King's College London

University Of London

ATTACH this paper to your script **USING THE STRING PROVIDED**

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

MSc Examination

7CCMCS06 Elements of Statistical Learning

Summer 2011

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. Consider the problem of the maximizing with respect to the parameters $\boldsymbol{\theta}$ the marginal likelihood $p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$ for a vector $\mathbf{x}$ of observed random variables. Here the variables $\mathbf{z}$ are hidden, i.e. unobserved, and the sum runs over all possible values of $\mathbf{z}$.

(a) [**10 points**] Prove that, for any probability distribution $q(\mathbf{z})$,

$$\ln p(\mathbf{x}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q||p)$$

where

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{z}} q(\mathbf{z}) \ln\left(\frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})}\right) \qquad \mathrm{KL}(q||p) = \sum_{\mathbf{z}} q(\mathbf{z}) \ln\left(\frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})}\right)$$

(b1) [**5 points**] The E-step of the EM algorithm consists of maximizing $\mathcal{L}(q, \boldsymbol{\theta}_{\mathrm{old}})$ over q at fixed $\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathrm{old}}$. Show that this gives $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$.

(b2) [**10 points**] The M-step consists of maximizing $\mathcal{L}(q, \boldsymbol{\theta})$ over $\boldsymbol{\theta}$ at fixed q. Show that for the q obtained from the E-step, this is equivalent to maximizing the function

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\mathrm{old}}) = \int d\mathbf{z}\, p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}_{\mathrm{old}}) \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$$

If the maximum occurs at $\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathrm{new}}$, show that $\ln p(\mathbf{x}|\boldsymbol{\theta}_{\mathrm{new}}) \geq \ln p(\mathbf{x}|\boldsymbol{\theta}_{\mathrm{old}})$.

(c) Consider now the problem of maximizing the likelihood of N observed datapoints $\mathbf{x} = (x_1, \ldots, x_N)$ under a Gaussian mixture model with K components, where

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{n=1}^{N}\left(\sum_{k=1}^{K} \pi_k (2\pi\sigma_k^2)^{-1/2} e^{-(x_n-\mu_k)^2/(2\sigma_k^2)}\right)$$

and $\boldsymbol{\theta} = (\pi_1, \ldots \pi_K, \mu_1, \ldots, \mu_K, \sigma_1, \ldots, \sigma_K)$ collects all parameters.

(c1) [**10 points**] Let $z_{nk} \in \{0, 1\}$ with $\sum_k z_{nk} = 1\ \forall n \in \{1, \ldots, N\}$ be 1-of-K variables indicating which mixture component k data point x_n is being generated from. Write down the appropriate $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$ and $p(\mathbf{z}|\boldsymbol{\theta})$, where $\mathbf{z} = (z_{11}, \ldots, z_{NK})$, and show that $p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$.

(c2) [**15 points**] Find the function $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\mathrm{old}})$. Add to this a Lagrange multiplier term $-\lambda \sum_{k=1}^{K} \pi_k$ to enforce the constraint $\sum_k \pi_k = 1$. Find the conditions for Q to have a maximum with respect to $\boldsymbol{\theta}$, and hence derive the EM update equations for $\boldsymbol{\theta}$.

See Next Page

2. Consider Bayesian linear regression. The output distribution given an input vector $\mathbf{x}$ and weights $\mathbf{w} \in \mathbb{R}^M$ is

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = [\beta/(2\pi)]^{1/2} e^{-\beta[t-\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})]^2/2}$$

Here $\boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_M(\mathbf{x}))$ is a vector of fixed basis functions, and t can be viewed as the clean output $\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$ corrupted by Gaussian noise of variance β^{-1}. The prior over the weights is $p(\mathbf{w}|\alpha) = [\alpha/(2\pi)]^{M/2} \exp(-\alpha\mathbf{w}^{\mathrm{T}}\mathbf{w}/2)$.

Assume you are given a data set of N training inputs $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and associated outputs $t_1, \ldots, t_N$, corrupted by i.i.d. noise as specified above. Abbreviate $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$. All probabilities below are conditional on the training inputs.

(a) **[15 points]** Write down the posterior distribution $p(\mathbf{w}|\mathbf{t}, \alpha, \beta)$. By completing the square, show that it is a Gaussian distribution $\mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$ with mean and covariance matrix

$$\mathbf{m}_N = \beta\mathbf{S}_N\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}, \qquad \mathbf{S}_N = (\alpha\mathbf{I} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}$$

Here $\mathbf{I}$ is the $M \times M$ identity matrix, and the matrix $\mathbf{\Phi}$ has entries $\Phi_{nj} = \phi_j(\mathbf{x}_n)$.

