Engineering Tripos Part IIB FOURTH YEAR

Paper 4F10: Statistical Pattern Processing

STATISTICAL PATTERN RECOGNITION

Solutions to Examples Paper 2

1. As the name implies linear classifiers only generate decision boundaries of the form $\mathbf{w}'\mathbf{x} + b = 0$. 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:

$$\begin{aligned} \alpha_i &\geq 0 \\ \sum_{i=1}^{m} \alpha_i y_i &= 0 \end{aligned}$$

   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

$$\begin{bmatrix} 1 \\ \sqrt{2} \\ -\sqrt{2} \\ -\sqrt{2} \\ 1 \\ 1 \end{bmatrix}; \quad \begin{bmatrix} 1 \\ -\sqrt{2} \\ +\sqrt{2} \\ -\sqrt{2} \\ 1 \\ 1 \end{bmatrix}; \quad \begin{bmatrix} 1 \\ \sqrt{2} \\ \sqrt{2} \\ \sqrt{2} \\ 1 \\ 1 \end{bmatrix}; \quad \begin{bmatrix} 1 \\ -\sqrt{2} \\ -\sqrt{2} \\ \sqrt{2} \\ 1 \\ 1 \end{bmatrix};$$

   The direction of the decision boundary is given by

$$\begin{aligned} \mathbf{w} &= \sum_{i=1}^{m} \alpha_i y_i \mathbf{\Phi}(\mathbf{x}_i) \\ &= \frac{1}{2}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ -\sqrt{2} \\ 0 \\ 0 \end{bmatrix}\right) \end{aligned}$$

To find b substitute into the expression

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

Select the point $[1, 1]'$

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

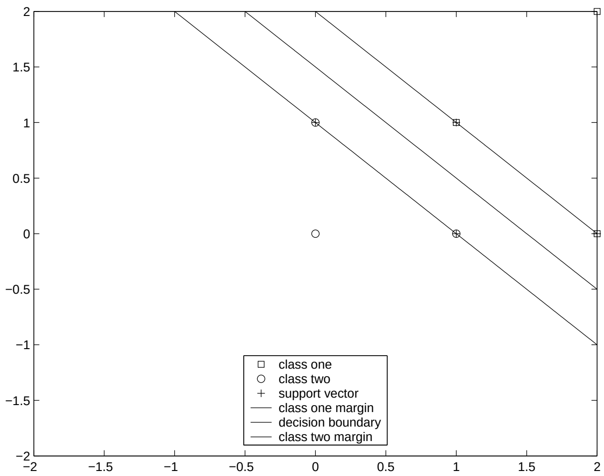
So $b = 0$. Check using the point $[1, -1]$

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

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

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

(c) There are multiple solutions for α (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 α, 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 $f(\mathbf{x}) + \epsilon \geq 0$. This is the same as the probability that $\epsilon \geq -\hat{\mathbf{w}}'\mathbf{x}$. Thus

$$P(y = +1|\mathbf{x}, \hat{\mathbf{w}}) = \int_{-\hat{\mathbf{w}}'\mathbf{x}}^{\infty} \mathcal{N}(\epsilon; 0, \sigma_{\mathtt{n}}^2) d\epsilon = 1 - \int_{-\infty}^{-\hat{\mathbf{w}}'\mathbf{x}} \mathcal{N}(\epsilon; 0, \sigma_{\mathtt{n}}^2) d\epsilon$$

Using the standard equalities

$$a = -\hat{\mathbf{w}}'\mathbf{x}/\sigma_{\mathtt{n}}$$

5. From the lecture notes

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}) &= \int p(\mathbf{y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}) d\mathbf{w} \\ &= \int \mathcal{N}(\mathbf{y}; \mathbf{X}'\mathbf{w}, \sigma_{\mathtt{n}}^2 \mathbf{I}) \mathcal{N}(\mathbf{w}; \mathbf{0}, \sigma_{\mathtt{w}}^2 \mathbf{I}) d\mathbf{w} \end{aligned}$$

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

$$\begin{aligned} &k \exp\left(-\frac{1}{2\sigma_{\mathtt{n}}^2}(\mathbf{y} - \mathbf{X}'\mathbf{w})'(\mathbf{y} - \mathbf{X}'\mathbf{w})) - \frac{1}{2\sigma_{\mathtt{w}}^2}\mathbf{w}'\mathbf{w}\right) = \\ &k \exp\frac{-1}{2}\left(\mathbf{w}'\left(\frac{1}{\sigma_{\mathtt{n}}^2}\mathbf{X}\mathbf{X}' + \frac{1}{\sigma_{\mathtt{w}}^2}\mathbf{I}\right)\mathbf{w} - \frac{2}{\sigma_{\mathtt{n}}^2}\mathbf{w}'\mathbf{X}\mathbf{y} + \frac{1}{\sigma_{\mathtt{n}}^2}\mathbf{y}'\mathbf{y}\right) \end{aligned}$$

