Name: ______________________________ Student No.: ______________

Do ALL questions.

Final exam
ECE1508F, Probability Models and Algorithms
November 30, 2004
Closed-book, calculators allowed, 2.5 hours, total of 60 marks

1. (20 marks) Free energy and the EM algorithm.

1a) In a model $P(h, v)$ with visible variable v and hidden variable h we approximate $P(h|v)$ with a distribution $Q(h)$. Give an expression for the free energy F. (2 marks)

1b) Explain why it is that minimizing F w.r.t. Q is the same as minimizing the relative entropy (a.k.a. KL-divergence), $D = \int_h Q(h) \log(Q(h)/P(h|v))$ w.r.t. Q. (2 marks)

1c) In the EM algorithm, there are 2 hidden variables, h and θ, where θ is the model parameter, and the joint distribution is $P(x, h, \theta)$. (For notational simplicity, assume there is only one training case.) Using an approximation $Q(h, \theta) = \delta(\theta - \hat{\theta})Q(h)$, the EM algorithm iteratively minimizes F by alternating between minimizing F w.r.t. $Q(h)$ while keeping $\hat{\theta}$ fixed (the E step), and minimizing F w.r.t. $\hat{\theta}$ while keeping $Q(h)$ fixed (the M step). Show that setting $Q(h) = P(h|v, \hat{\theta})$ minimizes F in the E step. (2 marks)

1d) Show that each iteration of EM either increases $P(x, \theta)$ or leaves it unchanged. (6 marks)

1e) Assuming the statement in (1d) is true and that $P(\theta) = const$, explain why EM is referred to as a technique for maximum likelihood estimation. (2 mark)

1f) Suppose in the E step, instead of setting $Q(h)$ to its optimal value $Q^*(h)$, you set it to an arithmetic average between this value and the previous value $Q^{\text{old}}(h)$: $Q(h) = (Q^*(h) + Q^{\text{old}}(h))/2$. Show that F based on $Q(h)$ is lower than F based on $Q^{\text{old}}(h)$, *i.e.*, that this new EM algorithm will also decrease F. (6 marks)

2. (10 marks) The sum-product algorithm in factor graphs.

2a) Suppose a function $\phi(x_1, x_2, x_3)$ receives messages $\mu_1(x_1)$ and $\mu_2(x_2)$ from x_1 and x_2. Give an expression for the message $\mu_3(x_3)$ sent from ϕ to x_3. (2 marks)

2b) Suppose an unobserved variable x appears in the arguments of functions $\phi_1(x, u)$, $\phi_2(x, v)$ and $\phi_3(x, w)$. x receives messages $\mu_1(x)$ from ϕ_1 and $\mu_2(x)$ from ϕ_2. Give an expression for the message $\mu_3(x)$ sent from x to ϕ_3. (2 marks)

.

2c) Repeat (2b), assuming that x is observed and has the value x^*. (1 marks)

2d) Suppose an unobserved variable x appears in the arguments of functions $\phi_1(x, u)$, $\phi_2(x, v)$ and $\phi_3(x, w)$. Give an expression for how the incoming messages $\mu_1(x)$, $\mu_2(x)$ and $\mu_3(x)$ arriving at x from these functions should be combined and normalized so as to estimate the distribution over x. (1 marks)

2e) For what type of factor graph and under what conditions regarding propagation of messages will (2d) give exact inference? (2 marks)

2f) Suppose we wish to compute $\max_{x_1} \max_{x_2} \max x_3 \max x_4 \phi_1(x_1, x_2)\phi_2(x_2, x_3)\phi_3(x_3, x_4)\phi_4(x_4, x_5)$. Describe a general method for how this computation can be carried out most efficiently for arbitrary ϕ's. (2 marks)

3. Approximate inference in the model from the midterm exam. (30 marks)

You have a training set $x^{(1)}, \ldots, x^{(T)}$ and you have knowledge that each training case is generated by selecting randomly one of J means and then adding Gaussian noise with variance selected from one of K variances. (Note that if $J = K$ and the selection of the variance is tied to the selection of the mean, this is a standard mixture of Gaussians.) So, in the model for a training case x, there are two hidden variables, the class of the mean $c \in \{1, \ldots, J\}$ and the class of the noise $s \in \{1, \ldots, K\}$. In this question, you will explore exact inference, mean field variational inference, Gibbs sampling inference, and learning.

3a) (Repeat of part of midterm question.) Write down $P(c, s, x)$ as a product of conditional distributions specified in terms of model parameters and terms like exp(). Use π_j as the probability of picking mean j, $\mu_1, \ldots, \mu_J$ as the J means, ρ_k as the probability of picking noise model k, and $\psi_1, \ldots, \psi_K$ as the K variances. Give an expression for the log-likelihood of the complete data, $\log P(c, s, x)$. (2 marks)

3b) Exact inference of c and s. Given x, describe how you would compute $P(c, s|x)$ using the value of x and the current values of the model parameters. Also, give an expression for the number of scalar binary operations needed in this computation, in terms of J and K. (4 marks)

3b) Using $Q^{(t)}(c,s)$ to denote the distribution $P(c,s|x^{(t)})$ found in (3a) for training case $x^{(t)}$, give an expression for the free energy for the entire training set, and derive an update equation for each model parameter, in terms of other model parameters, the x's and the Q-distributions. (4 marks)

3c) Mean field (variational) inference of c and s. Using the approximation $P(c, s|x) \approx Q(c)Q(s)$, give an expression for the free energy F of a single training case x, simplify F, explain how F can be efficiently computed, and give an expression for the number of scalar binary operations needed to compute F. (Hint: The number of computations should be significantly lower than in (3b).) (4 marks)

3d) By taking derivatives of F w.r.t. $Q(s)$ and $Q(c)$, find the mean field update for $Q(c)$ in terms of x, the model parameters and $Q(s)$; and find the update for $Q(s)$ in terms of x, the model parameters and $Q(c)$. (4 marks)

3e) Using $Q^{(t)}(c)$ and $Q^{(t)}(s)$ to denote the distributions found by iterating to convergence the updates in (3d) for training case $x^{(t)}$, give an expression for the free energy for the entire training set, and derive an update equation for each model parameter, in terms of other model parameters, the x's and the Q-distributions. (4 marks)

3f) Gibbs sampling inference of c and s. Given x, describe how Gibbs sampling would be used to obtain a sample of M pairs of the form $(c_1, s_1), \ldots, (c_M, s_M)$. Make sure to give the details of how each variable should be sampled, given the model parameters and the other variables, and justify your choice of points in the sampling procedure that are used to select the sample (4 marks).

3g) Denoting the sample of c-s pairs obtained for training case $x^{(t)}$ using the method in (3f) by $(c_1^{(t)}, s_1^{(t)}), \ldots, (c_M^{(t)}, s_M^{(t)})$, derive an update equation for each model parameter, in terms of other model parameters, the x's and the c-s pairs. (4 marks)