Q1.

From Bayes' theorem we have.

$$p(t|x) \propto p(x|t)p(t)$$

where $p(x|t) = \frac{1}{N_t}\sum_{n=1}^{N}\frac{1}{Z_k}k(x, x_n)\delta(t, t_n)$

Here N_t is the number of input vectors with label 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 archieved if, for each new input vector, $\tilde{x}$, we chose $\tilde{t}$ to maximise $p(\tilde{t}|\tilde{x})$, with equal class priors this is equivalent to maximizing $p(\tilde{x}|\tilde{t})$ and thus

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

Here we have dropped the factor $1/Z_k$ since it only acts as a common scaling factor. Using the encoding ~~schedu~~ scheme for the label, this classification rule can be written in the more compact form

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

Now we take $k(\tilde{x}, x_n) = x^T x_n$, which results in the kernel density.

$$p(x|t=+1) = \frac{1}{N_{+1}}\sum_{n:t_n=+1}x^T x_n = x^T\hat{x}^+$$

Here, the sum in the middle expression runs overall vectors x_n for which $t_n = +1$ and x^+ denotes the mean of these vectors, ~~with~~ the corresponding definition for the negative class. Note that this density is improper, since it cannot be normalized. However, we can still compare likelihoods under this density, resulting in the classification rule.

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

The same argument would of course also apply in the features space $\phi(x)$

Q2. If $p(t=1|y) = \sigma(y)$, then $p(t=-1|y) = 1 - p(t=1|y) = 1 - \sigma(y) = \sigma(-y)$.

Thus, given i.i.d. data $D = \{(t_1, x_n), \ldots, (t_N, x_N)\}$, we can write the corresponding likelihood as

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

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

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

Combining this with the regularization term $\lambda\|w\|^2$, we obtain

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

Q3.

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

where $E(w) = \frac{\beta}{2}\|t - \phi w\|^2 + \frac{1}{2}w^T A w$, and $A = \text{diag}(\alpha)$.

Completing the square over w, we get $E(w) = \frac{1}{2}(w-m)^T\Sigma^{-1}(w-m) + E(t)$.

where m and Σ are given by $m = \beta\Sigma\phi^T t$, $\Sigma = (A + \beta\phi^T\phi)^{-1}$

and $E(t) = \frac{1}{2}(\beta t^T t - m^T \Sigma^{-1} m)$.

Using $E(w)$, we can evaluate the integral in $E(t)$ to obtain.

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

Considering this as a function of t, that we only need to ~~deal~~ deal with the factor ~~exp{E(t)}~~ $\exp\{-E(t)\}$. Then we can re-write $E(t)$ as follows.

$$E(t) = \frac{1}{2}(\beta t^T t - m^T \Sigma^{-1} m) = \frac{1}{2}(\beta t^T t - \beta t^T \phi S \Sigma^{-1} S \phi^T t \beta).$$

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

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

This gives us the last terms on the r.h.s of the target equation; the two preceding terms are given implicitly, as they form the ~~re~~normalization constant for the posterior Gaussian distribution $p(t|X, \alpha, \beta)$.

Q4. Using the result from Q3, we can write $\ln p(t|X, \alpha, \beta)$ in the form of:

$$\ln p(t|X, \alpha, \beta) = \frac{N}{2} \ln \beta + \frac{1}{2} \sum_i^N \alpha_i - E(t) - \frac{1}{2} \ln |\Sigma| - \frac{N}{2} \ln(2\pi)$$

By making use of $E(t) = \frac{1}{2}(\beta t^T t - m^T \Sigma^{-1} m)$ and $\Sigma = (A + \beta \phi^T \phi)^{-1}$ together with $\frac{\partial}{\partial x} \ln |A| = Tr(A^{-1} \frac{\partial A}{\partial x})$, we can take the derivatives of this w.r.t α_i, yielding.

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

Setting this to zero and re-arranging, we obtain

$$\alpha_i = \frac{1 - \alpha_i \Sigma_{ii}}{m_i^2} = \frac{\gamma_i}{m_i^2} \quad \text{where we have used} \quad \gamma_i = 1 - \alpha_i \Sigma_{ii}$$

Similarly, for β we see that

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

Using $\Sigma \Sigma = (A + \beta \phi^T \phi)^{-1}$, we can rewrite the argument of the trace operator as

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

Here the first factor on the r.h.s of the last line equals $\gamma_i = 1 - \alpha_i \Sigma_{ii}$ ~~written~~ in matrix form. We can use this to set $\frac{\partial}{\partial \beta} \ln p(t|X, \alpha, \beta)$ equals to zero the re-arrange to obtain $(\beta^{new})^{-1} = \frac{\|t - \phi m\|^2}{N - \sum_i \gamma_i}$.

Q5. 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) $\|\phi(x_i) - \phi(x_j)\|^2 = \langle \phi(x_i), \phi(x_i) \rangle + \langle \phi(x_j), \phi(x_j) \rangle - 2 \langle \phi(x_i), \phi(x_j) \rangle$

$$= k(x_i, x_i) + k(x_j, x_j) - 2k(x_i, x_j)$$

$$= 1 + 1 - 2 \exp(-\frac{1}{2} \|x_i - x_j\|^2) < 2.$$

Q6. $\# \|x_{far} - x_i\| \gg 0. \quad \forall i \in SV \Rightarrow k(x_{far}, x_i) \approx 0 \quad \forall_i \in SV. \Rightarrow \sum_{i \in SV} y_i \alpha_i k(x_i, x) \approx 0 \Rightarrow f(x_{far}, \alpha, \hat{w}_0) \approx \hat{w}_0$