Engineering Tripos Part IIB FOURTH YEAR

Paper 4F10: Statistical Pattern Processing

STATISTICAL PATTERN RECOGNITION

Solutions to Examples Paper 2

1. Total number of weights in the system is

   - input to hidden layer: $(d+1)M$
   - the $L-1$ hidden to hidden: $(L-1)M(M+1)$
   - hidden to output $(M+1)K$

   The number of hidden layers determines the decision boundaries that can be produced (see lecture notes), the activation function determines the nature of the output - binary (step), sum to one (soft max), continuous (linear) etc. The number of hidden units should be large enough to model the problem, but small enough so that *generalisation* is not an issue.

2. $$\phi(z) = \frac{1}{1+\exp(-z)}, \quad \begin{aligned} \frac{\partial \phi(z)}{\partial z} &= \frac{\exp(-z)}{(1+\exp(-z))^2} \\ &= \phi(z)(1-\phi(z)) \end{aligned}$$

   The activation function affects the form of the error back propagation algorithm. The derivation given in lectures assumes a sigmoid, however the output layer can be more complex if a sum squared error is used with a softmax function (not not if used with a cross-entropy measure) since in this case the partial derivative for a particular weight in the output layer depends on all output values due to the normalisation in the softmax.

3. Need to compute the first and second moments. First moment given by

   $$\int_{-\infty}^{\infty} \phi(x)p(x)dx = \alpha \int_{-\infty}^{0} xp(x)dx + \int_{0}^{\infty} xp(x)dx = \int_{0}^{\infty} xp(x)dx - \alpha \int_{0}^{\infty} xp(x)dx$$

   It is possible to show that when $p(x) = \mathcal{N}(x; 0, \sigma^2)$

   $$\int_{0}^{\infty} xp(x)dx = \frac{\sigma}{2}\sqrt{\frac{2}{\pi}}$$

   Hence

   $$\int_{-\infty}^{\infty} \phi(x)p(x)dx = (1-\alpha)\frac{\sigma}{2}\sqrt{\frac{2}{\pi}} = (1-\alpha)\sigma\sqrt{\frac{1}{2\pi}}$$

and the second moment

$$\int_{-\infty}^{\infty}(\phi(x))^2 p(x)dx = \int_{-\infty}^{0}\alpha^2 x^2 p(x)dx + \int_{0}^{\infty} x^2\phi(x)p(x)dx = (1+\alpha^2)\sigma^2/2$$

So the total variance on the output is

$$\hat{\sigma}^2 = (1+\alpha^2)\sigma^2/2 - (1-\alpha)^2\frac{\sigma^2}{2\pi}$$

The simplest approcah is to ensure that the output variance matches the input variances for the initialisation (as discussed in lectures). This function has an added complexity as the mean is non-zero. If the network is deep this could result in a large offset for some layers. This could be addressed by consifdering an offset on the bias term initialisation, but is usually ignored.

4. If the gradient is approximately constant then we can write

$$\boldsymbol{\Delta}\mathbf{w}[\tau] = -\eta\,\boldsymbol{\nabla}E(\mathbf{w})|_{\mathbf{w}[0]} + \alpha\boldsymbol{\Delta}\mathbf{w}[\tau-1]$$

Substituting back yields

$$\begin{aligned}\boldsymbol{\Delta}\mathbf{w}[\tau] &= -\eta\,\boldsymbol{\nabla}E(\mathbf{w})|_{\mathbf{w}[0]} + \alpha\left(-\eta\,\boldsymbol{\nabla}E(\mathbf{w})|_{\mathbf{w}[0]} + (....)\right)\\ &= -\eta\left(1+\alpha+\alpha^2+...\right)\boldsymbol{\nabla}E(\mathbf{w})|_{\mathbf{w}[0]}\end{aligned}$$

If $\alpha < 1$ then the sum of the infinite GP give

$$\boldsymbol{\Delta}\mathbf{w}[\tau] = -\left(\frac{\eta}{1-\alpha}\right)\boldsymbol{\nabla}E(\mathbf{w})|_{\mathbf{w}[0]}$$

