King's College London

UNIVERSITY OF LONDON

ATTACH THIS PAPER TO YOUR SCRIPT **USING THE STRING PROVIDED**

Candidate No: .... **Desk No:** ....

MSC EXAMINATION

7CCMNN15 (CMNN15) ADVANCED NEURAL NETWORKS

SUMMER 2010

TIME ALLOWED: TWO HOURS

ALL QUESTIONS CARRY EQUAL MARKS. FULL MARKS WILL BE AWARDED FOR COMPLETE ANSWERS TO THREE QUESTIONS. ONLY THE BEST THREE QUESTIONS WILL COUNT TOWARDS GRADES A OR B, BUT CREDIT WILL BE GIVEN FOR ALL WORK DONE FOR LOWER GRADES.
FIGURES IN SQUARE BRACKETS GIVE AN INDICATION OF THE NUMBER OF POINTS PER SECTION.

NO CALCULATORS ARE PERMITTED.

DO NOT REMOVE THIS PAPER FROM THE EXAMINATION ROOM

TURN OVER WHEN INSTRUCTED

1. We consider unsupervised competitive learning processes in which N code-book vectors $\boldsymbol{m}_i \in \mathbb{R}^n$ evolve stochastically according to

$$\boldsymbol{m}_i(\ell{+}1) = \boldsymbol{m}_i(\ell) + \eta F_i[\boldsymbol{x}(\ell),\{\boldsymbol{m}(\ell)\}]\ (\boldsymbol{x}(\ell) - \boldsymbol{m}_i(\ell))$$

Here $\eta > 0$ and $\{\boldsymbol{m}(\ell)\}$ is a shorthand for $\boldsymbol{m}_1(\ell), \ldots, \boldsymbol{m}_N(\ell)$. The data vectors $\boldsymbol{x}(\ell) \in \mathbb{R}^n$ are drawn independently at random at each step $\ell = 0, 1, 2, \ldots$ according to a probability density $p(\boldsymbol{x})$.

(a) [**10 points**] Define Vector Quantization (VQ) in terms of Voronoi tessellations and show that it is of the form above.

(b) [**10 points**] Give the form of the function $F_i[\boldsymbol{x},\{\boldsymbol{m}\}]$ for Soft Vector Quantization (SVQ). Show that in an appropriate limit this reduces to the corresponding function for VQ.

(c) If we define a normalized time $t = \eta\ell$ then for $\eta \to 0$ the above discrete-time process reduces to the coupled deterministic equations

$$\frac{d}{dt}\boldsymbol{m}_i = \int d\boldsymbol{x}\ p(\boldsymbol{x})\ F_i[\boldsymbol{x},\{\boldsymbol{m}\}]\ (\boldsymbol{x} - \boldsymbol{m}_i) \qquad (*)$$

(c1) [**12 points**] Show that for SVQ the equations $(*)$ are of gradient form, i.e. that for an appropriate function $E[\{\boldsymbol{m}\}]$ we have $d\boldsymbol{m}_i/dt = -\boldsymbol{\nabla}_{\boldsymbol{m}_i} E[\{\boldsymbol{m}\}]$. The function E should depend on a trial distribution $q(\boldsymbol{x})$ that is determined by the code-book vector positions. Describe (without proofs) the meaning of the function E.

(c2) [**8 points**] Now consider exponentially weighted SVQ, defined by

$$F_i[\boldsymbol{x},\{\boldsymbol{m}\}] = \frac{e^{w_i - \beta|\boldsymbol{x}-\boldsymbol{m}_i|^2}}{\sum_{j=1}^N e^{w_j - \beta|\boldsymbol{x}-\boldsymbol{m}_j|^2}}$$

for $\beta > 0$ and weights w_i, $i = 1, \ldots, N$. Show that also for this case the equations $(*)$ are of gradient form, where in the trial distribution $q(\boldsymbol{x})$ each code-book vector contributes with weight $e^{w_i}/\sum_{j=1}^N e^{w_j}$.

(c3) [**10 points**] Assume that now also the weights w_i are allowed to evolve, by gradient descent $dw_i/dt = -\partial E/\partial w_i$ on the function E for weighted SVQ from (c2). Find an explicit expression for dw_i/dt and show that, if the process reaches a stationary state, then $\int d\boldsymbol{x}\, p(\boldsymbol{x}) F_i[\boldsymbol{x},\{\boldsymbol{m}\}] = e^{w_i}/\sum_{j=1}^N e^{w_j}$.