Completing the square in $\mathbf{w}$, this can be expressed as

$$k \exp\frac{-1}{2}\left((\mathbf{w} - \boldsymbol{\mu}_w)'\boldsymbol{\Sigma}_w^{-1}(\mathbf{w} - \boldsymbol{\mu}_w) - \boldsymbol{\mu}_w'\boldsymbol{\Sigma}_w^{-1}\boldsymbol{\mu}_w + \frac{1}{\sigma_{\mathtt{n}}^2}\mathbf{y}'\mathbf{y}\right)$$

where

$$\begin{aligned} \boldsymbol{\mu}_w &= \frac{1}{\sigma_{\mathtt{n}}^2}\left(\frac{1}{\sigma_{\mathtt{n}}^2}\mathbf{X}\mathbf{X}' + \frac{1}{\sigma_{\mathtt{w}}^2}\mathbf{I}\right)^{-1}\mathbf{X}\mathbf{y} \\ \boldsymbol{\Sigma}_w &= \left(\frac{1}{\sigma_{\mathtt{n}}^2}\mathbf{X}\mathbf{X}' + \frac{1}{\sigma_{\mathtt{w}}^2}\mathbf{I}\right)^{-1} \end{aligned}$$

Integrating over $\mathbf{w}$ yields a constant and

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}) &\propto \exp\frac{-1}{2}\left(\mathbf{y}'\left(\frac{1}{\sigma_{\mathtt{n}}^2}\mathbf{I} - \frac{1}{\sigma_{\mathtt{n}}^2}\mathbf{X}'\left(\frac{1}{\sigma_{\mathtt{n}}^2}\mathbf{X}\mathbf{X}' + \frac{1}{\sigma_{\mathtt{w}}^2}\mathbf{I}\right)^{-1}\mathbf{X}\frac{1}{\sigma_{\mathtt{n}}^2}\right)\mathbf{y}\right) \\ &= \mathcal{N}(\mathbf{y}; \mathbf{0}, \sigma_{\mathtt{w}}^2\mathbf{X}'\mathbf{X} + \sigma_{\mathtt{n}}^2\mathbf{I}) \end{aligned}$$

Using the equality given in the question.

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

$$\left[\begin{array}{c} f(\tilde{\mathbf{x}}) \\ \mathbf{y} \end{array}\right] \sim \mathcal{N}\left(\mathbf{0}, \left[\begin{array}{cc} k(\tilde{\mathbf{x}},\tilde{\mathbf{x}}) & \mathbf{k}(\tilde{\mathbf{x}},\mathbf{X})' \\ \mathbf{k}(\tilde{\mathbf{x}},\mathbf{X}) & \mathbf{K}(\mathbf{X},\mathbf{X})+\sigma_{\mathtt{n}}^2\mathbf{I} \end{array}\right]\right)$$

The conditional distribution can then be written as

$$f(\tilde{\mathbf{x}})|\mathbf{X},\mathbf{y},\tilde{\mathbf{x}} \sim \mathcal{N}(\mu_{\mathtt{f}},\sigma_{\mathtt{f}}^2)$$

where

$$\begin{aligned} \mu_{\mathtt{f}} &= \mathbf{k}(\tilde{\mathbf{x}},\mathbf{X})'\left[\mathbf{K}(\mathbf{X},\mathbf{X})+\sigma_{\mathtt{n}}^2\mathbf{I}\right]^{-1}\mathbf{y} \\ \mathrm{var}_n(f(\tilde{\mathbf{x}})) &= \sigma_{\mathtt{f}}^2 = k(\tilde{\mathbf{x}},\tilde{\mathbf{x}}) - \mathbf{k}(\tilde{\mathbf{x}},\mathbf{X})'\left[\mathbf{K}(\mathbf{X},\mathbf{X})+\sigma_{\mathtt{n}}^2\mathbf{I}\right]^{-1}\mathbf{k}(\tilde{\mathbf{x}},\mathbf{X}) \end{aligned}$$

