum_{n\in C_1} t_n x_n + \sum_{n\in C_2} t_n x_n = \frac{N}{N_1} m_1 N_1 + \left(-\frac{N}{N_2}\right) m_2 N_2 = N(m_1 - m_2)$$

$$\sum_{n=1}^{N} x_n x_n^T w - x_n m^T w = \left[\sum_{n=1}^{N} x_n x_n^T - x_n m^T\right]\cdot w$$

$$\sum_{n=1}^{N} x_n x_n^T - x_n m^T = \sum_{n\in C_1} x_n x_n^T - x_n m^T + \sum_{n\in C_2} x_n x_n^T - x_n m^T$$

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

$$= \sum_{n\in C_1} x_n x_n^T - \frac{N_1}{N} x_n m_1^T + \sum_{n\in C_2} x_n x_n^T - \frac{N_2}{N} x_n m_2^T + \sum_{n\in C_1} x_n \cdot \frac{N_2}{N} m_2^T + \sum_{n\in C_2} x_n \cdot \frac{N_1}{N} m_1^T$$

$$\sum_{n\in C_k}(x_n - m_k)(x_n - m_k)^T = \sum_{n\in C_k}(x_n x_n^T - m_k x_k^T - x_n m_k^T - m_k \cdot m_k^T)$$

$$= \sum_{n\in C_k}(x_n x_n^k - x_n m_k^T) - \sum_{n\in C_k} x_n^T m_k - m_k m_k^T = \sum_{n\in C_k}(x_n x_n^k - x_n m_k^T)$$

$$\therefore \sum_{n=1}^{N} x_n x_n^T - x_n m^T = \sum_{n\in C_1} x_n x_n^T - x_n m_1^T + \sum_{n\in C_2} x_n x_n^T - x_n m_2^T + \frac{N-N_1}{N}\sum_{n\in C_1} x_n m_1^T + \frac{N-N_2}{N}\sum_{n\in C_2} x_n m_2^T + \sum_{n\in C_1} x_n \frac{N_2}{N} m_2^T + \sum_{n\in C_2} x_n \cdot \frac{N_1}{N} m_1^T$$

$$= \sum_{n\in C_1}(x_n - m_1)(x_n - m_1)^T + \sum_{n\in C_2}(x_n - m_2)(x_n - m_2)^T + \frac{N-N_1}{N}\cdot N_1 m_1 m_1^T + \frac{N-N_2}{N} N_2 m_2 \cdot m_2^T + \frac{N_1 N_2}{N} m_1 m_2^T + \frac{N_1 N_2}{N} m_2 m_1^T$$

$$= S_W + \frac{N_1 N_2}{N} m_1 m_1^T + \frac{N_1 N_2}{N} m_2 m_2^T + \frac{N_1 N_2}{N} m_1 m_2^T + \frac{N_1 N_2}{N} m_2 m_1^T$$

$$= S_W + \frac{N_1 N_2}{N} S_B$$

$$\therefore \left(S_W + \frac{N_1 N_2}{N} S_B\right) w = N(m_1 - m_2)$$

4.7. Show that the logistic sigmoid function (4.59) satisfies the property $\sigma(-a) = 1 - \sigma(a)$ and that its inverse is given by $\sigma^{-1}(y) = \ln\{y/(1-y)\}$.

Sol. $\sigma(a) = \frac{1}{1+\exp(-a)}$ $\quad \sigma(-a) = \frac{1}{1+\exp(a)}$

$$1 - \sigma(a) = 1 - \frac{1}{1+\exp(-a)} = \frac{\exp(-a)}{1+\exp(-a)} = \frac{\exp(a)\cdot\exp(-a)}{\exp(a)[1+\exp(-a)]} = \frac{1}{1+\exp(a)} = \sigma(-a)$$

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

$\sigma(a) = \frac{1}{1+\exp(-a)}$, we can form it at $y = \frac{1}{1+\exp(-x)}$

so $1+\exp(-x) = y^{-1}$ $\qquad$ So $\sigma^{-1}(y) = \ln\frac{y}{1-y}$

$\exp(-x) = y^{-1} - 1 = \frac{1-y}{y}$

$-x = \ln\left(\frac{1-y}{y}\right)$

$x = \ln\left(\frac{1-y}{y}\right)^{-1} = \ln\frac{y}{1-y}$

4.8. Using (4.57) and (4.58), derive the result (4.65) for the posterior class probability in the two-class generative model with Gaussian densities, and verify the results (4.66) and (4.67) for the parameters w and w_0.

