Department of Electrical and Computer Engineering

ELE 888 / EE 8209 - Intelligent Systems (Machine Learning)

Midterm Examination - Wednesday, February 24th, 2016

VERSION A

Name: | **ID**:

Duration of midterm examination

120 minutes (2 hours)

Total marks possible for this midterm

100 marks

Aids Permitted

1. Non-programmable, scientific calculator with no graphing or text-based capabilities.
2. 8.5" x 11" (letter-sized) **one side, hand-written** formula/cheat sheet. No photocopies or computer-typed versions allowed.

Instructions for this midterm examination

Please read before starting and good luck!

1. This is a **closed book** in-class examination, but a **one side, hand-written** formula/cheet sheet is allowed.
2. This examination consists of **ONE (1)** concept/theoretical based and **THREE (3)** full-answer type questions.
3. Please write your answers in the Ryerson TRS response booklets. However, should you be asked to place answers in this exam paper, please do so.
4. Once you finish your exam, please hand in this paper as well as all TRS response booklets that contain your answers.
5. Each question is **not equally weighted**. The weights of each question and sub-part are shown beside them.
6. *Any assumptions should clearly be stated for consideration of any credit*. Simply writing down the answer *without* any explanation will not be granted any marks.
7. **Show all calculations and steps taken to arrive at your answers to get full marks.**

Instructor use ONLY. Please do not write anything here.				
Q1 Concept / Theoretical Based	**Q2** Full-Answer #1	**Q3** Full-Answer #2	**Q4** Full-Answer #3	**TOTAL**

Question 1 - 15 marks

1. **(1 mark)** ***What version is your midterm***?

2. **(1 mark)** What is the main difference between regression and classification?

3. **(2 marks)** Explain the difference between stochastic gradient descent and batch gradient descent. Which one do you believe is faster for very large training sets and why?

4. **(2 marks)** What happens if we set the learning rate α to be too large in gradient descent?

5. **(6 marks)** Suppose you are using logistic regression to create a classification model. For each of the following statements below, state whether they are TRUE or FALSE and explain why.
    - (a) Adding a new feature to the classification model always results in equal or better performance compared to the original classification model on examples seen in the training set.
    - (b) Introducing regularization to the classification model always results in equal or better performance compared to the original classification model on examples seen in the training set.
    - (c) Adding new features to the model helps prevent overfitting of the training set.

6. **(3 marks)** Suppose you ran linear regression with two features twice and you also introduced regularization where $\lambda = 0$ and $\lambda = 1$. For one of the times, you got the parameters $\theta = (13.4579, 0.94, 0.78)^T$ and the other time you got the parameters $\theta = (81.3562, 12.65, 9.54)^T$. However, you forgot which value of λ generated which set of parameters. Which set of parameters belongs to $\lambda = 0$ and $\lambda = 1$ and why?

Question 2 - 25 marks

Introduction

Suppose we have an electrical engineering problem concerning Light Emitting Diodes (LEDs) where our goal is to create a prediction model where given the current (in milliamps) x flowing through the diode, we want to determine what the voltage (in millivolts) y across the diode would be. We have made the following measurements on a particular LED chosen for our experiments.

Example	Current x (mA)	Voltage y (mV)
1	1	2
2	2.5	3.75
3	4	5.2
4	6	7.9

Because of the highly non-linear relationship between the current and voltage, we will be using new features to allow for higher accuracy. Specifically, we will use two features for our prediction model: $x_1 = x^2$ and $x_2 = \sqrt{x}$.

Parts to solve

1. **(2 marks)** Using the measurements above, complete the following table in order to create these two features. You may complete the table here or do this in your response booklets (**turn to the next page**):

Example	x_1	x_2
1		
2		
3		
4		

2. **(7 marks)** Because of the wide dynamic range, we will need to **feature normalize** the above table before we create our prediction model. Normalize the features above so that they exhibit **zero mean and unit variance**. You can complete the table below here or do this in your response booklets:

Example	x_1	x_2
1		
2		
3		
4		

3. **(12 marks)** With the **normalized features**, using a learning rate $\alpha = 0.75$, the **regularization parameter** $\lambda = 0.5$ and with the initial parameters $\theta_0 = \theta_1 = \theta_2 = 1$, compute $N = 1$ iteration and state the parameters $\theta_0, \theta_1, \theta_2$ as well as the cost $J(\theta)$.

4. **(4 marks)** Using the learned parameters found in the previous step after the second iteration, predict what the output voltage would be if the input current is 2 mA and 3 mA.

Question 3 - 45 marks

Introduction

Suppose we have a machine learning problem focusing in the medical diagnosis area. We are concerned with classifying whether someone is either (1) not ill, (2) has a cold or (3) has the flu. For our training examples, we will be using two input features where x_1 is the temperature (in Celsius) of the subject and x_2 is the number of solid meals the subject has consumed 24 hours before the temperature of the subject was taken. We have measured the temperatures and food intake amounts for six (6) patients seen below, as well as the corresponding diagnoses:

Example	Temperature x_1 (^{o}C)	# of solid meals x_2	Status	Label y
1	37	3	Not ill	1
2	37.2	2	Not ill	1
3	36.8	1	Cold	2
4	37.3	1	Cold	2
5	38	1	Flu	3
6	38.5	0	Flu	3

Please turn to the next page!

Parts to solve

1. **(9 marks)** Because of the wide dynamic range, we will need to **feature normalize** the above table before we create our prediction model. Normalize the features above so that they exhibit **zero mean and unit variance**. You can complete the table below here or do this in your response booklets:

Example	x_1	x_2
1		
2		
3		
4		
5		
6		

2. **(30 marks)** With the **normalized features**, using a learning rate of $\alpha = 1$ and with the initial parameters $\theta_0 = \theta_1 = \theta_2 = 0.5$, use $N = 1$ iterations and find the parameters for each class $\Theta_1, \Theta_2, \Theta_3$ using gradient descent under the One-Vs-All multiclass classification. State the parameters at each iteration. Remember, you have ***three (3)*** sets of parameters to report at each iteration.

3. **(6 marks)** Using the following test data, determine what the corresponding class / labels would be using the parameters you found in the previous step.

Example	x_1	x_2
1	37.5	2
2	38	0

Please turn to the next page!

Question 4 - 15 marks

Introduction

Suppose we have a machine learning problem where we have 10 training examples and two features. We also know that the examples fall into two classes and the features describing this problem are Gaussian distributed. The training example features as well as the classes they belong to are shown below:

Example	x_1	x_2	Class
1	0.8663	0.5371	ω_2
2	-0.6757	1.1227	ω_1
3	0.4452	1.8456	ω_2
4	1.5644	0.0891	ω_1
5	0.7014	-0.5329	ω_2
6	0.6513	1.0900	ω_2
7	1.2643	-0.8235	ω_1
8	1.1680	-0.7951	ω_1
9	-0.5350	0.1214	ω_1
10	-0.1050	0.8343	ω_2

We also know that over all training examples, the variances for both features are drawn from a Gaussian distribution where $\sigma^2 = 2$ and there is a positive correlation of 0.5 between the two features.

Parts to solve

1. **(2 marks)** Compute the prior probablities for both classes: $P(\omega_1), P(\omega_2)$.
2. **(4 marks)** Compute the mean feature vectors μ_1, μ_2.
3. **(1 mark)** State the **combined** covariance matrix Σ. Remember that this problem considers the covariance matrix as being shared between the two classes.
4. **(8 marks)** Given the following input features, determine which class input example belongs to:

Example	x_1	x_2
1	0.5	-0.5
2	1	1