# UNIVERSITY COLLEGE LONDON

# EXAMINATION FOR INTERNAL STUDENTS

| | | |
|---|---|---|
| MODULE CODE | : | **COMPGI13** |
| ASSESSMENT PATTERN | : | **COMPGI13B** |
| MODULE NAME | : | **Advanced Topics in Machine Learning** |
| DATE | : | **16-March-09** |
| TIME | : | **14:00** |
| TIME ALLOWED | : | **2 Hours 30 Minutes** |

Advanced Topics in Machine Learning, GI13, 2009

Answer TWO of FOUR questions.

Marks for each part of each question are indicated in square brackets

Calculators are NOT permitted

1. This question concerns the regression model

$$y = w^{\mathsf{T}}\phi(x) + \varepsilon$$

where $\phi(x) = (\phi_1(x), \phi_2(x), \ldots, \phi_D(x))$ is a vector function of the input $x$. The term $\varepsilon$ denotes additive zero mean Gaussian noise with variance $s^2$ so that

$$p(y|x,w) = \mathcal{N}(y; w^{\mathsf{T}}\phi(x), s^2)$$

A multi-variate Gaussian distribution with mean $\mu$ and covariance $\Sigma$ is defined as

$$\mathcal{N}(x;\mu,\Sigma) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} e^{-\frac{1}{2}(x-\mu)^{\mathsf{T}}\Sigma^{-1}(x-\mu)}$$

a. For a zero mean unit covariance Gaussian prior distribution

$$p(w) = \mathcal{N}(w; 0, I)$$

where 0 is a $D$-dimensional zero vector, and $I$ is the $D \times D$ identity matrix, show that

$$p(y|x) = \int p(y|x,w)p(w)dw = \mathcal{N}(y; 0, \sigma^2)$$

where

$$\sigma^2 = \phi(x)^{\mathsf{T}}\phi(x) + s^2$$

(You may make use of the fact that the distribution of a linearly transformed Gaussian random variable is a Gaussian distribution).

[10 marks]

b. Given a set of training data $\mathcal{D} = \{(x^n, y^n), n = 1, \ldots, N\}$, and a zero mean unit covariance Gaussian weight prior $p(w) = \mathcal{N}(w; 0, I)$, show that the posterior weight distribution is given by

$$p(w|\mathcal{D}) \propto \exp\left(-\frac{1}{2}w^{\mathsf{T}}w - \frac{1}{2s^2}\sum_{n=1}^{N}(y^n - w^{\mathsf{T}}\phi(x^n))^2\right)$$

and hence show that this is a Gaussian distribution with mean

$$\mu = \left(I + \frac{1}{s^2}\sum_{n=1}^{N}\phi(x^n)\phi(x^n)^{\mathsf{T}}\right)^{-1}\frac{1}{s^2}\sum_{n=1}^{N}y^n\phi(x^n)$$

and covariance

$$\Sigma = \left(I + \frac{1}{s^2}\sum_{n=1}^{N}\phi(x^n)\phi(x^n)^{\mathsf{T}}\right)^{-1}$$

[20 marks]

c. Show that for

$$p(w|\mathcal{D}) = \mathcal{N}(w; \mu, \Sigma)$$

then for a novel input $x^*$, the distribution of the corresponding output $y*$ is

$$p(y^*|x^*, \mathcal{D}) \equiv \int p(y^*|x^*, w)p(w|\mathcal{D})dw = \mathcal{N}(y^*|\mu^*, \sigma_*^2)$$

where

$$\mu^* \equiv \mu^{\mathsf{T}}\phi(x^*)$$

and

$$\sigma_*^2 \equiv \phi(x^*)^{\mathsf{T}}\Sigma\phi(x^*) + s^2$$

[20 marks]

[Total 50 marks]

2. In conditional PLSA the aim is to find a decomposition of a matrix with elements $p(i|j)$. Each element of the matrix $p$ is positive and

$$\sum_i p(i|j) = 1$$

so that each column of $p$ sums to 1. This question derives an approximate decomposition of the matrix $p$ in the form

$$p(i|j) \approx \sum_k \tilde{p}(i|k)\tilde{p}(k|j)$$

a. The Kullback-Leibler divergence is defined as

$$KL(q,p) = \sum_x (q(x)\log q(x) - q(x)\log p(x))$$

for distributions $q(x)$, $p(x)$. Consider the bound

$$\log x \leq x - 1$$

By replacing $x$ in the above bound with $q(x)/p(x)$, show that

$$KL(q(x), p(x)) \geq 0$$

[10 marks]

b. By taking the $KL$ divergence

$$KL(p(\cdot|j), \tilde{p}(\cdot|j))$$

show that minimising this $KL$ divergence with respect to $\tilde{p}(\cdot|j)$ is equivalent to maximising the likelihood

$$\sum_i p(i|j) \log \tilde{p}(i|j)$$

[5 marks]

