1. As the name implies linear classifiers only generate decision boundaries of the form \( w'x + b = 0 \). Non linear mappings of the feature can increase the effective dimensionality. A linear decision boundary in this mapped space will be non-linear in the original space. Note there is an increase in the number of model parameters that need to be trained for the decision boundary.

A mapping will exist if the points have distinct labels (i.e. no point has multiple class labels associated with it.)

2. The conditions that must be satisfied are:

\[
\alpha_i \geq 0 \\
\sum_{i=1}^{m} \alpha_i y_i = 0
\]

The solution from the lecture notes are

\[
\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \frac{1}{8}
\]

By inspection the conditions are satisfied. Consider the value of the mapped points

\[
\begin{pmatrix}
1 \\
\sqrt{2} \\
-\sqrt{2} \\
-\sqrt{2} \\
1
\end{pmatrix}
; \\
\begin{pmatrix}
1 \\
-\sqrt{2} \\
+\sqrt{2} \\
-\sqrt{2} \\
1
\end{pmatrix}
; \\
\begin{pmatrix}
1 \\
\sqrt{2} \\
\sqrt{2} \\
\sqrt{2} \\
1
\end{pmatrix}
; \\
\begin{pmatrix}
1 \\
-\sqrt{2} \\
-\sqrt{2} \\
\sqrt{2} \\
1
\end{pmatrix}

The direction of the decision boundary is given by

\[
w = \sum_{i=1}^{m} \alpha_i y_i \Phi(x_i)
\]

\[
= \frac{1}{2} \begin{pmatrix}
0 \\
0 \\
0 \\
-\sqrt{2}
\end{pmatrix}
\]
To find $b$ substitute into the expression

$$\alpha_i ((y_i ((\mathbf{w}, \mathbf{x}_i) + b) - 1)) = 0$$

Select the point $[1, 1]'$

$$\frac{1}{8} (-1 \times (-1 + b) - 1) = 0$$

So $b = 0$. Check using the point $[1, -1]$

$$\frac{1}{8} (1 \times (1 + b) - 1) = 0$$

This is correct (also other points satisfy this). The equation of the decision boundary is

$$x_1 x_2 = 0$$

3. (a) The decision boundary and margins are shown below.

(b) There are four support vectors

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

(c) There are multiple solutions for $\alpha$ (though a unique decision boundary) to this as it is an under-specified problem. If the Lagrange multiplier for the fourth point is set to zero then the associated values of $\alpha$, 4, 2, 2 and associated class labels 1, $-1$ and $-1$, Again it is possible to check that these points satisfy the training criteria.
4. From the question, this is the same as the probability that \( f(x) + \epsilon \geq 0 \). This is the same as the probability that \( \epsilon \geq -\hat{w}'x \). Thus

\[
P(y = +1|x, \hat{w}) = \int_{-\hat{w}'x}^{\infty} N(\epsilon; 0, \sigma_n^2)d\epsilon = 1 - \int_{-\infty}^{-\hat{w}'x} N(\epsilon; 0, \sigma_n^2)d\epsilon
\]

Using the standard equalities

\[
a = -\hat{w}'x/\sigma_n
\]

5. From the lecture notes

\[
p(y|X) = \int p(y|X, w)p(w)dw = \int N(y; \hat{X}'w, \sigma_n^2I)N(w; 0, \sigma_w^2I)dw
\]

Considering just the term inside the integral, this may be expressed as \((k\) is a constant\)

\[
k \exp\left(-\frac{1}{2\sigma_n^2} (y - \hat{X}'w)'(y - \hat{X}'w) - \frac{1}{2\sigma_w^2}w'w\right) = k \exp\left(-\frac{1}{2} \left( w' \left( \frac{1}{\sigma_n^2}XX' + \frac{1}{\sigma_w^2}I \right) - \frac{2}{\sigma_n^2}w'Xy + \frac{1}{\sigma_n^2}y'y \right) \right)
\]

Completing the square in \( w \), this can be expressed as

\[
k \exp\left(-\frac{1}{2} \left( (w - \mu_w)'\Sigma_w^{-1}(w - \mu_w) - \mu_w'\Sigma_w^{-1}\mu_w + \frac{1}{\sigma_n^2}y'y \right) \right)
\]

where

\[
\mu_w = \frac{1}{\sigma_n^2} \left( \frac{1}{\sigma_n^2}XX' + \frac{1}{\sigma_w^2}I \right)^{-1}Xy
\]

\[
\Sigma_w = \left( \frac{1}{\sigma_n^2}XX' + \frac{1}{\sigma_w^2}I \right)^{-1}
\]