2. We consider Bayesian regression. A neural network produces an output $t \in \mathbb{R}$ for every input vector $\boldsymbol{\xi} \in \mathbb{R}^N$, subject to zero mean additive noise. It is parametrized by a weight vector $\boldsymbol{w} \in \mathbb{R}^M$, such that

$$p(t|\boldsymbol{\xi}, \boldsymbol{w}) = P(t - f(\boldsymbol{\xi}, \boldsymbol{w}))$$

where $P(z)$ is some probability distribution over z with zero mean and variance σ^2. The data used in training this system consist of p pairs of inputs and corresponding outputs, $D = \{(\boldsymbol{\xi}^1, t_1), \ldots, (\boldsymbol{\xi}^p, t_p)\}$.

(a) [**16 points**] Give an expression for the predictive distribution $p(t|\boldsymbol{\xi}, D)$, in terms of the posterior $p(\boldsymbol{w}|D)$. Derive the following expressions for the predictive mean $t^\star(\boldsymbol{\xi}) = \int dt\, t\, p(t|\boldsymbol{\xi}, D)$ and variance $[\Delta t^\star(\boldsymbol{\xi})]^2 = \int dt\, t^2\, p(t|\boldsymbol{\xi}, D) - [\int dt\, t\, p(t|\boldsymbol{\xi}, D)]^2$:

$$\begin{aligned} t^\star(\boldsymbol{\xi}) &= \int d\boldsymbol{w}\ f(\boldsymbol{\xi}, \boldsymbol{w}) p(\boldsymbol{w}|D) \\ [\Delta t^\star(\boldsymbol{\xi})]^2 &= \sigma^2 + \int d\boldsymbol{w}\ f^2(\boldsymbol{\xi}, \boldsymbol{w}) p(\boldsymbol{w}|D) - \left[\int d\boldsymbol{w}\ f(\boldsymbol{\xi}, \boldsymbol{w}) p(\boldsymbol{w}|D)\right]^2 \end{aligned}$$

(b) Now consider radial basis function networks with $f(\boldsymbol{\xi}, \boldsymbol{w}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{\xi}) = \sum_{i=1}^M w_i \phi_i(\boldsymbol{\xi})$, where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_M)$ is a vector of M basis functions. Also assume a Gaussian prior on $\boldsymbol{w}$, $p(\boldsymbol{w}) = (2\pi)^{-M/2} (\det \boldsymbol{C})^{-1/2} e^{-\boldsymbol{w} \cdot \boldsymbol{C}^{-1} \boldsymbol{w}/2}$ with a covariance matrix $\boldsymbol{C}$. Finally, let the noise be Gaussian, $P(z) = (2\pi\sigma^2)^{-1/2} \exp[-z^2/(2\sigma^2)]$.

(b1) [**20 points**] Show that the posterior distribution $p(\boldsymbol{w}|D)$ is a Gaussian with means and covariances (abbreviating $\langle \ldots \rangle = \int d\boldsymbol{w}\ \ldots\ p(\boldsymbol{w}|D)$)

$$\langle \boldsymbol{w} \rangle = \boldsymbol{A}^{-1} \boldsymbol{c}, \quad \boldsymbol{c} = \sigma^{-2} \sum_{\mu=1}^p t_\mu \boldsymbol{\phi}(\boldsymbol{\xi}^\mu) \qquad \langle w_i w_j \rangle - \langle w_i \rangle \langle w_j \rangle = (\boldsymbol{A}^{-1})_{ij}$$

where $\boldsymbol{A}$ is an $M \times M$ matrix with elements

$$A_{ij} = (\boldsymbol{C}^{-1})_{ij} + \sigma^{-2} \sum_{\mu=1}^p \phi_i(\boldsymbol{\xi}^\mu) \phi_j(\boldsymbol{\xi}^\mu)\ .$$

(b2) [**14 points**] Use the results of (a) and (b1) to derive

$$t^\star(\boldsymbol{\xi}) = \boldsymbol{\phi}(\boldsymbol{\xi}) \cdot \boldsymbol{A}^{-1} \boldsymbol{c}, \qquad \Delta t^\star(\boldsymbol{\xi}) = \sqrt{\sigma^2 + \boldsymbol{\phi}(\boldsymbol{\xi}) \cdot \boldsymbol{A}^{-1} \boldsymbol{\phi}(\boldsymbol{\xi})}$$

You may, if you wish, use without proof the identity (where $\boldsymbol{A}$ is a symmetric and positive definite matrix):

$$\frac{\int d\boldsymbol{u}\ u_i u_j e^{-\frac{1}{2}\boldsymbol{u} \cdot \boldsymbol{A} \boldsymbol{u}}}{\int d\boldsymbol{u}\ e^{-\frac{1}{2}\boldsymbol{u} \cdot \boldsymbol{A} \boldsymbol{u}}} = (\boldsymbol{A}^{-1})_{ij}$$

3. We consider Bayesian classification. A neural network produces a binary output $t \in \{-1, 1\}$ for every input vector $\boldsymbol{\xi} \in \mathbb{R}^N$. It implements a noisy classifier parametrized by a weight vector $\boldsymbol{w} \in \mathbb{R}^N$, such that

$$p(t|\boldsymbol{\xi}, \boldsymbol{w}) = \frac{1}{2}[1 + t\, g(\boldsymbol{\xi}, \boldsymbol{w})]$$

for a suitable function $g(\boldsymbol{\xi}, \boldsymbol{w})$. The data used in training this system consist of p pairs of inputs and corresponding outputs: $D = \{(\boldsymbol{\xi}^1, t_1), \ldots, (\boldsymbol{\xi}^p, t_p)\}$.