Explain why a valid measure of the accuracy of the approximation $\tilde{p}$ can be taken to be

$$\sum_{i,j} p(i|j) \log \tilde{p}(i|j)$$

[5 marks]

c. By considering

$$KL(q(\cdot|i,j), \tilde{p}(\cdot|i,j))$$

show that

$$\log \tilde{p}(i|j) \geq \sum_k q(k|i,j) \left(-\log q(k|i,j) + \log \tilde{p}(i,k|j)\right)$$

[5 marks]

and hence

$$\sum_{i,j} p(i|j) \log \tilde{p}(i|j) \geq \sum_{k,i,j} q(k|i,j) p(i|j) \left(-\log q(k|i,j) + \log \tilde{p}(i,k|j)\right) \tag{1}$$

[5 marks]

d. Using $\tilde{p}(i,k|j) = \tilde{p}(i|k)\tilde{p}(k|j)$ in the above bound, equation (1), first assume that $\tilde{p}(i|k)$ is fixed and isolate the contribution from $\tilde{p}(k|j)$. Show that, for fixed $\tilde{p}(i|k)$, and $q(k|i,j)$, the optimal setting for $\tilde{p}(k|j)$ to maximise the bound is

$$\tilde{p}(k|j) \propto \sum_i p(i|j)q(k|i,j)$$

[5 marks]

Similarly, for fixed $q(k|i,j)$ and $\tilde{p}(k|j)$, show that optimally

$$\tilde{p}(i|k) \propto \sum_j p(i|j)q(k|i,j)$$

[5 marks]

Finally show that for fixed $\tilde{p}(i|k)$, $\tilde{p}(k|j)$, show that optimally

$$q(k|i,j) = \tilde{p}(k|i,j)$$

where

$$\tilde{p}(k|i,j) \propto \tilde{p}(i|k)\tilde{p}(k|j)$$

[5 marks]

Using the above results, suggest an EM style iterative algorithm for training conditional PLSA.

[5 marks]

[Total 50 marks]

3. This question pertains to regularisation methods with kernels. Consider the following optimisation problem

$$\min_{w \in \mathbb{R}^d} \left\{ \sum_{i=1}^{m} (w^\top x_i - y_i)^2 + \gamma \|w\|^2 \right\}, \tag{2}$$

where $\|\cdot\|$ denotes the $L_2$ norm, $\gamma > 0$ is a fixed real number, $x_1, \ldots, x_m \in \mathbb{R}^d$ are given inputs and $y_1, \ldots, y_m \in \mathbb{R}$ given outputs.

a. State the *Representer Theorem* for problem (2). Prove the Representer Theorem.

[14 marks]

b. Let $G$ be the $m \times m$ matrix with entries $G_{ij} = x_i^\top x_j$, for $i, j = 1, \ldots, m$. Derive a problem equivalent (dual) to (2), which involves only matrix $G$ (and does not involve the inputs $x_i$).

[12 marks]

c. Let $W \in \mathbb{R}^{d \times n}$ be a $d \times n$ matrix and assume that $d > n$. Define the *trace norm* $\|W\|_{tr}$ of $W$.

[12 marks]

d. Show that the function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, defined as

$$K(a,b) := a^\top b \qquad \text{for all } a, b \in \mathbb{R}^d$$

is a symmetric positive semidefinite kernel.

[12 marks]

[Total 50 marks]

4. This question pertains to convex functions and convex optimisation.

a.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function. When do we say that $f$ is a *convex* function? When do we say that $f$ is *strictly convex*?

[12 marks]

b. Consider the function $f : \mathbb{R}^d \to \mathbb{R}$ defined as

$$f(w) = \sum_{i=1}^{m} (w_i - 1)^2 + \|w\|^2 \qquad \text{for every } w \in \mathbb{R}^d.$$

Show that $f$ is convex, using properties of convex functions or in another way (Note: $w_i$ denotes the $i$-th component of vector $w$).

[14 marks]

c. Consider the function $f : \mathbb{R}^d \to \mathbb{R}$ defined as

$$f(w) = \sum_{i=1}^{m} \max\{w_i, 0\} + \|w\|^2 \qquad \text{for every } w \in \mathbb{R}^d.$$

Show that $f$ is convex, using properties of convex functions or in another way. Is $f$ strictly convex? Explain your answer.

[12 marks]

d. Give an example of a *quadratic program*. Give an example of a *linear program*.

[12 marks]

[Total 50 marks]

END OF PAPER