NAME: _____ STUDENT NO.: _____

/

UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE AND ENGINEERING
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Final Examination, April 19, 2012
Time allowed: 2.5 hours

ECE521H1S — Inference Algorithms

Exam Type A: No additional notes, books or data permitted.
Calculator Type 2: All non-programmable electronic calculators allowed.
Examiners: B. J. Frey, H. Xiong

Useful results

In the M step of EM, parameter θ in a BN with a term $P(x_j|X_{\pi(j)},\theta)$ is updated by solving

$$\sum_t \sum_{x_j^{(t)}} \sum_{X_{\pi(j)}^{(t)}} Q(x_j^{(t)}, X_{\pi(j)}^{(t)}) \frac{\partial}{\partial \theta} \log P(x_j^{(t)}|X_{\pi(j)}^{(t)},\theta) = 0.$$

If the child x_j or parents $X_{\pi(j)}$ are observed, then we only sum over the unobserved parents or child, and Q only depends on the unobserved variables.

Instructions

- Write your name and student number above.
- **Answer all questions.**
- Different questions have **very different value** and a total of **90 marks** is available.
- Make sure you have a complete exam paper, with 10 pages, including this one.
- Answer each question directly on the examination paper.
- Indicate clearly where your work can be found.
- Show your work! State assumptions, show all steps, and present all results clearly.

EXAMINER'S REPORT

1.		/12
2.		/14
3.		/14
4.		/30
5.		/20
Total:		/90

1. (12 marks) Consider a regression problem where $\mathbf{x}$ is the vector of features, y is the target, θ is the set of parameters, T is the number of training cases and N is the number of features. *Answer true "T" or false "F" for each question below.*

(a) Regular linear regression produce unique solutions only when $N \leq T$. ☐

(b) If the features are multiplied by a non-zero constant before regular linear regression is applied, the predictions for the training cases will depend on the value of the constant. ☐

(c) Ridge regression consists of combining regular linear regression with an L_2 regularization term. ☐

(d) Ridge regression produces unique solutions only when $N \leq T$. ☐

(e) If the features are multiplied by a non-zero constant before ridge regression is applied, the predictions for the training cases will depend on the value of the constant. ☐

(f) Decreasing the learning rate in gradient descent guarantees that the algorithm will converge to a better local minimum. ☐

(g) Generative models (such as mixtures of Gaussians) try to model $P(y|\mathbf{x})$. ☐

(h) Discriminative models (such as neural networks, logistic regression) try to model $P(y|\mathbf{x})$. ☐

(i) Bayesian inference models $P(\theta|y, \mathbf{x})$. ☐

(j) In general, inference in a tree-shaped Markov random field is NP-hard in the number of variables. ☐

(k) In general, it is easy to generate independent samples from a Bayesian network. ☐

(l) The set of conditional independencies that can be represented by Bayesian networks and factor graphs is identical. ☐

2. (14 marks) In Bayesian logistic regression, suppose we have a model $P(y|\theta, \mathbf{x})$, $y \in \{0, 1\}$ and a prior $P(\theta)$. Two data points are collected and they have feature vectors $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$ and measured targets $y^{(1)}$ and $y^{(2)}$.

(a) (2 marks) Give a formula for evaluating the likelihood $P(y|\theta, \mathbf{x})$ of a single training case.

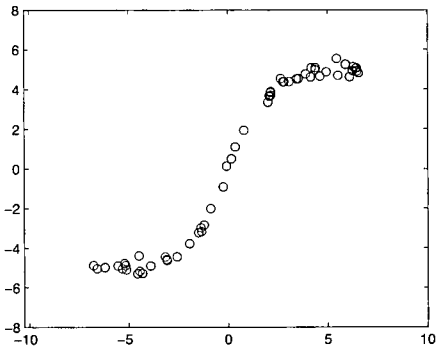
(b) (4 marks) Give a formula for the posterior distribution, $P(\theta|\mathbf{x}^{(1)}, y^{(1)}, \mathbf{x}^{(2)}, y^{(2)})$, in terms of the likelihood and prior. Make sure to include the normalizing constant.

