[redacted]

Department of Electrical, Computer, and Biomedical Engineering

ELE 888-Intelligent Systems

Mid-term Examination-Wednesday, June 13th, 2018

Aids Permitted

1. Non-programmable, scientific calculator with no graphing or text-based capabilities.
2. $8.5'' \times 11''$ (letter-sized) double-sided, hand-written formula/cheat sheet. No photocopies or computer-typed version allowed.

Instructions for this midterm examination

Please read before starting and good luck!

1. This is a closed book in-class examination, However a double sided, hand-written formula/cheat sheet is allowed.
2. This examination consists of **ONE (1)** concept/theoretical based questions, One set of multiple choice questions and **TWO (2)** 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 responses booklets that contain your answers..
5. Any assumptions should clearly be stated for consideration of any credit. Simply writing down the answer without any explanation will not be granted marks.
6. Show all calculations and steps taken to arrive at your answers to get full marks.

Question 1 [20 marks] (definitions)

1- Describe Supervised Learning and Unsupervised Learning with examples.
2- Define the concept of skewed data in machine learning?
3- Explain what the problem of over fitting is and explain how you would reduce over fitting in machine learning problems.
4- What is the purpose of the learning rate α, what happens if this value is too small? What happens if this value is too large? Is there a strategy you can use to select the right learning rate to ensure that the gradient descent algorithm converges
5- Is it possible that using Gradient descent in linear regression we come up with different values for minimum of the cost function? Why?
6- What is the hypothesis /model in machine learning? Draw a Block Diagram and explain
7- Is there any difference between the sigmoid and hyperbolic tangent (tanh)activation function? Clarify
8- What is the similarity between logistic regression and Neural Network?

9- Is it important how we initialize the parameters for gradient decent? What effect does the initialization have in gradient descent? How about initializing the parameters in Neural Networks?

Multiple Choice Questions [20 marks] – 2 negative marks for each 4 questions:

1. A computer program is said to learn from experience E with respect to some task T and some performance measure P if its performance on T, as measured by P, improves with experience E. Suppose we feed a learning algorithm a lot of historical weather data, and have it learn to predict weather. In this setting, what is T?
   - The weather prediction task.
   - The process of the algorithm examining a large amount of historical weather data.
   - None of these.
   - The probability of it correctly predicting a future date's weather.

2. Suppose you are working on stock market prediction. You would like to predict whether or not a certain company will declare bankruptcy within the next 7 days (by training on data of similar companies that had previously been at risk of bankruptcy). Would you treat this as a classification or a regression problem?
   - Regression
   - Classification.

3. Some of the problems below are best addressed using a supervised learning algorithm, and the others with an unsupervised learning algorithm. Which of the following would you apply supervised learning to? (Select all that apply.) In each case, assume some appropriate Dataset is available for your algorithm to learn from.
   - Have a computer examine an audio clip of a piece of music, and classify whether or not there are vocals (i.e., a human voice singing) in that audio clip, or if it is a clip of only musical instruments (and no vocals).
   - Given genetic (DNA) data from a person, predict the odds of him/her developing diabetes over the next 10 years.
   - Given data on how 1000 medical patients respond to an experimental drug (such as effectiveness of the treatment, side effects, etc.), discover whether there are different categories or "types" of patients in terms of how they respond to the drug, and if so what these categories are.
   - Given a large dataset of medical records from patients suffering from heart disease, try to learn whether there might be different clusters of such patients for which we might tailor separate treatments.

4. Many substances that can burn (such as gasoline and alcohol) have a chemical structure based on carbon atoms; for this reason they are called hydrocarbons. A chemist wants to understand how the number of carbon atoms in a molecule affects how much energy is released when that molecule combusts (meaning that it is burned). The chemist obtains the dataset below. In the column on the right, "kJ/Mol" is the unit measuring the amount of energy released.