Sol. $$p(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)} = \frac{1}{1 + \frac{P(x|C_2)P(C_2)}{P(x|C_1)P(C_1)}}$$

$$\frac{P(x|C_2)P(C_2)}{P(x|C_1)P(C_1)} = \exp\{-\tfrac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2) + \ln P(C_2) + \tfrac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \ln P(C_1)\}$$

$$= \exp\{-\tfrac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_2 - \tfrac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + \tfrac{1}{2}x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_1 + \tfrac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \ln\frac{P(C_2)}{P(C_1)}\}$$

$$= \exp\{x^T\Sigma^{-1}(\mu_2-\mu_1) + \tfrac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 - \tfrac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + \ln\frac{P(C_2)}{P(C_1)}\}$$

$$P(C_1|x) = \sigma(w^Tx + w_0) = \frac{1}{1+e^{-(w^Tx+w_0)}} = \frac{1}{1+e^{-(x^Tw+w_0)}}$$

$$\therefore\ w = \Sigma^{-1}(\mu_1-\mu_2)$$

$$w_0 = -\tfrac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \tfrac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + \ln\frac{P(C_1)}{P(C_2)}$$

4.9. Consider a generative classification model for K classes defined by prior class probabilities $p(C_k) = \pi_k$ and general class-conditional densities $p(\phi|C_k)$ where ϕ is the input feature vector. Suppose we are given a training data set $\{\phi_n, t_n\}$ ~~where ϕ is the input feature input~~ where $n=1,\dots,N$, and t_n is a ~~bina~~ binary target vector of length K that uses the 1-of-K coding theme, so that it has components $t_{nj} = I_{jk}$ if pattern n is from class C_k. Assuming that the data points are drawn independently from this model, show that the maximum-likelihood solution for the prior probabilities is given by $\pi_k = \frac{N_k}{N}$, where N_k is the number of data points assigned to ~~the~~ class C_k.

Sol. $$p(\overset{t}{\cancel{\phi}}|\pi) = \prod_{k=1}^{K}(\pi_k)^{t_k}$$

~~max~~ likelihood function.

$$P(\phi, t|\pi) = \prod_{n=1}^{N}\prod_{k=1}^{K}[P(C_k)P(\phi_k|C_k)]^{t_k}$$

So, we should maximum likelihood function, ~~the~~ drop terms independent of π_k. after logarithm, the optimization problem can be formed as

$$\max \prod_{n=1}^{N}\prod_{k=1}^{K}(\pi_k)^{t_k} \quad \text{s.t.} \prod_{k=1}^{K}\pi_k = 1$$

$\Rightarrow$

$$\text{so, } \prod_{n=1}^{N}\prod_{k=1}^{K}(\pi_k)^{t_k} = \prod_{k=1}^{K}\pi_k^{N_k}$$

$$\therefore \ln\prod_{k=1}^{K}\pi_k^{N_k} = \sum_{k=1}^{K}N_k\ln\pi_k$$

$\Rightarrow$

$$\max \sum_{k=1}^{K}N_k\ln\pi_k \quad \text{s.t.} \sum_{k=1}^{K}\pi_k = 1$$

the ~~Lagrange~~ Lagrange function is

$$L(\lambda,\pi_k)=\sum_{k=1}^{K}N_k\ln\pi_k+\lambda\left(\sum_{k=1}^{K}\pi_k-1\right)$$

$$\frac{\partial L}{\partial\pi_k}=\frac{N_k}{\pi_k}+\lambda=0 \quad\Rightarrow\quad \pi_k=-\frac{N_k}{\lambda}$$

$$\sum_{k=1}^{K}\pi_k=\sum_{k=1}^{K}\frac{N_k}{-\lambda}=1 \quad\Rightarrow\quad -\frac{N}{\lambda}=1 \quad\Rightarrow\quad \lambda=-N$$