(c) (4 marks) We want to classify a third data point with a feature vector $\mathbf{x}'$. According to the Bayesian method, what is the expression used to obtain the predicted distribution of y' given $\mathbf{x}'$ and the training data, $P(y'|\mathbf{x}', \mathbf{x}^{(1)}, y^{(1)}, \mathbf{x}^{(2)}, y^{(2)})$?

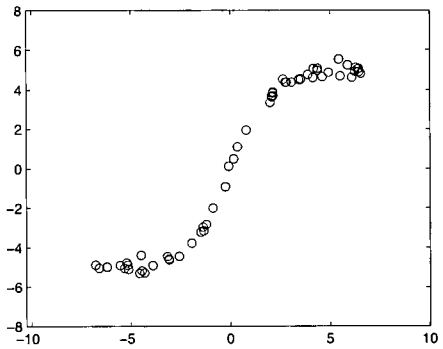
(d) (4 marks) Suppose we have M samples $\theta_1, \ldots, \theta_M$ from the posterior distribution. Give a formula for estimating the expected value of the prediction y'.

3. (14 marks) The plots below show a set of zero-mean 2D datapoints.

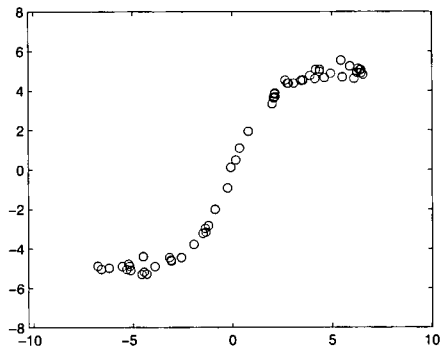
(a) (3 marks) Draw a line showing the principal subspace that would be obtained by 1D PCA.



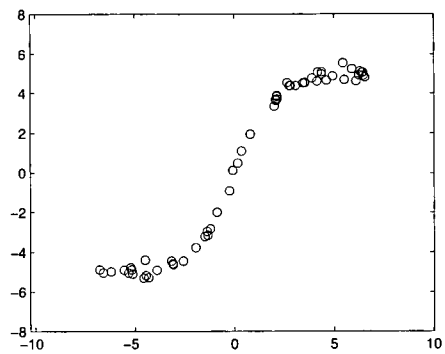
(b) (3 marks) Draw the subspace of the reconstructed data obtained after training a layered auto-encoder with 2 linear input units, 1 linear hidden unit and 2 linear output units.



(c) (4 marks) Draw the subspace for a layered auto-encoder with 2 linear input units, 1 linear unit in the first hidden layer, 2 sigmoidal units in the second hidden layer, and 2 linear output units.



(d) (4 marks) EM is used to train a mixture of *three* full-covariance Gaussians. Sketch the Gaussians, using ovals to indicate their locations and orientations.



4. (30 marks) Naive Bayes is a generative probability model used in classification tasks, where it is assumed that input features $x_1, x_2, \ldots x_N$ are independent given the class label y. Here, we consider a model with K classes, $y \in \{0, \ldots, K-1\}$ and binary features, $x_i \in \{0, 1\}$.

(a) (3 marks) Draw the Bayesian network for the Naive Bayes model $P(y, x_1, \ldots, x_N)$.

(b) (4 marks) Draw the Markov random field and the factor graph for the Naive Bayes model.

(c) (2 marks) How many free parameters does the model have? Justify your answer. (A parameter is *not* free if its value is determined by other parameters.)

(d) (3 marks) Suppose there are two classes ($K = 2$, $y \in \{0, 1\}$) and two features ($N = 2$) and the model is parameterized as follows: $P(y) = \pi^y (1-\pi)^{1-y}$, $P(x_i|y) = \lambda_{y,i}^{x_i}(1-\lambda_{y,i})^{(1-x_i)}$. Provide an expression for the likelihood function $P(y^{(1)}, \ldots, y^{(T)}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}|\pi, \lambda)$.