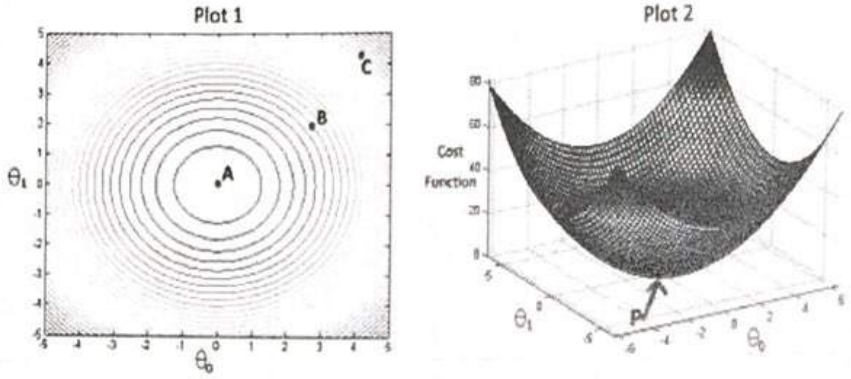
Name of molecule	Number of hydrocarbons in molecule (x)	Heat release when burned (kJ/mol) (y)
methane	1	-890
ethene	2	-1411
ethane	2	-1560
propane	3	-2220
cyclopropane	3	-2091
butane	4	-2878
pentane	5	-3537
benzene	6	-3268
cycloexane	6	-3920
hexane	6	-4163
octane	8	-5471
napthalene	10	-5157

You would like to use linear regression $h_\theta(x) = \theta_0 + \theta_1 x)$ to estimate the amount of energy released (y) as a function of the number of carbon atoms (x). Which of the following do you think will be the values you obtain for θ_0 band θ_1? You should be able to select the right answer without actually implementing linear regression.

$\theta_0 = -569.6, \theta_1 = -530.9$
$\theta_0 = -569.6, \theta_1 = 530.9$
$\theta_0 = -1780.0, \theta_1 = 530.9$
$\theta 0 = -1780.0, \theta 1 = -530.9$

5. In the given figure, the cost function J (θ_0,θ_1) has been plotted against θ_0 and θ_1, as shown in 'Plot 2'. The contour plot for the same cost function is given in 'Plot 1'. Based on the figure, choose the correct options (check all that apply).

Plots for Cost Function $J(\theta_0, \theta_1)$



If we start from point B, gradient descent with a well-chosen learning rate will eventually help us reach at or near point A, as the value of cost function $J(\theta_0,\theta_1)$ is maximum at point A. Point P (The global minimum of plot 2) corresponds to point C of Plot 1.

If we start from point B, gradient descent with a well-chosen learning rate will eventually help us reach at or near point A, as the value of cost function $J(\theta_0,\theta_1)$ is minimum at A. Point P (the global minimum of plot 2) corresponds to point A of Plot 1.

If we start from point B, gradient descent with a well-chosen learning rate will eventually help us reach at or near point C, as the value of cost function $J(\theta_0,\theta_1)$ is minimum at point C.

6. You run gradient descent for 15 iterations with α=0.3 and compute J(θ) after each iteration. You find that the value of J(θ) increases over time. Based on this, which of the following conclusions seems most plausible?

   α=0.3 is an effective choice of learning rate.

   Rather than use the current value of α, it'd be more promising to try a larger value of (say α=1.0).

   Rather than use the current value of α, it'd be more promising to try a smaller value of \alphaα (say α=0.1).

7. Suppose you have a dataset with m=50 examples and n=15 features for each example. You want to use multivariate linear regression to fit the parameters θ to our data. Should you prefer gradient descent or the normal equation?

   The normal equation, since it provides an efficient way to directly find the solution.

   Gradient descent, since it will always converge to the optimal θ.

   Gradient descent, since $(X^TX)^{-1}$ will be very slow to compute in the normal equation.

   The normal equation, since gradient descent might be unable to find the optimal θ.

8. Which of the following statements are true? Check all that apply.

For logistic regression, sometimes gradient descent will converge to a local minimum (and fail to find the global minimum).

The cost function $J(\theta)$ for logistic regression trained with $m \geq 1$ examples is always greater than or equal to zero.

Linear regression always works well for classification if you classify by using a threshold on the prediction made by linear regression.

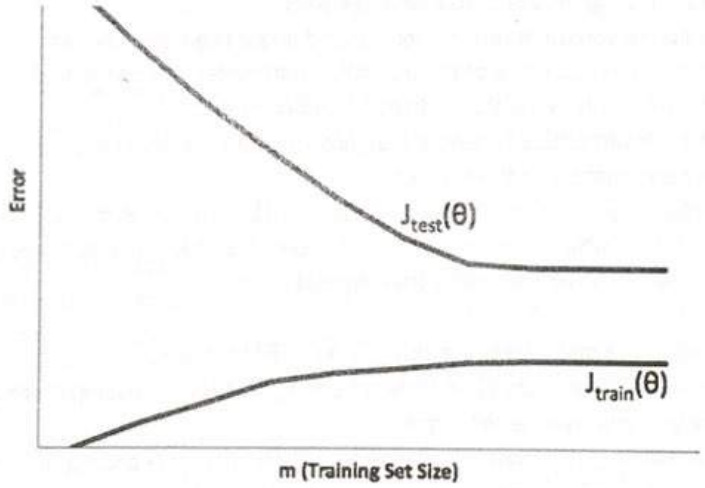
The sigmoid function $g(z)=(1/1+e-z)$ is never greater than one (>1).

