NAME: _____ STUDENT NO.: _____

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

Midterm Examination, February 28, 2012
Time allowed: 110 minutes

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 probability distributions

Gaussian pdf

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

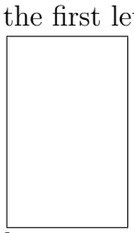
Binomial pmf

$$P(k) = \binom{N}{k} p^k (1-p)^{N-k}$$

Instructions

- Make sure you have a complete exam paper, with 7pages, including this one.
- Write your name and student number above, and enter the first letter of your last name in the box below.
- Answer **all** questions, and note the value of each question. A total of **60 marks** is available.
- Answer each question directly on the examination paper, using the back of each page if necessary. Indicate clearly where your work can be found.
- Show your work! State assumptions, show all steps, and present all results clearly.

Enter the first letter of



your last name here.

EXAMINER'S REPORT

1.		/8
2.		/6
3.		/10
4.		/12
5.		/8
6.		/16
7.		/10
Total:		/70

1. (8 marks) Each set below lists the observed values for x in a training set. For each, indicate whether it is ordinal, and the data type (binary, categorical, integer, real-valued, vector-valued). The first case is answered for you; the last two cases are tricky.

Set of observed values	Ordinal?	Data type
$x \in \{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}$	Yes	Integer
$x \in \{0.1, 0.12, 0.18, 0.21, 0.34, 0.36, 0.39, 0.44, 0.52, 0.55\}$		
$x \in \{\text{Cow}, \text{Motorcycle}, \text{Airplane}, \text{Dog}\}$		
$x \in \{0, 1, 2\}$, where 0=Horse, 1=Bicycle, 2=Barn		
$x \in \{\text{Heads}, \text{Tails}\}$		
$x \in \{\text{Black}, \text{Grey}, \text{White}\}$		
$x \in \mathcal{R}^n$		
$x \in \{3.14159, 2.7183, 1.4142\}$		
$x \in \left\{ \begin{pmatrix} \text{Penny} \\ \text{Heads} \end{pmatrix}, \begin{pmatrix} \text{Penny} \\ \text{Tails} \end{pmatrix}, \begin{pmatrix} \text{Nickel} \\ \text{Heads} \end{pmatrix}, \begin{pmatrix} \text{Nickel} \\ \text{Tails} \end{pmatrix} \right\}$		

2. (6 marks) Suppose you would like to train a neural network that uses a *step function* instead of the sigmoid activation function:

$$g(z) = \begin{cases} 1 & \text{if } z \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

(a) (2 marks) Can the back-propagation algorithm be used to infer the parameters? Justify your answer.

(b) (4 marks) Can the importance sampling method discussed in class be used to infer the parameters? Justify your answer.

3. (10 marks) For a regression task with training cases $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(T)}, y^{(T)})$ and $x^{(t)} \in \mathcal{R}^n$, $y^{(t)} \in \mathcal{R}$, suppose you would like to evaluate a predictor for y given by $f(x)$.

(a) (2 marks) Write an expression for the sum of squared prediction error in terms of the x's, the y's and the function $f(x)$.

(b) (2 marks) Write an expression for the sum of absolute loss.

(c) (2 marks) Interpreting the predictor $f(x)$ as the mean of a Gaussian distribution over y with variance σ^2, give an expression for the data likelihood in terms of the x's, the y's, the function $f()$ and basic operators such as exp() and $\sqrt{\ }$. You should assume that the training cases are IID.

(d) (4 marks) Show that the logarithm of the likelihood from (c) equals the negative sum of squared error from (a), up to additive and multiplicative constants that don't depend on the predictor, $f()$.

4. (12 marks) The parameter derivative is shown below for four different learning scenarios. Note that $g(z) = 1/(1 + \exp(-z))$ and $g'()$ is the derivative of $g()$.

In each case, specify the type of *model*, the type of *loss function*, and the type of *parameter penalty*. The first case is answered for you.

Parameter derivative	Model	Loss function	Param penalty
$\frac{\partial E}{\partial \theta_i} = -\sum_t (y^{(t)} - \theta \cdot x^{(t)}) x_i^{(t)}$	Linear regression	Squared error	None
$\frac{\partial E}{\partial \theta_i} = -\sum_t (y^{(t)} - \theta \cdot x^{(t)}) x_i^{(t)} + \lambda \theta_i$			
$\frac{\partial L}{\partial \theta_i} = \sum_t (y^{(t)} - g(\theta \cdot x^{(t)})) x_i^{(t)}$			
$\frac{\partial L}{\partial \theta_i} = \sum_t (y^{(t)} - g(\theta \cdot x^{(t)})) x_i^{(t)} + \lambda \text{sgn}(\theta_i)$			
$\frac{\partial E}{\partial w_{\ell m}} = \sum_t \frac{\partial E}{\partial x_m^{(t)}} g(x_\ell^{(t)})$, where $\frac{\partial E}{\partial x_m^{(t)}} = \sum_{k=m+1}^{n} \frac{\partial E}{\partial x_k^{(t)}} w_{km} g'(x_m^{(t)})$			

5. (8 marks) You are going to describe different neural network learning algorithms using the following types of computer operations:

(1) Update parameters.
(2) Repeat following commands until training error converges.
(3) Repeat following commands until validation error increases.
(4) Apply back-propagation to obtain parameter derivatives for one case.
(5) Loop over training cases and execute following commands.
(6) Randomly initialize parameters.
(7) Apply forward propagation to obtain prediction for one case.
(8) End repeat.
(9) End loop over training cases.
(10) Add current parameter derivatives to accumulated derivatives.