$\therefore\ \pi_k=-\frac{N_k}{-N}=\frac{N_k}{N}$ $\quad\therefore$ the maximum-likelihood solution for the prior probabilities is $\pi_k=\frac{N_k}{N}$

4.10. Consider the classification model of Exercise 4.9 and now suppose that the class-conditional densities are given by Gaussian distributions with a shared covariance matrix, so that $P(\phi|C_k)=N(\phi|\mu_k,\Sigma)$. Show that the maximum likelihood solution for the mean of the ~~mean~~ Gaussian distribution for class C_k is given by $\mu_k=\frac{1}{N_k}\sum_{n=1}^{N}t_{nk}\phi_n$. which represents the mean of those feature vectors assigned to class C_k. Similarly, show that the maximum likelihood solution for the shared covariance matix is given by $\Sigma=\sum_{k=1}^{K}\frac{N_k}{N}S_k$. where $S_k=\frac{1}{N_k}\sum_{n=1}^{N}t_{nk}(\phi_n-\mu_k)(\phi_n-\mu_k)^T$. Thus Σ is given by a weighted average of the covariances of the data associated with each class, in which the weighting coefficients are given by the prior probabilities of the classes.

Sol. $P(\phi,t|\pi)=\prod_{n=1}^{N}\prod_{k=1}^{K}[P(C_k)P(\phi|C_k)]^{t_{nk}}$

$$\log P(\phi,t|\pi)=\sum_{n=1}^{N}\sum_{k=1}^{K}t_{k}[\log P(C_k)+\log P(\phi|C_k)]$$

$\log P(C_k)$ is independent to μ_k and Σ. so, we only to maximum $\sum_{n=1}^{N}\sum_{k=1}^{K}[t_{k}\cdot\log P(\phi,C_k)]$

$$\sum_{n=1}^{N}\sum_{k=1}^{K}t_{k}\cdot\left(-\frac{1}{2}(\phi-\mu_k)^T\Sigma^{-1}(\phi-\mu_k)\right)-\frac{1}{2}\ln|\Sigma|-\frac{D}{2}\ln 2\pi)]$$

$$\frac{\partial}{\partial\mu_k}\left\{\sum_{n=1}^{N}\sum_{k=1}^{K}t_{k}\left(-\frac{1}{2}(\phi-\mu_k)^T\Sigma^{-1}(\phi-\mu_k)-\frac{1}{2}\ln|\Sigma|-\frac{D}{2}\ln 2\pi\right)\right\}=\sum_{n=1}^{N}-\frac{1}{2}\cdot t_{nk}\cdot 2\cdot\Sigma^{-1}(\mu_k-\phi_n)=0$$

$$\sum_{n=1}^{N}t_{nk}\Sigma^{-1}(\mu_k-\phi_n)=0 \quad\Rightarrow\quad \sum_{n=1}^{N}t_{nk}\mu_k=\sum_{n=1}^{N}t_{nk}\phi_n$$

$\because\ \sum_{n=1}^{N}t_{nk}=N_k$

$\therefore\ \mu_k=\frac{1}{N_k}\sum_{n=1}^{N}t_{nk}\phi_n$

$$\frac{\partial}{\partial \Sigma}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left[-\frac{1}{2}(\phi-\mu_k)^T\Sigma^{-1}(\phi-\mu_k)-\frac{1}{2}\ln|\Sigma|-\frac{D}{2}\ln 2\pi\right]$$

$$=\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left[\frac{1}{2}\Sigma^{-1}(\phi_n-\mu_k)(\phi_n-\mu_k)^T\Sigma^{-1}-\frac{1}{2}\Sigma^{-1}\right]=0$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\Sigma^{-1}(\phi_n-\mu_k)(\phi_n-\mu_k)^T=\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} I$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}(\phi_n-\mu_k)(\phi_n-\mu_k)^T=\sum_{k=1}^{K} N_k\Sigma$$

