Carnegie Mellon University in Qatar

AI for Medicine

15-182 - Spring 2023

Assignment 3

Name: _____

Andrew ID: _____

Due on: March 16, 2023 by midnight

Instructions:

- This assignment consists of two problems. Solve them both.
- Submit your solution through Gradescope.

Question	Points	Score
Learning a Linear Regression Model	60	
Malignant or Benign Tumour	40	
Total:	100	

Problem 1: Learning a Linear Regression Model (60 Points)

As an example of supervised learning, consider the four points (1, 2), (2, 1), (3, 4), and (4, 3). These points can be thought of as a training set, wherein each point (x, y) includes a feature x and an associated label y. For instance, the point (1, 2) assumes $x = 1$ and $y = 2$; the other points can be interpreted similarly.

Suppose we want to "learn" the **hypothesis function** $h_\theta(x) = \theta_0 + \theta_1 x$ that best fits the points of the training set. A natural interpretation of "best fits" is that a **cost function**, say, the **Mean Squared Error** (MSE) of the value of $h_\theta(x)$ compared with the value of y over all the given points in the training set is minimized.

Answer the following questions *after you show all your work.*

12pts (a) What will be the value of MSE if:
- i. $\theta_0 = 0$ and $\theta_1 = 1$
- ii. $\theta_0 = 1$ and $\theta_1 = 0$
- iii. $\theta_0 = 1$ and $\theta_1 = 3$
- iv. $\theta_0 = 1$ and $\theta_1 = 3/5$

20pts (b) Starting off with $\theta_0 = 0$ and $\theta_1 = 0$, what will be the values of θ_0 and θ_1 after 3, 6, and 9 rounds of running gradient descent, assuming:
- i. The learning rate $\alpha = 0.5$
- ii. The learning rate $\alpha = 0.2$
- iii. The learning rate $\alpha = 0.7$

4pts (c) For each number of rounds and learning rate in part (b), state whether the learnt line was underfitting, properly fitting, or overfitting the data.

8pts (d) How do the learning rate, number of rounds, and amount of training data impact fitting the data? Explain with proper reasoning.

8pts (e) Solve this problem (i.e., finding the optimal θ_0 and θ_1 that best fit the given training set) through a system of equations rather than gradient descent. What are the optimal solutions for θ_0 and θ_1? (*Hint: think of taking the derivatives of the cost function with respect to θ_0 then with respect to θ_1*). Show all your work.

8pts (f) Say, we are trying to learn regression parameters for a dataset which we know was generated from a polynomial of a certain degree, but we do not know what this degree is. Assume the data was actually generated from a polynomial of degree 3 with some added noise, that is, $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \epsilon$. Assume we use 50 {x,y} pairs for training and additional 50 {x,y} pairs for testing. Since we do not know the degree of the polynomial, we will train two models, **A** which learns parameters for a polynomial of degree 2, and **B** which learns parameters for a polynomial of degree 4. Which model, **A** or **B**, is likely to fit the test data better? Explain with proper reasoning.

Assignment continues on the next page(s)

Problem 2: Malignant or Benign Tumour (40 Points)

Let us consider training a model to recognize whether a tumour is malignant or benign in a given input (e.g., an MRI represented in a specific format). To this end, suppose you are given a training set that consists of pairs (x, y), where x is a vector of 0's and 1's, with each component x_i in the vector corresponding to the presence ($x_i = 1$) or absence ($x_i = 0$) of a particular feature in the input. The value of y is +1 if the input is known to have a malignant tumour and -1 if it is known to have a benign one.

In this problem, we will assume that there are five features, namely, x_1, x_2, x_3, x_4, and x_5 that dictate whether a tumour is malignant or benign in any given input. The table below shows a training set with six examples and their corresponding classes.

	x_1	x_2	x_3	x_4	x_5	y
Example **1**	1	1	0	1	1	+1
Example **2**	0	0	1	1	0	-1
Example **3**	0	1	1	0	0	+1
Example **4**	1	0	0	1	0	-1
Example **5**	1	0	1	0	1	+1
Example **6**	1	0	1	1	0	-1

Figure 1: Six examples, each represented as a vector of 1's and 0's, where 1 or 0 indicates the presence or absence of a certain feature x_i, respectively. The label y indicates whether the example has a malignant (y = +1) or benign (y = -1) tumour.

Suppose we want to "learn" the **hypothesis function** $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5$ that best fits the examples of the training set. For this sake, let us consider using Mean Squared Error (MSE) as a **cost function** and transforming the problem into a classification problem via using *thresholding* (*Note*: As we will see later in class when covering logistic regression, this is not necessarily a good way to do classification). In particular, if $h_\theta(x) > 0$, the tumour of the given input will be considered malignant, and if $h_\theta(x) < 0$, the tumour will be deemed benign (the special case where $h_\theta(x) = 0$ will be assumed "wrong").

25pts (a) Assuming a learning rate $\alpha = 0.5$, what will be the values of θ_0, θ_1, θ_2, θ_3, θ_4, and θ_5 after one round of running gradient descent, if we start off with $\theta_0 = 0$, $\theta_1 = 0$, $\theta_2 = 0$, $\theta_3 = 0$, $\theta_4 = 0$, and $\theta_5 = 0$? Show all your work.

8pts (b) Based on the above defined classification approach and the θs learnt from one round of running gradient descent, how many of the given examples will be classified correctly? Show your work.

7pts (c) Based on the above defined classification approach and the θs learnt from one round of running gradient descent, what will be the tumour type of this new unknown example $[x_1 = 0, x_2 = 1, x_3 = 0, x_4 = 0, x_5 = 1]$?