For each of the following methods, list the required computer operations in proper order, *eg*, "8, 6, 5, 3, . . . ".

(a) Batch learning without early stopping:

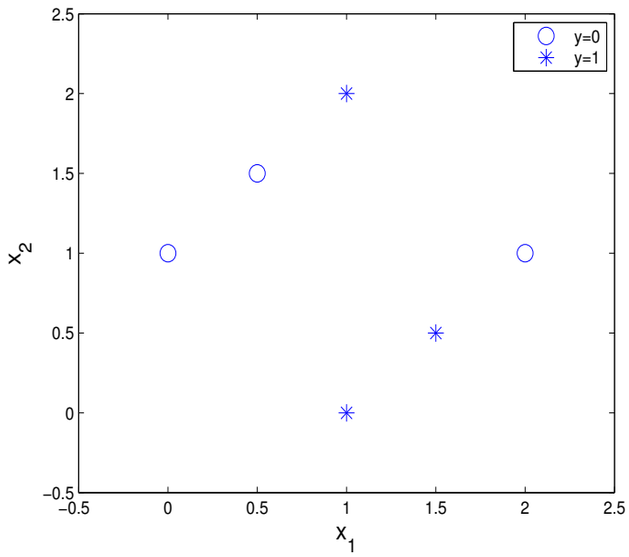
(b) Online learning without early stopping:

(c) Batch learning *with* early stopping:

(d) Online learning *with* early stopping:

6. (16 marks) You are using logistic regression to classify the data shown below. You use a gradient-based maximum likelihood (ML) algorithm and obtain the following model:

$$P(y = 1|x_1, x_2) = \frac{1}{1 + e^{-(.69x_1 - .69x_2)}}.$$



(a) (2 marks) Derive an equation that describes the decision boundary *and* draw the decision boundary along with the predicted class (0 or 1) on the plot shown above using a solid line labelled "a".

(b) (1 mark) What is the error rate of this model on the training data?

(c) (3 marks) Is this the best possible error rate that a logistic regression model can achieve? Justify your answer. (If you want to draw the decision boundary for another model to justify your answer, use a solid line labelled "c".)

(d) (4 marks) Prove that the model shown at the top of the page is the maximum likelihood model. (Make sure to take into account all three parameters that a general two-input logistic regression model has.)

(e) (4 marks) Instead of using simple logistic regression, you decide to use a neural network with two sigmoidal hidden variables and a sigmoidal output. The two hidden variables are connected to x_1, x_2 and a bias (an additional input fixed to 1), whereas the output variable is connected to the two hidden variables and a bias. You find that when you train this model to maximize the Bernoulli likelihood, you fit the training data perfectly and achieve a training classification rate of 100%. Give the set of weights for a neural network that achieves this and briefly explain how it works. (Hint: This can be done by setting the weight on either x_1 or x_2 to zero.)

(f) (2 marks) Draw the decision boundary of this neural network on the above plot, using a dashed line labelled "f".

7. (10 marks) Suppose we have k models $h_1, h_2, ...h_k$ for a binary prediction problem. Each of these models makes a binary prediction, $\hat{y} = h_i(x)$, based on the input feature vector x. Let the error probability $\epsilon(h_i) = \mathbb{E}_x\big(\sum_{y=0}^{1} I(h_i(x) \neq y)p(y|x)\big)$, where I is the indicator function and the expectation is taken over the distribution of x, $p(x)$. We would like to study estimates of the error probability using a dataset of 10 i.i.d. training cases $\big(x^{(1)}, y^{(1)}\big), \big(x^{(2)}, y^{(2)}\big), ...\big(x^{(10)}, y^{(10)}\big)$ drawn from the joint distribution of x and y. That is, $x^{(t)}$ is drawn from $p(x)$ and then $y^{(t)}$ is drawn from from $p(y|x^{(t)})$.

(a) (1 mark) Write down the equation for the estimated error of model i, $\hat{\epsilon}(h_i)$, in terms of the training cases.

(b) (1 mark) Different randomly drawn datasets will give different values of $\hat{\epsilon}(h_i)$. How many possible values can $\hat{\epsilon}(h_i)$ take on?

(c) (2 marks) Give an expression for the distribution of $\hat{\epsilon}(h_i)$.

(d) (2 marks) Suppose $\epsilon(h_1) = 0.15$. Let A_i be the event that we underestimate the error of h_i by more than 10%. What is the probability $P(A_1)$ of this event?

(e) (2 marks) In general, are events A_i and A_j independent? Justify your answer.

(f) (2 marks) There are 3 models h_1, h_2, h_3, and we pick one of the models $\hat{h}$ based on the training error. Suppose $\epsilon(h_1) = 0.15$, $\epsilon(h_2) = 0.14$ and $\epsilon(h_3) = 0.16$. Compute an upper bound on the probability that we underestimate the error of $\hat{h}$ by more than 10%. You should use the union bound: $P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$.