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

$$\left[\mathbf{K}(\mathbf{X},\mathbf{X})+\sigma_{\mathtt{n}}^2\mathbf{I}\right]^{-1} = \left[\begin{array}{cc} \mathbf{K}(\bar{\mathbf{X}},\bar{\mathbf{X}})+\sigma_{\mathtt{n}}^2\mathbf{I} & \mathbf{k}(\mathbf{x}_n,\bar{\mathbf{X}}) \\ \mathbf{k}(\mathbf{x}_n,\bar{\mathbf{X}})' & k(\mathbf{x}_n,\mathbf{x}_n)+\sigma_{\mathtt{n}}^2 \end{array}\right]^{-1} = \left[\begin{array}{cc} \mathbf{A} & \mathbf{b} \\ \mathbf{b}' & c \end{array}\right]^{-1} = \left[\begin{array}{cc} \mathbf{E} & \mathbf{g} \\ \mathbf{g}' & h \end{array}\right]$$

and

$$\mathbf{k}(\tilde{\mathbf{x}},\mathbf{X}) = \left[\begin{array}{c} \mathbf{k}(\tilde{\mathbf{x}},\bar{\mathbf{X}}) \\ \mathbf{k}(\tilde{\mathbf{x}},\mathbf{x}_n) \end{array}\right]$$

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

$$d_n = \mathbf{k}(\tilde{\mathbf{x}},\bar{\mathbf{X}})'\mathbf{E}\mathbf{k}(\tilde{\mathbf{x}},\bar{\mathbf{X}}) + 2k(\tilde{\mathbf{x}},\mathbf{x}_n)\mathbf{g}'\mathbf{k}(\tilde{\mathbf{x}},\bar{\mathbf{X}}) + k(\tilde{\mathbf{x}},\mathbf{x}_n)hk(\tilde{\mathbf{x}},\mathbf{x}_n)$$

and for the $n-1$ samples

$$d_{n-1} = \mathbf{k}(\tilde{\mathbf{x}},\bar{\mathbf{X}})'(\mathbf{K}(\bar{\mathbf{X}},\bar{\mathbf{X}})+\sigma_{\mathtt{n}}^2\mathbf{I})^{-1}\mathbf{k}(\tilde{\mathbf{x}},\bar{\mathbf{X}})$$

Consider the form for the partitioned n example case. Let

$$k_a = \mathbf{k}(\tilde{\mathbf{x}},\bar{\mathbf{X}})'\mathbf{A}^{-1}\mathbf{b}$$

and

$$k_b = \frac{1}{k(\mathbf{x}_n,\mathbf{x}_n)+\sigma_{\mathtt{n}}^2 - \mathbf{k}(\mathbf{x}_n,\bar{\mathbf{X}})'\left(\mathbf{K}(\bar{\mathbf{X}},\bar{\mathbf{X}})+\sigma_{\mathtt{n}}^2\mathbf{I}\right)^{-1}\mathbf{k}(\mathbf{x}_n,\bar{\mathbf{X}})}$$

Then

$$\begin{aligned} d_n &= d_{n-1} + k_a k_b k_a - 2k(\tilde{\mathbf{x}},\mathbf{x}_n)k_b k_a + k(\tilde{\mathbf{x}},\mathbf{x}_n)k_b k(\tilde{\mathbf{x}},\mathbf{x}_n) \\ &= d_{n-1} + (k(\tilde{\mathbf{x}},\mathbf{x}_n) - k_a)^2 k_b \end{aligned}$$

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

$$d_n \geq d_{n-1} \quad \text{therefore} \quad \mathrm{var}_n(f(\tilde{\mathbf{x}})) \leq \mathrm{var}_{n-1}(f(\tilde{\mathbf{x}}))$$

7. Using the equality from question (6), the following equality (assuming that $i = n$)

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

Examining the variance term and substituitin in $\lambda_i \to \infty$

$$\begin{aligned} \mathbf{\Sigma}_w &= \left(\frac{1}{\sigma_{\mathrm{n}}^2}\mathbf{\Phi}'\mathbf{\Phi} + \mathbf{\Lambda}\right)^{-1} \\ &= \begin{bmatrix} \mathbf{A}^{-1} & \mathbf{0} \\ \mathbf{0}' & 0 \end{bmatrix} \end{aligned}$$

Thus the variance of elemene i goes towards zero. For the mean

$$\begin{aligned} \boldsymbol{\mu}_w &= \frac{1}{\sigma_{\mathrm{n}}^2}\mathbf{\Sigma}_w\mathbf{\Phi}'\boldsymbol{y} \\ &= \frac{1}{\sigma_{\mathrm{n}}^2}\begin{bmatrix} \mathbf{A}^{-1} & \mathbf{0} \\ \mathbf{0}' & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_1)'\mathbf{y} \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}_n)'\mathbf{y} \end{bmatrix} \\ &= \frac{1}{\sigma_{\mathrm{n}}^2}\begin{bmatrix} \mathbf{A}^{-1}\begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_1)'\mathbf{y} \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}_{n-1})'\mathbf{y} \end{bmatrix} \\ 0 \end{bmatrix} \end{aligned}$$