If the solution is oscillating then we can write (approximately)

$$\begin{aligned}\boldsymbol{\Delta}\mathbf{w}[\tau] &= -\eta\,\boldsymbol{\nabla}E(\mathbf{w})|_{\mathbf{w}[0]} + \alpha\left(+\eta\,\boldsymbol{\nabla}E(\mathbf{w})|_{\mathbf{w}[0]} + (....)\right)\\ &= -\eta\left(1-\alpha+\alpha^2-...\right)\boldsymbol{\nabla}E(\mathbf{w})|_{\mathbf{w}[0]}\\ &= -\eta\left((1-\alpha)(1+\alpha^2+\alpha^4...)\right)\boldsymbol{\nabla}E(\mathbf{w})|_{\mathbf{w}[0]}\\ &= -\eta\frac{(1-\alpha)}{(1-\alpha^2)}\,\boldsymbol{\nabla}E(\mathbf{w})|_{\mathbf{w}[0]}\\ &= -\left(\frac{\eta}{1+\alpha}\right)\boldsymbol{\nabla}E(\mathbf{w})|_{\mathbf{w}[0]}\end{aligned}$$

5. (a) The Hessian may be used to obtain the Newton direction. This requires computing $\mathbf{H}^{-1}\mathbf{g}$. (see lecture notes for more details).

(b)

$$\frac{\partial E}{\partial w_{ij}} = \sum_{p=1}^{n} (y(x_p) - t(x_p)) \frac{\partial y(x_p)}{\partial w_{ij}}$$

and

$$\frac{\partial^2 E}{\partial w_{ij} \partial w_{lk}} = \sum_{p=1}^{n} \frac{\partial y(x_p)}{\partial w_{lk}} \frac{\partial y(x_p)}{\partial w_{ij}} + \sum_{p=1}^{n} (y(x_p) - t(x_p)) \frac{\partial^2 y(x_p)}{\partial w_{ij} \partial w_{lk}}$$

For the conditions described the network will train so that

$$y(x_p) = t(x_p)$$

In this condition the second term is zero.

(c) From the conditions given

$$\mathbf{H}_{N+1} = \mathbf{H}_N + \mathbf{g}^{(N+1)} (\mathbf{g}^{(N+1)})'$$

Consider the inverse

$$\begin{aligned}
\mathbf{H}_{N+1}^{-1} &= \left(\mathbf{H}_N + \mathbf{g}^{(N+1)} (\mathbf{g}^{(N+1)})' \right)^{-1} \\
&= \mathbf{H}_N^{-1} - \mathbf{H}_N^{-1} \mathbf{g}^{(N+1)} \left(1 + \mathbf{g}^{(N+1)\prime} \mathbf{H}_N^{-1} (\mathbf{g}^{(N+1)}) \right)^{-1} (\mathbf{g}^{(N+1)})' \mathbf{H}_N^{-1} \\
&= \mathbf{H}_N^{-1} - \frac{\mathbf{H}_N^{-1} \mathbf{g}^{(N+1)} (\mathbf{g}^{(N+1)})' \mathbf{H}_N^{-1}}{1 + \mathbf{g}^{(N+1)\prime} \mathbf{H}_N^{-1} (\mathbf{g}^{(N+1)})}
\end{aligned}$$

The calculation of the inverse can be computationally expensive for large numbers of weights (naive implementation $\mathcal{O}(W^3)$). This scheme directly calculates the inverse. An initial value is needed for this scheme ($\mathbf{H}_0$). The simplest approach is to use a diagonal matrix with very small values on the leading diagonal (easy to invert and will not distort the final results).

6. 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.)

7. 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 $[1, 1]'$

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

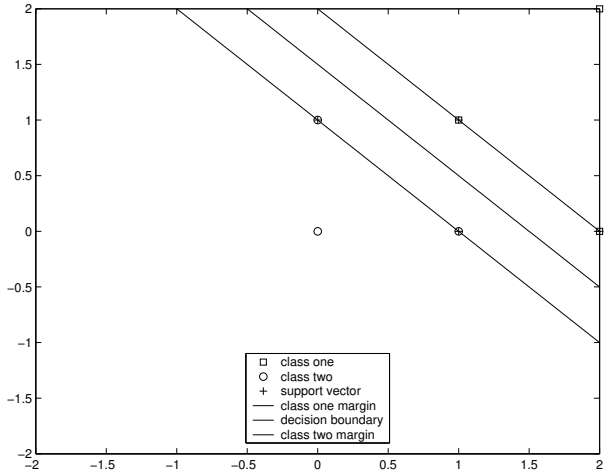
So $b = 0$. Check using the point $[1, -1]$

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

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

$$x_1 x_2 = 0$$

8. (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.

9. 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.

10. (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

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

As h_n gets small

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

(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]

11. (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