(a1) [**6 points**] Give an expression for $p(t|\boldsymbol{\xi}, D)$, the conditional output distribution given the data D, in terms of $p(\boldsymbol{w}|D)$.

(a2) [**6 points**] Assume that the network prediction $t^\star(\boldsymbol{\xi})$ for the classification of $\boldsymbol{\xi}$ and its uncertainty $\Delta t^\star(\boldsymbol{\xi})$ are defined as usual:

$$t^\star(\boldsymbol{\xi}) = \begin{cases} 1 & \text{if } p(1|\boldsymbol{\xi}, D) > 1/2 \\ -1 & \text{if } p(-1|\boldsymbol{\xi}, D) > 1/2 \end{cases} \qquad \Delta t^\star(\boldsymbol{\xi}) = \begin{cases} p(-1|\boldsymbol{\xi}, D) & \text{if } t^\star(\boldsymbol{\xi}) = 1 \\ p(1|\boldsymbol{\xi}, D) & \text{if } t^\star(\boldsymbol{\xi}) = -1 \end{cases}$$

Explain the precise meaning of $\Delta t^\star(\boldsymbol{\xi})$.

(a3) [**10 points**] Prove the following statements:

$$t^\star(\boldsymbol{\xi}) = \text{sgn}(I(\boldsymbol{\xi}, D)) \qquad \Delta t^\star(\boldsymbol{\xi}) = \frac{1}{2} - \frac{1}{2}|I(\boldsymbol{\xi}, D)|$$

in which $I(\boldsymbol{\xi}, D) = \int d\boldsymbol{w}\, g(\boldsymbol{w} \cdot \boldsymbol{\xi})\, p(\boldsymbol{w}|D)$.

(b) [**6 points**] Assume now that $g(\boldsymbol{\xi}, \boldsymbol{w}) = (1 - 2\epsilon)\text{sgn}(\boldsymbol{\xi} \cdot \boldsymbol{w})$, with $0 \leq \epsilon \leq \frac{1}{2}$. Show that for $\epsilon = 0$ we have a noise-free classifier $t = \text{sgn}(\boldsymbol{\xi} \cdot \boldsymbol{w})$, and hence give an interpretation of ϵ in terms of noise strength.

(c) Let g be as in (b), and assume further that $\boldsymbol{w}, \boldsymbol{\xi} \in \mathbb{R}^2$ with $|\boldsymbol{w}| = |\boldsymbol{\xi}| = 1$. Let input-output pairs be parameterized as $t_\mu \boldsymbol{\xi}^\mu = (\cos(\phi_\mu), \sin(\phi_\mu))$ and the weight vector as $\boldsymbol{w} = (\cos(\omega), \sin(\omega))$, with all angles in the range $[0, 2\pi)$. Assume a uniform prior over ω, $p(\omega) = 1/(2\pi)$, and consider a data set D of $p = 2$ examples with $\phi_1 = 0$ and $\phi_2 = \pi/2$.

(c1) [**8 points**] For $\epsilon = 0$, show that the posterior is

$$p(\omega|D) = \begin{cases} 2/\pi & \text{for } \ 0 < \omega < \pi/2 \\ 0 & \text{for } \ \pi/2 < \omega < 2\pi \end{cases}$$

Hint: Use that $t_\mu \text{sgn}(\boldsymbol{\xi}^\mu \cdot \boldsymbol{w}) = \text{sgn}(\cos(\phi_\mu - \omega))$.

(c2) [**12 points**] For $\epsilon = 0$ and a test input-output pair parameterized as $t\boldsymbol{\xi} = (\cos(\phi), \sin(\phi))$, show that the predictive distribution has the form

$$p(t|\boldsymbol{\xi}, D) = \begin{cases} 1 & \text{for} \quad 0 < \phi < \pi/2 \\ \frac{2}{\pi}(\pi - \phi) & \text{for} \quad \pi/2 < \phi < \pi \\ 0 & \text{for} \quad \pi < \phi < 3\pi/2 \\ \frac{2}{\pi}(\phi - 3\pi/2) & \text{for} \quad 3\pi/2 < \phi < 2\pi \end{cases}$$

[**2 points**] Explain why, even though $\epsilon = 0$, this does not have the form of a noise-free classifier where $p(t|\boldsymbol{\xi}, D) \in \{0, 1\}$ everywhere.