It is clear that all elements associated with $\mathbf{x}_n$ 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 ω_1 for the root node

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

The left node of the hypothesised split

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

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

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

The entropy cost function can then be written as

$$\mathcal{I}(N_L) = P(\omega_1|N_L)\log(P(\omega_1|N_L)) + (1 - P(\omega_1|N_L))\log(1 - P(\omega_1|N_L))$$

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

9. (a)

$$\begin{aligned}
\mathcal{E}\{\tilde{p}(x)\} &= \mathcal{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\phi\left(\frac{x-x_i}{h_n}\right)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\mathcal{E}\left\{\phi\left(\frac{x-x_i}{h_n}\right)\right\} \\
&= \frac{1}{h_n}\mathcal{E}\left\{\phi\left(\frac{x-x_i}{h_n}\right)\right\}
\end{aligned}$$

As $\phi()$ is Gaussian distributed then

$$\begin{aligned}
\mathcal{E}\left\{\mathcal{N}\left(\frac{x-x_i}{h_n};0,1\right)\right\} &= \int \mathcal{N}\left(\left(\frac{x-v}{h_n}\right);0,1\right)\mathcal{N}(v;\mu,\sigma^2)dv \\
&= \int h_n\mathcal{N}(x;v,h_n^2)\mathcal{N}(v;\mu,\sigma^2)dv \\
&= h_n\mathcal{N}(x;\mu,\sigma^2+h_n^2)
\end{aligned}$$

Hence

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

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

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

As h_n gets small

$$\mathrm{var}[\tilde{p}(x)] \approx \frac{1}{2nh_n\sqrt{\pi}}p(x)$$

(c)

$$\begin{aligned}
p(x) - \mathcal{E}\{\tilde{p}(x)\} &= \mathcal{N}(x;\mu,\sigma^2) - \mathcal{N}(x;\mu,\sigma^2+h_n^2) \\
&= \left(1 - \frac{\mathcal{N}(x;\mu,\sigma^2+h_n^2)}{\mathcal{N}(x;\mu,\sigma^2)}\right)\mathcal{N}(x;\mu,\sigma^2) \\
&= \left(1 - \sqrt{\frac{\sigma^2}{\sigma^2+h_n^2}}\exp\left(\frac{h_n^2(x-\mu)^2}{2(\sigma^2+h_n^2)\sigma^2}\right)\right)\mathcal{N}(x;\mu,\sigma^2)
\end{aligned}$$

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

$$\begin{aligned}
p(x) - \mathcal{E}\{\tilde{p}(x)\} &\approx \left(1 - \left(1 - \frac{h_n^2}{2(\sigma^2+h_n^2)}\right)\left(1 + \frac{h_n^2(x-\mu)^2}{2(\sigma^2+h_n^2)\sigma^2}\right)\right)p(x) \\
&\approx \left(\frac{h_n^2}{2\sigma^2} - \frac{h_n^2}{2\sigma^2}\left(\frac{x-\mu}{\mu}\right)^2\right)p(x) \\
&= \frac{h_n^2}{2\sigma^2}\left(1 - \left(\frac{(x-\mu)}{\mu}\right)^2\right)p(x)
\end{aligned}$$

[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

$$\begin{aligned}
\frac{\partial}{\partial\mu_i}\log(P(\mathbf{X}_{1:T})) &= \sum_{t=1}^{T}\frac{1}{p(x_t)}c_i\frac{\partial}{\partial\mu_i}\mathcal{N}(x_t;\mu_i,\sigma_i^2) \\
&= \sum_{t=1}^{T}P(i|x_t)\frac{1}{\sigma_i^2}(x_t-\mu_i)
\end{aligned}$$

(c) From part (b) (note it assumed that $i \neq j$)

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

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

$$\begin{aligned}\frac{\partial}{\partial\mu_j}P(i|x_t) &= -\frac{c_i\mathcal{N}(x_t;\mu_i,\sigma_i^2)}{(p(x_t))^2}c_j\frac{\partial}{\partial\mu_j}\mathcal{N}(x_t;\mu_j,\sigma_j^2)\\ &= -P(i|x_t)P(j|x_t)\frac{1}{\sigma_j^2}(x_t-\mu_j)\end{aligned}$$

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

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

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