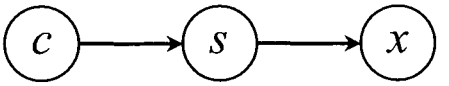
(e) (4 marks) Suppose we have $T = 6$ training cases (y, x_1, x_2): (011), (001), (000), (000), (111), (100). Write down the maximum likelihood estimates (numerical values) of the parameters π, $\lambda_{0,1}$, $\lambda_{0,2}$, $\lambda_{1,1}$, $\lambda_{1,2}$.

(f) (2 marks) You observe additional data, so that the entire dataset is (011), (001), (000), (000), (011), (011), (011), (000), (000), (011), (000), (000), (111), (100), (101), (110), (111), (101), (110), (100), (110), (101). Do you think the Naive Bayes assumption of independence is correct for this dataset? Justify your answer.

(g) (4 marks) Consider a model estimated from *different* data, that has parameters $\pi = 1/4$, $\lambda_{0,1} = 1/2$, $\lambda_{0,2} = 1/8$, $\lambda_{1,1} = 3/4$, $\lambda_{1,2} = 3/4$. To classify a test case with features $x_1 = 0, x_2 = 0$, we need to compute $P(y|x_1 = 0, x_2 = 0)$. Work out the numerical values for this distribution and state what the best guess for the class label, y, would be.

(h) (8 marks) Suppose we observe $x_1 = 0$, but we do not observe x_2. We would like to know what the value of x_2 is. *Using the model provided in part (g)*, apply the sum-product algorithm to find the distribution of x_2 given $x_1 = 0$. Draw the factor graph below – *fill the page!* – and write down the numerical values for all of the messages.

5. (18 marks) You would like to train the Bayesian network shown below using maximum likelihood estimation. For all training cases, variables c and x are observed, but the variable s is hidden, so you will need to use EM. c and s are discrete with $c \in \{1, \ldots, K\}$ and $s \in \{1, \ldots, J\}$. x is real-valued and it is assumed to be normally distributed with a mean and variance that depend on s. The dataset is $(c^{(1)}, x^{(1)}), \ldots, (c^{(T)}, x^{(T)})$.



The model is parameterized as follows:

$$P(c) = \alpha_c,$$
$$P(s|c) = \lambda_{c,s}$$
$$P(x|s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-(x-\mu_s)^2/2\sigma_s^2}$$

(a) (2 marks) Write down the normalization conditions for the parameters $\alpha_1, \ldots, \alpha_K$ and the parameters $\lambda_{c,1}, \ldots, \lambda_{c,J}$, for $c = 1, \ldots, K$.

(b) (4 marks) The variable c is observed, so estimating α is straightforward and does not require EM. Provide an expression for the maximum likelihood estimates of $\alpha_1, \ldots, \alpha_K$ in terms of the training data.

(c) (2 marks) The E step of the EM algorithm requires that you compute $Q(s) = P(s|c, x)$ for each training case, where here the superscript "(t)" has been omitted for clarity. Explain how $Q(s), s = 1, \ldots, J$ is computed using the model parameters and the observed variables c and x. (Make sure to show how the distribution is normalized.)

(d) (4 marks) For the M step, use the general form described in class for deriving EM updates for Bayesian networks to derive the update for the parameters $\lambda_{1,1}, \ldots, \lambda_{KJ}$ in terms of $Q(s^{(1)}) \ldots Q(s^{(T)})$ and $c^{(1)} \ldots c^{(T)}$. (Note that you may need to use a Lagrange multiplier to enforce normalization of λ.)

(e) (5 marks) Derive the M step updates for the parameters $\mu_1, \ldots, \mu_J$ in terms of $Q(s^{(1)}) \ldots Q(s^{(T)})$ and $x^{(1)} \ldots x^{(T)}$.

(f) (5 marks) Derive the M step updates for the parameter $\sigma_1^2, \ldots, \sigma_J^2$ in terms of $Q(s^{(1)}) \ldots Q(s^{(T)})$, $x^{(1)} \ldots x^{(T)}$ and the parameters μ. (Hint: Take the derivative w.r.t. $\log \sigma_j^2$, *not* σ_j^2.)