Question 1

(a) From Bayes' theorem, $p(t|x) \propto p(x|t)\,p(t)$

$p(x) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{h^D} k\left(\frac{x - x_n}{h}\right)$ $\quad p(x|t) = \frac{1}{N_t}\sum_{n=1}^{N} \frac{1}{Z_k} k(x, x_n)\delta(t, t_n)$

N_t is the number of input vectors with label t ($t=+1$ or -1) and $N = N_{+1} + N_{-1}$. $\delta(t, t_n)$ equals 1 if $t = t_n$ and 0 otherwise. Z_k is the normalization constant for the kernel. The minimum misclassification-rate is achieved if, for each new input vector, $\tilde{x}$, we chose $\tilde{t}$ to maximize $p(\tilde{t}|\tilde{x})$. With equal class priors, this is equivalent to maximizing $p(\tilde{x}|\tilde{t})$ and thus

$$\tilde{t} = \begin{cases} +1 & \text{iff } \frac{1}{N_{+1}}\sum_{i:t_i=+1} k(\tilde{x}, x_i) \geq \frac{1}{N_{-1}}\sum_{j:t_j=-1} k(\tilde{x}, x_j) \\ -1 & \text{otherwise} \end{cases}$$

Dropped the factor $1/Z_k$, because it's a scaling factor. Use the encode scheme.

$$\tilde{t} = \text{sign}\left(\sum_{n=1}^{N} \frac{t_n}{N_{t_n}} k(\tilde{x}, x_n)\right)$$

(b) let $k(x, x_n) = x^T x_n$ $\quad p(x|t=+1) = \frac{1}{N_{+1}}\sum_{n:t_n=+1} x^T x_n = \bar{x}^T \bar{x}^+$

The sum in the middle expression runs over all vectors x_n for which $t_n = +1$ and $\bar{x}^+$ indicates the mean of these vectors. and similar definition for the negative class.

$$\tilde{t} = \begin{cases} +1 & \text{if } \bar{x}^T x^+ \geq \bar{x}^T x^- \\ -1 & \text{otherwise} \end{cases}$$

(c) The same argument also applies for feature space $\phi(x)$

$$\tilde{t} = \begin{cases} +1 & \text{if } \phi(\tilde{x})^T \phi(\bar{x})^+ \geq \phi(\tilde{x})^T \phi(\bar{x})^- \\ -1 & \text{otherwise} \end{cases}$$

Question 2

$p(t=1|y) = \sigma(y) \quad \Rightarrow p(t=-1|y) = 1 - p(t=1|y) = 1 - \sigma(y) = \sigma(-y)$

data $D = \{(t_1, x_1), \ldots, (t_N, x_N)\}$

$$p(D) = \prod_{t_n=1} \sigma(y_n) \prod_{t_n=-1} \sigma(-y_n) = \prod_{n=1}^{N} \sigma(t_n y_n)$$

where $y_n = y(x_n)$. Taking the negative logarithm of this, we get

$$-\ln p(D) = -\ln \prod_{n=1}^{N} \sigma(t_n y_n) = \sum_{n=1}^{N} \ln \sigma(t_n y_n) = \sum_{n=1}^{N} \ln(1 + e^{(-t_n y_n)})$$

Combining with regularization term $\lambda\|w\|^2$, we get

$$\sum_{n=1}^{N} E_{LR}(y_n t_n) + \lambda\|w\|^2$$

Question 3

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

$$p(w|\alpha) = \prod_{i=1}^{M} N(w_i|0, \alpha_i^{-1})$$

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

$$p(t|X, \alpha, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \frac{1}{(2\pi)^{M/2}} \prod_{i=1}^{M} \alpha_i \int \exp\{-E(w)\}\, dw$$

where $E(w) = \frac{\beta}{2}(w-m)^T \Sigma^{-1}(w-m) + E(t)$ $\quad A = \text{diag}(\alpha)$

$E(w) = \frac{1}{2}(w-m)^T \Sigma^{-1}(w-m) + E(t)$ $\quad m = \beta\Sigma\phi^T t$ $\quad \Sigma = (A + \beta\phi^T\phi)^{-1}$