$$\Sigma=\frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N} t_{nk}(\phi_n-\mu_k)(\phi_n-\mu_k)^T=\sum_{k=1}^{K}\frac{N_k}{N}\cdot\frac{1}{N_k}\sum_{n=1}^{N} t_{nk}(\phi_n-\mu_k)(\phi_n-\mu_k)^T$$

$$\Sigma=\sum_{k=1}^{K}\frac{N_k}{N}S_N \qquad S_N=\frac{1}{N_k}\sum_{n=1}^{N} t_{nk}(\phi_n-\mu_k)(\phi_n-\mu_k)^T$$

4.12. Verify the relation (4.88) for the derivate of the logistic sigmoid function defined by (4.59)

Sol. the sigmoid function $y=\frac{1}{1+e^{-x}}$

$$y'=\frac{-e^{-x}\times(-1)}{(1+e^{-x})^2}=\frac{e^{-x}}{(1+e^{-x})^2}=\frac{1}{1+e^{-x}}\cdot\frac{e^{-x}}{1+e^{-x}}=\frac{1}{1+e^{-x}}\left(1-\frac{1}{1+e^{-x}}\right)=y(1-y)$$

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

4.13. By making use of the result (4.88) for the derivative of the logistic sigmoid, show that the derivative of the error function (4.90) for the logistic regression model is given by (4.91)

Sol. $E(w)=-\ln p(t|w)=-\sum_{n=1}^{N}\{t_n\ln y_n+(1-t_n)\ln(1-y_n)\}$ where $y_n=\sigma(a_n)$ $a_n=w^T\phi_n$

$$\frac{\partial E}{\partial w_i}=-\sum_{n=1}^{N}\left\{t_n\frac{1}{y_n}y_n(1-y_n)\phi_{ni}+(1-t_n)\frac{-1}{1-y_n}y_n(1-y_n)\phi_{ni}\right\}$$

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

$$=-\sum_{n=1}^{N}(t_n-t_ny_n+t_ny_n-y_n)\phi_{ni}$$

$$=\sum_{n=1}^{N}(y_n-t_n)\phi_{ni}$$

$$\therefore \frac{\partial E}{\partial w}=\sum_{n=1}^{N}(y_n-t_n)\phi_n$$

4.15. Show that Hessian matrix H for the logistic regression model, given by (4.97), is positive definite. Here R is a diagonal matrix with elements $y_n(1-y_n)$, and y_n is the output of the logistic regression model for input vector x_n. Hence show that the error function is a concave function of w and that it has a unique minimum.

Sol. Assuming that the argument to the sigmoid function (4.87) is finite, the diagonal elements ~~$(y_n(1-y_n))^{1/2}$, and thus $\phi^T R \phi$ is~~ of R will be strictly positive. Then

$$v^T \Phi^T R \Phi v = (v^T \Phi^T R^{1/2})(R^{1/2} \Phi v) = \|R^{1/2} \Phi v\|^2 > 0$$

where $R^{1/2}$ is a diagonal matrix with elements $(y_n(1-y_n))^{1/2}$, and thus $\phi^T R \phi$ is positive definite.

Now consider a Taylor expansion of $E(w)$ around a minima, w^*,

$$E(w) = E(w^*) + \frac{1}{2}(w-w^*)^T H (w-w^*)$$

where the linear term has vanished since w^* is a minimum. Now let

$$w = w^* + \lambda v$$

where v is an ~~ab~~ arbitrary, non-zero vector in the weight space and consider

$$\frac{\partial^2 E}{\partial \lambda^2} = v^T H v > 0$$

This shows that $E(w)$ is convex. Moreover, at the minimum of $E(w)$,

$$H(w-w^*) = 0$$

and since H is positive definite, H^{-1} exists and $w = w^*$ must be the unique minimum.