(b) **[15 points]** Show that the predictive distribution $p(\hat{t}|\hat{\mathbf{x}}, \mathbf{t}, \alpha, \beta)$ for the output $\hat{t}$ at test input $\hat{\mathbf{x}}$ is a Gaussian distribution $\mathcal{N}(\hat{t}|m(\hat{\mathbf{x}}), v(\hat{\mathbf{x}}))$ with mean and variance

$$m(\hat{\mathbf{x}}) = \mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\hat{\mathbf{x}}), \qquad v(\hat{\mathbf{x}}) = \beta^{-1} + \boldsymbol{\phi}(\hat{\mathbf{x}})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\hat{\mathbf{x}})$$

Discuss how this result differs from what would be obtained by predicting with the maximum likelihood weight estimate $\mathbf{w} = \mathbf{m}_N$.

(c) **[20 points]** Show that the marginal likelihood of the observed training data is

$$\ln p(\mathbf{t}|\alpha, \beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}\mathbf{t}^{\mathrm{T}}\mathbf{t} + \frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N + \frac{1}{2}\ln|\mathbf{S}_N|$$

Explain why maximizing this quantity with respect to α and β is a reasonable method for setting these hyperparameters.

You may, if you wish, use without proof the following property of the linear Gaussian model: if $\mathbf{x}$ is a vector of Gaussian random variables and $\mathbf{y}$ is Gaussian conditionally on $\mathbf{x}$, so that $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma})$ and $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{V})$, then $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{V} + \mathbf{A}\mathbf{\Sigma}\mathbf{A}^{\mathrm{T}})$.

See Next Page

3. Consider Bayesian discriminative binary classification. We encode class $\mathcal{C}_1$ as output $t = 1$, and class $\mathcal{C}_2$ as $t = 0$. The output distribution given an input vector $\mathbf{x}$ and weights $\mathbf{w} \in \mathbb{R}^M$ is

$$p(t = 1|\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})), \qquad p(t = 0|\mathbf{x}, \mathbf{w}) = 1 - p(t = 1|\mathbf{x}, \mathbf{w})$$

Here $\boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_M(\mathbf{x}))$ is a vector of fixed basis functions. The squashing function σ obeys $0 < \sigma(a) < 1$, is monotonically increasing, and has the symmetry $\sigma(-a) = 1 - \sigma(a)$. The prior over the weights is $p(\mathbf{w}) = [\alpha/(2\pi)]^{M/2} \exp(-\alpha \mathbf{w}^{\mathrm{T}}\mathbf{w}/2)$.

Assume you are given a data set of N training inputs $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and associated outputs $t_1, \ldots, t_N$. Abbreviate $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$. All probabilities below are conditional on the training inputs.

(a) **[20 points]** Show that the posterior distribution $p(\mathbf{w}|\mathbf{t})$ has the form $p(\mathbf{w}|\mathbf{t}) = \exp[-E(\mathbf{w})]/Z$ with

$$E(\mathbf{w}) = \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_{n=1}^{N} \left[t_n \ln \sigma(a_n) + (1 - t_n) \ln \sigma(-a_n)\right], \qquad a_n = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)$$

Find the gradient of $E(\mathbf{w})$. Show that the Hessian of $E(\mathbf{w})$ is

$$\nabla\nabla E(\mathbf{w}) = \alpha \mathbf{I} + \sum_{n=1}^{N} \boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}[t_n g(a_n) + (1 - t_n) g(-a_n)]$$

with $g(a) = [\sigma'(a)/\sigma(a)]^2 - \sigma''(a)/\sigma(a)$ and $\mathbf{I}$ the $M \times M$ identity matrix. Hence show that, if the function $-\ln(\sigma(a))$ is convex, the Hessian of E is positive definite. What does this imply about uniqueness of the maximum a posteriori (MAP) weights $\mathbf{w}_{\mathrm{MAP}}$?

(b) **[15 points]** Assuming that $\mathbf{w}_{\mathrm{MAP}}$ has been determined numerically, state the Laplace approximation $q(\mathbf{w})$ to the posterior $p(\mathbf{w}|\mathbf{t})$. Show that the resulting approximate predictive distribution for the output $\hat{t}$ at test input $\hat{\mathbf{x}}$ is

$$q(\hat{t} = 1|\hat{\mathbf{x}}, \mathbf{t}) = \int \sigma(a)\mathcal{N}(a|\mathbf{w}_{\mathrm{MAP}}^{\mathrm{T}}\boldsymbol{\phi}(\hat{\mathbf{x}}), \boldsymbol{\phi}(\hat{\mathbf{x}})^{\mathrm{T}}\mathbf{A}^{-1}\boldsymbol{\phi}(\hat{\mathbf{x}}))\, da$$

for an appropriate matrix $\mathbf{A}$.

(c) **[15 points]** For the case of the inverse probit squashing function, $\sigma(a) = \int_{-\infty}^{a} \mathcal{N}(x|0, 1)\, dx$, find the integral

$$\int \sigma(a)\mathcal{N}(a|\mu, \sigma^2)\, da$$

explicitly. Hence give an explicit expression for $q(\hat{t} = 1|\hat{\mathbf{x}}, \mathbf{t})$ for this case.

Hint: You may want to change variables to $z = (a - \mu)/\sigma$, differentiate the integral with respect to μ, and then integrate over μ again at the end.