Name: ______________________________ Student No.: __________

University of Toronto
Faculty of Applied Science and Engineering
Department of Electrical and Computer Engineering

Midterm Examination, February 12, 2013
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: Brendan J. Frey, Hui Y. Xiong, Jimmy Ba.

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 9pages, 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 **x 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.

Examiner's Report

1.		/15
2.		/15
3.		/15
4.		/15
Total:		/60

1. (15 marks) True and false questions: 0 marks for the wrong answer, 0.5 marks if you leave the box blank, 1 mark for the correct answer.

Notation: X are the feature vectors, $x_j^{(i)}$ is the jth feature for the ith training case, $y^{(i)}$ is the target for the ith training case, $\hat{\theta}$ is the set of learned parameters, $\hat{y}^{(i)}$ is the predicted response for the ith training case, N is the number of training cases and K is the number of features.

(a) Regular linear regression doesn't produce unique solutions when $K > N$. □

(b) For regular linear regression with $K < N$, if the training targets are scaled by a factor of 10, the new $\hat{\theta}$ will also be scaled by a factor of 10. □

(c) Suppose we perform linear regression, but with an L_2 penalty on θ. If the training targets are scaled by a factor of 10, the new $\hat{\theta}$ will also be scaled by a factor of 10. □

(d) Suppose we perform linear regression, but with an L_2 penalty on θ. If we flip the sign of the training targets, the new $\hat{\theta}$ will be equal to the negative of the original $\hat{\theta}$. □

(e) In logistic regression, if we multiply θ by a factor of 2, the decision boundary will move by e^2. □

(f) In a neural network with a single hidden layer and binary targets, if we multiply θ by a factor of 2, the decision boundary does not change. □

(g) In the squared error cost function of a neural network with a single hidden layer and binary targets, there can be many local minima. □

(h) Early stopping usually helps gradient descent to converge to a better local minimum. □

(i) Suppose we define feature sensitivity to be the change in the prediction for a small change in the feature, for training case i, $d\hat{y}^{(i)}/dx_j^{(i)}$. For logistic regression, the feature sensitivity of a particular feature will have the same sign for all training cases. □

(j) For neural networks with one hidden layer, the feature sensitivity of a particular feature will have the same sign for all training cases. □

(k) The following is related to overfitting: A large number of features and a small number of training cases. □

(l) The following is related to overfitting: The training error (or likelihood) oscillates when performing gradient descent. □

(m) The following is related to overfitting: The training error is very low and the test error is very high. □

(n) The following is related to overfitting: Gradient descent converges to different solutions when initialized differently. □

(o) For neural networks, as we increase the number of hidden units, the difference between the training error and the test error becomes smaller. □

2. (15 marks) Multiple choice questions; circle *one* choice for each question (3 marks each):

(1) Which of the following is *not* a reason why minimizing the squared error in linear regression is equivalent to maximum likelihood learning under Gaussian noise?

a. The log of the Gaussian pdf is a parabola.
b. The Gaussian noise is assumed to be independent between training cases.
c. After taking the logarithm, products become sums.
d. The sum of parabolas is also a parabola.

(2) When performing gradient descent in a neural network, the training error oscillates. What should you do?

a. Consider adding a L_2 regularization term.
b. Increase the number of hidden units.
c. Reduce the learning rate.
d. Start using early stopping.

(3) Which of the following models is relatively unlikely to severely overfit a training set of 100 examples?

a. 3-nearest neighbours, 100 features.
b. Regular linear regression, 100 features.
c. Regular linear regression, 50 highly correlated features.
d. A single hidden layer neural network with 10 hidden units and 20 features.

(4) After performing gradient descent to convergence on a training set of 50 examples, our neural network gets a classification rate of 100% on the training data. However, the model gets only 50% on the test data. The targets are binary and the two classes are balanced (same number of 0s and 1s). What should we do?

a. The model is fine, because it gets 100% on the training data.
b. We should consider having more hidden units.
c. We should consider having more hidden layers.
d. We should consider including an L_2 penalty on the parameters.

(5) Which of the following is *not* what early stopping does?

a. Keeps the final parameters close to where they started.
b. Produces models with worse training error.
c. Improves the performance of the final model on training data.
d. Avoids overfitting.

3. (15 marks) Suppose you train a model with no real features but only a constant bias θ, so that predictions are the same for all training cases.

(a) (7 marks) Show that for linear regression, the value of θ that minimizes the squared error is the average of the training targets.

Important: In your proof, use N for the number of training cases, $y^{(i)}$ for the target value of training case i, E for the squared error, and $\bar{y}$ for the average of the training targets. Therefore you need to show that $\theta = \bar{y}$ is what minimizes E.

(b) (8 marks) Show that for logistic regression, the value of θ that maximizes the likelihood is given by the logarithm of the ratio of the number of targets that are 1 to the number of targets that are 0.

Important: In your proof, please use N for the number of training cases, m for the number of training cases that are labeled 1, and L for the log likelihood of the entire training dataset. Therefore you need to show that $\theta = \log \frac{m}{N-m}$.

3. (15 marks) We train a neural network with 2 features that are fully connected to 2 hidden units by a 2 by 2 weight matrix W, where w_{ij} is the connection from the ith feature to the jth hidden unit. Sigmoidal hidden units are used. The two hidden units are connected to a linear output unit through a 2 by 1 weight matrix V. There aren't any biases. We have only one training case whose feature vector is $x = [-1 \quad 2]$ and target is $y = 2$. We initialize $w_{ij} = 0, \forall i, j$ and $v_1 = 2, v_2 = 0$.

(a) (2 marks) What is the predicted value, $\hat{y}$?

(b) (2 marks) If we assume there is Gaussian noise with variance $\sigma^2 = 2$, what is the value of the training data log-likelihood?

(c) (4 marks) What is the gradient of the log-likelihood with respect to v_1? v_2?

(d) (4 marks) What is the gradient of the log likelihood with respect to w_{11}, w_{12}, w_{21} and w_{22}?

(e) (3 marks) If we perform gradient descent with a learning rate of 0.01, what will be the values of the parameters after the first step?

(f) (3 marks) If we keep doing gradient descent, do you think the log likelihood will approach a fixed (asymptotic) value? If so, what value is it?

4. A coin was given to us by a stranger and it has an unknown head probability θ. We toss the coin twice and see 2 heads. We assume that the tosses are independent.

(a) (4 marks) Our first hypothesis H_1 is that *a priori*, θ is uniformy distribution over $[0, 1]$. What is the probability that the next flip is a head?

(b) (4 marks) Our second hypothesis H_2 is that the stranger had 3 *fair* coins, but that the first coin has heads on both sides, the second coin has tails on both sides and the third coin has a head on one side and a tail on the other. In this scenario, we assume that the stranger picks one of the three coins with equal probability and then hands it to us. Under this hypothesis, what is the probability that the next flip is a head?

(c) (4 marks) Compute $P(D|H_1)$ and $P(D|H_2)$, where D is the data (2 heads). Note that computing $P(D|H_1)$ will involve integrating over θ.

(d) (2 marks) Which hypothesis H_1 or H_2 does our data favour?

(e) (4 marks) If our prior belief is that the two hypothesis are equally probable, what is the probability that the next flip is a head?