$E(t) = \frac{\beta}{2}(\beta t^T t - m^T \Sigma m)$

$\int \exp\{-E(w)\}\, dw = \exp\{-E(t)\}(2\pi)^{\frac{M}{2}} |\Sigma|^{\frac{1}{2}}$

$C = \beta^{-1} I + \phi A^{-1} \phi^T$

Using $(A + BD^{-1}C)^{-1} = A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1}$

We can rewrite $E(t) = \frac{1}{2}(\beta t^T t - m^T \Sigma^{-1} m) = \frac{1}{2}(\beta t^T t - \beta t^T \phi \Sigma \Sigma^{-1} \Sigma \phi^T t \beta)$

$= \frac{1}{2} t^T (\beta I - \beta\phi\Sigma\phi^T\beta) t = \frac{1}{2} t^T (\beta^{-1} I + \phi A^{-1} \phi^T)^{-1} t = \frac{1}{2} t^T C^{-1} t$

This is the last term of the right handside. The first and second terms are given implicitly, as they form the normalization constant for the posterior Gaussian distribution $p(t|X, \alpha, \beta)$

Question 4

Using results from Question 3

we can rewrite $\ln(t|X, \alpha, \beta) = \ln N(t|0, C) = -\frac{1}{2}\{N\ln(2\pi) + \ln|C| + t^T C^{-1} t\}$

$$= \frac{N}{2}\ln\beta + \frac{1}{2}\sum_{i}^{M} \ln\alpha_i - E(t) - \frac{1}{2}\ln|\Sigma| - \frac{N}{2}\ln(2\pi)$$

By taking the derivative w.r.t. α_i:

$$\frac{\partial}{\partial\alpha_i} \ln p(t|X, \alpha, \beta) = \frac{1}{2\alpha_i} - \frac{1}{2}\Sigma_{ii} - \frac{1}{2} m_i^2$$

Setting the derivative to zero

$$\alpha_i = \frac{1-\alpha_i \Sigma_{ii}}{m_i^2} = \frac{\gamma_i}{m_i^2} \qquad \gamma_i = 1-\alpha_i \Sigma_{ii}$$

$$\frac{\partial}{\partial \beta} \ln p(t \mid X, \alpha, \beta) = \frac{1}{2}\left(\frac{N}{\beta} - \|t-\Phi m\|^2 - \mathrm{Tr}[\Sigma \Phi^T \Phi]\right)$$

$$\begin{aligned} \Sigma \Phi^T \Phi &= \Sigma \Psi^T \Phi + \beta^{-1} \Sigma A - \beta^{-1} \Sigma A \\ &= \Sigma (\Phi^T \Phi \beta + A) \beta^{-1} - \beta^{-1} \Sigma A \\ &= (A + \beta \Phi^T \Phi)^{-1} (\Phi^T \Phi \beta + A) \beta^{-1} - \beta^{-1} \Sigma A \\ &= (I - A\Sigma) \beta^{-1} \end{aligned}$$

$(I - A\Sigma)$ equals $\gamma_i = 1 - \alpha_i \Sigma_{ii}$ in matrix form

Set $\frac{\partial}{\partial \beta} \ln p(t \mid X, \alpha, \beta) = 0$

By rearranging $(\beta^{new})^{-1} = \frac{\|t-\Phi m\|^2}{N - \sum_i \gamma_i}$

Question 5

(a) $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle = \langle \phi(x_j), \phi(x_i) \rangle = K(x_j, x_i)$

(b) $\|\phi(x_i) - \phi(x_j)\|^2$

$= \langle \phi(x_i), \phi(x_i) \rangle + \langle \phi(x_j), \phi(x_j) \rangle - 2 \cdot \langle \phi(x_i), \phi(x_j) \rangle = K(x_i, x_i) + K(x_j, x_j) - 2 \cdot K(x_i, x_j) = (+1) - 2 \exp(-\frac{1}{2}\|x_i - x_j\|^2) < 2$

Question 6

$\|x_{far} - x_i\| \to \infty \quad \forall i \in SV$

$\Rightarrow K(x_{far}, x_i) \approx 0 \quad \forall i \in SV$

$\Rightarrow \sum_{i \in SV} y_i \alpha_i K(x_i, x) \approx 0$

$\Rightarrow f(x, \alpha, \hat{w}_0) \approx \hat{w}_0$