Integrating over \( w \) yields a constant and

\[
p(y|X) \propto \exp\left(-\frac{1}{2} \left( y' \left( \frac{1}{\sigma_n^2}I - \frac{1}{\sigma_n^2}X' \left( \frac{1}{\sigma_n^2}XX' + \frac{1}{\sigma_w^2}I \right)^{-1}X \frac{1}{\sigma_n^2} \right) y \right) \right) = N(y; 0, \sigma_n^2XX' + \sigma_w^2I)
\]

Using the equality given in the question.
6. Using the standard form of the partition matrix in the lecture notes

\[
\begin{bmatrix}
  f(\hat{x}) \\
y
\end{bmatrix} 
\sim \mathcal{N}
\left(0,
\begin{bmatrix}
  k(\hat{x}, \hat{x}) & k(\hat{x}, X) \\
k(\hat{x}, X) & K(X, X) + \sigma_n^2 I
\end{bmatrix}
\right)
\]

The conditional distribution can then be written as

\[
f(\hat{x}) | X, y, \hat{x} \sim \mathcal{N}(\mu_t, \sigma_t^2)
\]

where

\[
\mu_t = k(\hat{x}, X)' \left[ K(X, X) + \sigma_n^2 I \right]^{-1} y
\]

\[
\text{var}_n(f(\hat{x})) = \sigma_t^2 = k(\hat{x}, \hat{x}) - k(\hat{x}, X)' \left[ K(X, X) + \sigma_n^2 I \right]^{-1} k(\hat{x}, X)
\]

Now consider the following partition where $\bar{X}$ excludes observation $x_n$ from the training data

\[
\left[ K(X, X) + \sigma_n^2 I \right]^{-1} = \left[ \begin{bmatrix} K(\bar{X}, \bar{X}) + \sigma_n^2 I & k(x_n, \bar{X}) \\ k(x_n, X) & k(x_n, x_n) + \sigma_n^2 \end{bmatrix} \right]^{-1} = \begin{bmatrix} A & b \\ b' & c \end{bmatrix} = \begin{bmatrix} E & g \\ g' & h \end{bmatrix}
\]

and

\[
k(\hat{x}, X) = \begin{bmatrix} k(\hat{x}, \bar{X}) \\ k(\hat{x}, x_n) \end{bmatrix}
\]

Looking at the second term in the variance only, for the $n$ samples this can be expressed as

\[
d_n = k(\hat{x}, \bar{X})' Eyk(\hat{x}, \bar{X}) + 2k(\hat{x}, x_n)g'k(\hat{x}, \bar{X}) + k(\hat{x}, x_n)hk(\hat{x}, x_n)
\]

and for the $n - 1$ samples

\[
d_{n-1} = k(\hat{x}, \bar{X})' (K(X, X) + \sigma_n^2 I)^{-1} k(\hat{x}, \bar{X})
\]

Consider the form for the partitioned $n$ example case. Let

\[
k_a = k(\hat{x}, \bar{X})' A^{-1} b
\]

and

\[
k_b = \frac{1}{k(x_n, x_n) + \sigma_n^2 - k(x_n, \bar{X})' \left( K(X, X) + \sigma_n^2 I \right)^{-1} k(x_n, \bar{X})}
\]

Then

\[
d_n = d_{n-1} + k_n k_b k_a - 2k(\hat{x}, x_n)k_b k_a + k(\hat{x}, x_n)k_b k(\hat{x}, x_n)
\]

\[
d_n = d_{n-1} + (k(\hat{x}, x_n) - k_a)^2 k_b
\]

by definition the second term is non-negative (definition of semi-positive definite), so

\[
d_n \geq d_{n-1} \text{ therefore } \text{var}_n(f(\hat{x})) \leq \text{var}_{n-1}(f(\hat{x}))
\]
7. Using the equality from question (6), the following equality (assuming that $i = n$)

$$
\begin{bmatrix}
A & b \\
b' & \infty
\end{bmatrix}^{-1}
= 
\begin{bmatrix}
A^{-1} & 0 \\
0' & 0
\end{bmatrix}
$$

Examining the variance term and substituting in $\lambda_i \rightarrow \infty$

$$
\Sigma_w = \left( \frac{1}{\sigma_n^2} \Phi'\Phi + A \right)^{-1}
= 
\begin{bmatrix}
A^{-1} & 0 \\
0' & 0
\end{bmatrix}
$$

Thus the variance of element $i$ goes towards zero. For the mean

$$
\mu_w = \frac{1}{\sigma_n^2} \Sigma_w \Phi' y
= 
\frac{1}{\sigma_n^2} \begin{bmatrix}
A^{-1} & 0 \\
0' & 0
\end{bmatrix} 
\begin{bmatrix}
\phi(x_1)' y \\
\vdots \\
\phi(x_n)' y
\end{bmatrix}
= 
\frac{1}{\sigma_n^2} \begin{bmatrix}
A^{-1} \\
\phi(x_1)' y \\
\vdots \\
\phi(x_{n-1})' y \\
0
\end{bmatrix}
$$

It is clear that all elements associated with $x_n$ have been removed from the prediction. This is the basis for the sparse representation of RVMs.