9. Suppose you ran logistic regression twice, once with $\lambda=0$, and once with $\lambda=1$. One of the times, you got parameters $\begin{bmatrix} 26.29 \\ 65.41 \end{bmatrix}$, and the other time you got $\begin{bmatrix} 2.75 \\ 1.32 \end{bmatrix}$. However, you forgot which value of λ corresponds to which value of θ. Which one do you think corresponds to $\lambda=1$?

$\begin{bmatrix} 2.75 \\ 1.32 \end{bmatrix}$

$\begin{bmatrix} 26.29 \\ 65.41 \end{bmatrix}$

10. You train a learning algorithm, and find that it has unacceptably high error on the test set. You plot the learning curve, and obtain the figure below. Is the algorithm suffering from high bias, high variance, or neither?



High bias

Neither

High variance

11. Suppose you have implemented regularized logistic regression to classify what object is in an image (i.e., to do object recognition). However, when you test your hypothesis on a new set of images, you

find that it makes unacceptably large errors with its predictions on the new images. However, your Hypothesis performs well (has low error) on the training set. Which of the following are promising steps to take? Check all that apply.

- Try decreasing the regularization parameter λ.
- Try increasing the regularization parameter λ
- Try using a smaller set of features.
- Try evaluating the hypothesis on a cross validation set rather than the test set.

12. Suppose you have implemented regularized logistic regression to predict what items customers will purchase on a web shopping site. However, when you test your hypothesis on a new set of customers, you find that it makes unacceptably large errors in its predictions. Furthermore, the hypothesis performs poorly on the training set. Which of the following might be promising steps to take? Check all that apply.

- Try adding polynomial features.
- Try evaluating the hypothesis on a cross validation set rather than the test set.
- Try decreasing the regularization parameter λ.
- Use fewer training examples.

13. Which of the following statements are true? Check all that apply.

- Suppose you are training a regularized linear regression model. The recommended way to choose what value of regularization parameter λ to use is to choose the value of λ which gives the lowest **test set** error.
- Suppose you are training a regularized linear regression model. The recommended way to choose what value of regularization parameter λ to use is to choose the value of λ which gives the lowest **cross validation** error.
- The performance of a learning algorithm on the training set will typically be better than its performance on the test set
- Suppose you are training a regularized linear regression model. The recommended way to choose what value of regularization parameter λ to use is to choose the value of λ which gives the lowest **training set** error.

14. Which of the following statements are true? Check all that apply.

- If a learning algorithm is suffering from high variance, adding more training examples is likely to improve the test error.
- If a learning algorithm is suffering from high bias, only adding more training examples may **not** improve the test error significantly.
- A model with more parameters is more prone to overfitting and typically has higher variance.
- If the training and test errors are about the same, adding more features will **not** help improve the results.

15. You are working on a spam classification system using regularized logistic regression. "Spam" is a positive class (y = 1) and "not spam" is the negative class (y = 0). You have trained your classifier and

there are m = 1000 examples in the cross-validation set. The chart of predicted class vs. actual class is:

	Actual Class: 1	**Actual Class: 0**
Predicted Class: 1	85	890
Predicted Class: 0	15	10

For reference:

- Accuracy = (true positives + true negatives) / (total examples)
- Precision = (true positives) / (true positives + false positives)
- Recall = (true positives) / (true positives + false negatives)
- F_1 score = (2 * precision * recall) / (precision + recall)

What is the classifier's accuracy (as a value from 0 to 1)? Write your answer in the below. If necessary, provide at least two values after the decimal point.

The value is: ----------

16. Suppose a massive dataset is available for training a learning algorithm. Training on a lot of data is likely to give good performance when two of the following conditions hold true. Which are the two?

    The classes are not too skewed.

    Our learning algorithm is able to represent fairly complex functions (for example, if we train a neural network or other model with a large number of parameters).

    A human expert on the application domain can confidently predict y when given only the features x (or more generally, if we have some way to be confident that x contains sufficient information to predict y accurately).