4. Consider a zero-mean Gaussian process with covariance function $C(\boldsymbol{\xi}, \boldsymbol{\xi}')$. The clean outputs y_μ corresponding to fixed training outputs $\boldsymbol{\xi}^\mu$ ($\mu = 1, \ldots, p$) then have joint distribution

$$p(\boldsymbol{y}) = (2\pi)^{-p/2}(\det \boldsymbol{C})^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{y} \cdot \boldsymbol{C}^{-1}\boldsymbol{y}\right)$$

where the matrix $\boldsymbol{C}$ has entries $C_{\mu\nu} = C(\boldsymbol{\xi}^\mu, \boldsymbol{\xi}^\nu)$ and $\boldsymbol{y} = (y_1, \ldots, y_p)$. Each clean output is corrupted independently by noise. Given the clean output, the noisy output t_μ has distribution $p(t_\mu|y_\mu) = (2\pi\sigma^2)^{-1/2} \exp[-(t_\mu - y_\mu)^2/(2\sigma^2)]$.

You may use throughout that for any symmetric and positive definite $n \times n$ matrix $\boldsymbol{A}$ and for $\boldsymbol{w}, \boldsymbol{u} \in \mathbb{R}^n$

$$\int d\boldsymbol{w}\; e^{-\frac{1}{2}\boldsymbol{w}\cdot\boldsymbol{A}\boldsymbol{w}+i\boldsymbol{w}\cdot\boldsymbol{u}} = \left[\frac{(2\pi)^n}{\det \boldsymbol{A}}\right]^{1/2} e^{-\frac{1}{2}\boldsymbol{u}\cdot\boldsymbol{A}^{-1}\boldsymbol{u}} \qquad (*)$$

(a) [**4 points**] Explain why the joint distribution of the noisy outputs $\boldsymbol{t} = (t_1 \ldots t_p)$ is given by $p(\boldsymbol{t}) = \int d\boldsymbol{y}\, p(\boldsymbol{y}) \prod_{\mu=1}^{p} p(t_\mu|y_\mu)$.

[**4 points**] Use the relation $(*)$ to show that

$$p(t_\mu|y_\mu) = \int \frac{dk_\mu}{2\pi} e^{-ik_\mu(t_\mu - y_\mu) - \sigma^2 k_\mu^2/2}$$

(b) [**6 points**] Show that, with $\boldsymbol{k} = (k_1 \ldots k_p)$,

$$p(\boldsymbol{t}) = \int d\boldsymbol{k}\, d\boldsymbol{y}\, \frac{\exp[-i\boldsymbol{k}\cdot(\boldsymbol{t}-\boldsymbol{y}) - \frac{\sigma^2}{2}\boldsymbol{k}^2 - \frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}\boldsymbol{y}]}{(2\pi)^p(2\pi)^{p/2}(\det \boldsymbol{C})^{1/2}}$$

[**8 points**] Use $(*)$ to show that

$$p(\boldsymbol{t}) = \int d\boldsymbol{k}\, \frac{\exp[-i\boldsymbol{k}\cdot\boldsymbol{t} - \frac{\sigma^2}{2}\boldsymbol{k}^2 - \frac{1}{2}\boldsymbol{k}\cdot\boldsymbol{C}\boldsymbol{k}]}{(2\pi)^p}$$

[**8 points**] Use $(*)$ again to deduce

$$p(\boldsymbol{t}) = (2\pi)^{-p/2}(\det \boldsymbol{K})^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{t}\cdot\boldsymbol{K}^{-1}\boldsymbol{t}\right)$$

where the matrix $\boldsymbol{K}$ has entries $K_{\mu\nu} = C_{\mu\nu} + \sigma^2\delta_{\mu\nu}$.

(c) [**6 points**] Explain how the result for $p(\boldsymbol{t})$ could have been obtained from the fact that $t_\mu = y_\mu + z_\mu$ with appropriate noise variables z_μ.

(d) [**14 points**] Consider the special case where $C_{\mu\nu} = \delta_{\mu\nu}$. Simplify $p(\boldsymbol{t})$ to an expression depending on $\boldsymbol{t}$ only via $\boldsymbol{t}^2$, using the fact that the matrix $\boldsymbol{K}$ is now a multiple of the identity matrix. Show that the noise level σ^2 that maximizes $\ln p(\boldsymbol{t})$ obeys

$$1 + \sigma^2 = \frac{1}{p} \sum_{\mu=1}^{p} t_\mu^2$$

Give an interpretation of this result.