8. From the root node the data is split from $-\infty$ to $x_1$. Assume that the split for the node occurs at $x_s$. The posterior probability for class $\omega_1$ for the root node

$$
P(\omega_1|N) = \frac{\int_{x_s}^{x_1} \mathcal{N}(x; 0, 1) dx}{\int_{-\infty}^{x_s} \mathcal{N}(x; 0, 1) + \mathcal{N}(x; 1, 1) dx}
$$

The left node of the hypothesised split

$$
P(\omega_1|N_L) = \frac{\int_{x_s}^{x_1} \mathcal{N}(x; 0, 1) dx}{\int_{-\infty}^{x_s} \mathcal{N}(x; 0, 1) + \mathcal{N}(x; 1, 1) dx}
$$

For the right node

$$
P(\omega_1|N_R) = \frac{\int_{x_s}^{x_1} \mathcal{N}(x; 0, 1) dx}{\int_{x_s}^{x_1} \mathcal{N}(x; 0, 1) + \mathcal{N}(x; 1, 1) dx}
$$

The fractions assigned to the left node is

$$
n_L = \frac{\int_{x_s}^{x_1} \mathcal{N}(x; 0, 1) + \mathcal{N}(x; 1, 1) dx}{\int_{-\infty}^{x_1} \mathcal{N}(x; 0, 1) + \mathcal{N}(x; 1, 1) dx}
$$
The entropy cost function can then be written as

\[ I(N_L) = P(\omega_1|N_L) \log(P(\omega_1|N_L)) + (1 - P(\omega_1|N_L)) \log(1 - P(\omega_1|N_L)) \]

These can then be directly substituted into to the overall expression.

9. (a)

\[
E\{\tilde{p}(x)\} = E \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_n} \phi \left( \frac{x - x_i}{h_n} \right) \right\} \\
= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_n} E \left\{ \phi \left( \frac{x - x_i}{h_n} \right) \right\} \\
= \frac{1}{h_n} E \left\{ \phi \left( \frac{x - x_i}{h_n} \right) \right\} 
\]

As \( \phi() \) is Gaussian distributed then

\[
E \left\{ N \left( \frac{x - x_i}{h_n}; 0, 1 \right) \right\} = \int N \left( \left( \frac{x - v}{h_n} \right); 0, 1 \right) N(v; \mu, \sigma^2) dv \\
= \int h_n N(x; v, h_n^2) N(v; \mu, \sigma^2) dv \\
= h_n N(x; \mu, \sigma^2 + h_n^2) 
\]

Hence

\[
E\{\tilde{p}(x)\} = N(x; \mu, \sigma^2 + h_n^2) 
\]

(b) Since each of the individual samples is independent, then the total variance is a combination of the individual variances. Hence

\[
\text{var}[\tilde{p}(x)] = \frac{1}{n^2} \sum_{i=1}^{n} \left( \frac{1}{h_n} \right)^2 E \left\{ \phi^2 \left( \frac{x - x_i}{h_n} \right) \right\} - (E\{\tilde{p}(x)\})^2 \\
= \frac{1}{n^2} \sum_{i=1}^{n} \left( \int \left( N(x; v, h_n^2) \right)^2 N(v; \mu, \sigma^2) dv - (E\{\tilde{p}(x)\})^2 \right) \\
= \frac{1}{n} \left( \frac{1}{2h_n \sqrt{\pi}} N(x; \mu, \sigma^2 + \frac{h_n^2}{2}) - \left( N(x; \mu, \sigma^2 + h_n^2) \right)^2 \right) \\
= \frac{1}{n} \left( \frac{1}{2h_n \sqrt{\pi}} N(x; \mu, \sigma^2 + \frac{h_n^2}{2}) - \frac{1}{2 \sqrt{\sigma^2 + h_n^2}} N(x; \mu, \frac{\sigma^2 + h_n^2}{2}) \right) 
\]