    When we are willing to include high order polynomial features of x (such as $x_1^2, x_1x_2, x_1x_2^2$, $x_1^2x_2$, etc....)

17. Suppose you have trained a logistic regression classifier which is outputting $h_\theta(x)$.Currently, you predict 1 if $h_\theta(x) \geq$ threshold, and predict 0 if $h_\theta(x) <$ threshold, where currently the threshold is set to 0.5.Suppose you increase the threshold to 0.7. Which of the following are true? Check all that apply.

    The classifier is likely to now have higher precision

    The classifier is likely to have unchanged precision and recall, and thus the same F_1 score.

    The classifier is likely to now have higher recall.

    The classifier is likely to have unchanged precision and recall, but higher accuracy.

18. Suppose you are working on a spam classifier, where spam emails are positive examples (y=1) and non-spam emails are negative examples (y=0). You have a training set of emails in which 99% of the emails are non-spam and the other 1% is spam. Which of the following statements are true? Check all that apply.

A good classifier should have both a high precision and high recall on the cross validation set.

If you always predict non-spam (output y=0), your classifier will have an accuracy of 99%.

If you always predict non-spam (output $y=0$), your classifier will have 99% accuracy on the training set, but it will do much worse on the cross validation set because it has over fit the training data.

If you always predict non-spam (output $y=0$), your classifier will have 99% accuracy on the training set, and it will likely perform similarly on the cross validation set.

19. Which of the following statements are true? Check all that apply.

If your model is under fitting the training set, then obtaining more data is likely to help.

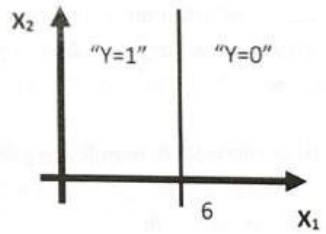
After training a logistic regression classifier, you **must** use 0.5 as your threshold for predicting whether an example is positive or negative.

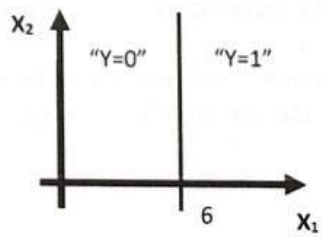
The "error analysis" process of manually examining the examples which your algorithm got wrong can help suggest what are good steps to take (e.g., developing new features) to improve your algorithm's performance.

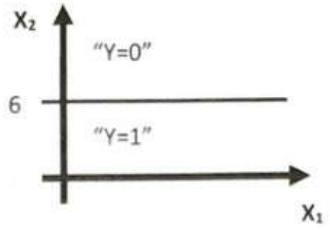
Using a **very large** training set makes it unlikely for model to over fit the training data.

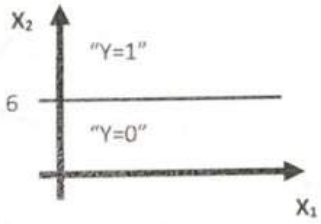
It is a good idea to spend a lot of time collecting a **large** amount of data before building your first version of a learning algorithm.

20 .Suppose you train a logistic classifier $h_\theta = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$. suppose $\theta_0 = 6, \theta_1 = -1, \theta_2 = 0$, which of the following figures represents the decision boundary found by your classifier?







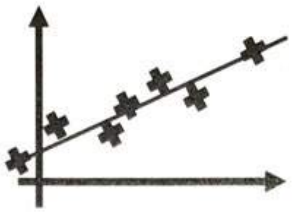


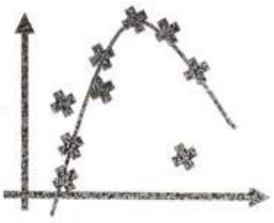
21. Suppose you ran logistic regression twice, once with λ =0, and once with λ=1, for (n) times you got:

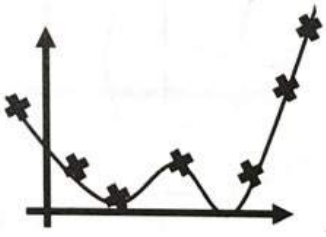
Parameters $\theta = \begin{bmatrix}26.29\\65.41\end{bmatrix}$, and the other time you got $\theta = \begin{bmatrix}2.75\\1.32\end{bmatrix}$, however, you forgot which values of λ corresponding to which value of θ, which one do you think corresponds to λ=1?