As \( h_n \) gets small

\[
\text{var}[\tilde{p}(x)] \approx \frac{1}{2nh_n \sqrt{\pi}} p(x) 
\]
\( p(x) - \mathcal{E}\{\tilde{p}(x)\} = \mathcal{N}(x; \mu, \sigma^2) - \mathcal{N}(x; \mu, \sigma^2 + h_n^2) \)
\[
= \left(1 - \frac{\mathcal{N}(x; \mu, \sigma^2 + h_n^2)}{\mathcal{N}(x; \mu, \sigma^2)}\right) \mathcal{N}(x; \mu, \sigma^2)
\]
\[
= \left(1 - \frac{\frac{\sigma^2}{\sigma^2 + h_n^2}}{\exp \left(\frac{h_n^2(x - \mu)^2}{2(\sigma^2 + h_n^2)\sigma^2}\right)}\right) \mathcal{N}(x; \mu, \sigma^2)
\]

As \( h_n \) gets small
\[
\sqrt{\frac{\sigma^2}{\sigma^2 + h_n^2}} = \sqrt{1 - \frac{h_n^2}{\sigma^2 + h_n^2}} \approx 1 - \frac{h_n^2}{2(\sigma^2 + h_n^2)}
\]
\[
\exp \left(\frac{h_n^2(x - \mu)^2}{2(\sigma^2 + h_n^2)\sigma^2}\right) \approx 1 + \frac{h_n^2(x - \mu)^2}{2(\sigma^2 + h_n^2)\sigma^2}
\]

Hence
\[
p(x) - \mathcal{E}\{\tilde{p}(x)\} \approx \left(1 - \left(1 - \frac{h_n^2}{2(\sigma^2 + h_n^2)}\right) \left(1 + \frac{h_n^2(x - \mu)^2}{2(\sigma^2 + h_n^2)\sigma^2}\right)\right) p(x)
\]
\[
\approx \left(\frac{h_n^2}{2\sigma^2} - \frac{h_n^2}{2\sigma^2} \left(\frac{x - \mu}{\mu}\right)^2\right) p(x)
\]
\[
= \frac{h_n^2}{2\sigma^2} \left(1 - \left(\frac{x - \mu}{\mu}\right)^2\right) p(x)
\]

[Note this should strictly be done more carefully including order of expressions]

10. (a) The feature-space maps the variable length verification sequences into a fixed dimensionality. SVMs (empirically) generalise well for large dimensional feature spaces which will occur when \( M \) gets large. For the case given the dimensionality is \( M + (M(M+1))/2 \) [noting the symmetry in the second derivative (the function is twice differentiable and continuous)].

(b) We need
\[
\frac{\partial}{\partial \mu_i} \log \left(\prod_{t=1}^{T} \sum_{m=1}^{M} c_m \mathcal{N}(x_t; \mu_m, \sigma_m^2)\right) = \sum_{t=1}^{T} \frac{\partial}{\partial \mu_i} \log \left(\sum_{m=1}^{M} c_m \mathcal{N}(x_t; \mu_m, \sigma_m^2)\right)
\]

This was discussed in the Mixture Model lectures. This can be simply written as
\[
\frac{\partial}{\partial \mu_i} \log(P(X_1:T)) = \sum_{t=1}^{T} \frac{1}{p(x_t)} c_i \frac{\partial}{\partial \mu_i} \mathcal{N}(x_t; \mu_i, \sigma_i^2)
\]
\[
= \sum_{t=1}^{T} P(i|x_t) \frac{1}{\sigma_i^2} (x_t - \mu_i)
\]
(c) From part (b) (note it assumed that $i \neq j$)

$$\frac{\partial^2}{\partial \mu_j \partial \mu_i} \log(p(X_{1:T})) = \sum_{t=1}^{T} \frac{\partial}{\partial \mu_j} \left( P(i|x_t) \frac{1}{\sigma_i^2} (x_t - \mu_i) \right)$$

Only the posterior is a function of the mean of component $j$. This can be calculated

$$\frac{\partial}{\partial \mu_j} P(i|x_t) = -\frac{c_i N(x_t; \mu_i, \sigma_i^2)}{(p(x_t))^2} \frac{\partial}{\partial \mu_j} N(x_t; \mu_j, \sigma_j^2)$$

$$= -P(i|x_t)P(j|x_t) \frac{1}{\sigma_j^2} (x_t - \mu_j)$$

It is simple to see that the form in the question is simply obtained.

$$\frac{\partial^2}{\partial \mu_j \partial \mu_i} \log(p(X_{1:T})) = -\sum_{t=1}^{T} P(i|x_t)P(j|x_t) \frac{(x_t - \mu_j)(x_t - \mu_i)}{\sigma_i^2 \sigma_j^2}$$

These second order statistics have the potential for additional information as they are not a linear transform of the first order statistics. Furthermore it is not possible to obtain this form from a standard kernel operation on the first order statistics due to the summation over time.

Mark Gales
November 2003, 2007