- $\theta = \begin{bmatrix}2.75\\1.32\end{bmatrix}$
- $\theta = \begin{bmatrix}26.29\\65.41\end{bmatrix}$

22.In which one of the following figures do you think the hypothesis has overfit the training set?







Question 1 (multi variable Linear Regression & Regularization)

Suppose we have an thermal consumption problem concerning a specific IC (Integrated Circuit) operation, where our goal is to create a prediction model where given the current (in milliamps) flowing through the IC, we want to determine what the temperature changes (ΔT) (in C°) across the microprocessor would be. We have made the following measurements on an IC chosen for our experiments.

k (sample)	Current (mA)	Temperature changes (C°)	Output
1	2.25	1.224	2
2	7.5625	1.563	3.75
3	20.25	2.1213	5.5
4	33.0625	2.3979	7.5

Note : Use the following feature -normalized formula for each feature (x_k) at sample k :

$$x_{k(normlized)} = \frac{x_k - \mu_k}{s_k}$$

Where: (m= No. of samples)

$$\mu_k = \frac{1}{m}\sum_{i=1}^{m} x_k^{(i)}$$

$$s_k = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(x_k^{(i)} - \mu_k)^2}$$

- **Questions to Solve**

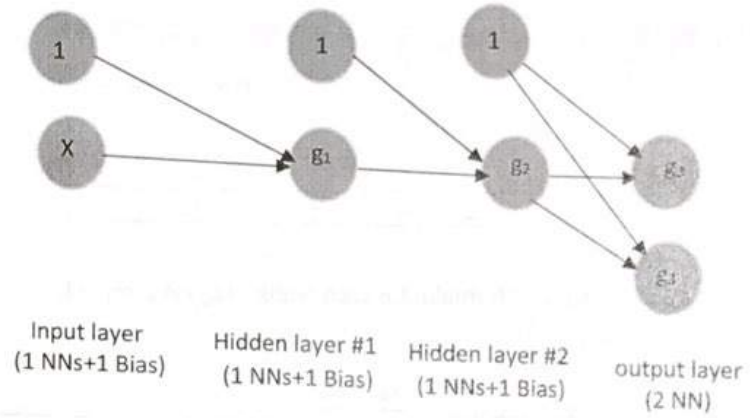
1. Because of the wide dynamic range, Normalize the features above before we create our prediction model. Normalize the features above so that they exhibit zero mean and unit variance.

2. With the normalized features, using a learning rate α=1, the regularization parameter λ=0.75 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(θ).

3. Using the learned parameters found in the previous step after the first iteration, predict what the output temperature changes would be if the input current is 1 mA and 3 mA.

Question 2 (NN method)

Suppose we have the following Neural Network architecture shown below; The network is composed of 1 input layer, 2 hidden layers and 1 output layer. There are 1 input feature, 1 hidden neurons in the first hidden layer, 1 hidden neuron in the second hidden layer and 2 outputs neuron in the output layer. Take note that we didn't count the bias units in the total number of neurons. Assume that each layer except the input layer has same activation functions .



The activation function we will use is the hyperbolic tangent function: **tanh**. Recall that this activation function as well as its derivative corresponds to:

$$g_1 = g_2 = g_3 = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

$$g' = (1 - (g(z))^2)$$

Suppose we have the following two training example that we are to use to train this neural network:

Sample	X	Y_1	Y_2
1	2	-0.5	0.5

Where X is input feature and their expected outputs are y_1 & y_2. Additionally, learning rate α=0.6, compute **One Epoch** and the following parameters for each sample at every epoch:

$$W^{(1)} = \begin{bmatrix} 0.6 \\ 0.4 \end{bmatrix}; W^{(2)} = \begin{bmatrix} 0.02 \\ 0.05 \end{bmatrix}, W^{(3)} = \begin{bmatrix} 0.2 & 0.1 \\ 0.6 & 0.5 \end{bmatrix}$$

Questions to Solve

1- Redraw the above neural network with weights
2- State what the updated weight matrices are:

$(w^{(1)}, w^{(2)}, w^{(3)}, X^{(0)}, S^{(1)}, X^{(1)}, S^{(2)}, X^{(2)}, S^{(3)}, X^{(3)}, \delta^{(1)}, \delta^{(2)}, \delta^{(3)}, \frac{\partial e}{\partial w^{(1)}}, \frac{\partial e}{\partial w^{(2)}}, , \frac{\partial e}{\partial w^{(3)}})$