Name – Surname:
Number:

Q.I want to predict the grade of a student by looking at his/her attendance. A past data from one of my lectures is given in the following:

	X (no. of lectures attended)	Y (Overall Grade)
Ismail	10	95
Ebru	8	80
Tugce	2	50

Let Model 1 be the model with $\theta_0 = 40$, $\theta_1 = 5$ and Model 2 be the model with $\theta_0 = 42$, $\theta_1 = 5$. Recall that

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

a) (2) What is h(5) for model 1?

b) (4) What is $J(\theta_0, \theta_1)$ for model 1?

c) (6) What is $J(\theta_0, \theta_1)$ for Model 2?

d) (3) Will you prefer model 1 or model 2? Why?

e) (12) Up to now we always used the squared difference between the prediction (hypothesis) and the observed value in calculating the total cost value J(θ). However, there are some researchers who suggests using the **absolute difference** between the prediction and the real value. Let's try to see what happens if we work that way. In another words, assume the cost function is given by:

$$J'(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left|h_\theta\left(x^{(i)}\right) - y^{(i)}\right|$$

Please find the total cost of two models with respect to the new cost function $J'(\theta)$

f) (3) Does your answer to (d) change with this new cost function? Please explain.

Q. Please write the type of the following problems. Use LR for Linear Regression, LOG for Logistic Regression, and UL for Unsupervised Learning (UL). Each incorrect answer will **cost you -2** points::

	Given email labeled as spam/not spam, learn a spam filter.
	Given a set of news articles found on the web, group them into set of articles about the same story
	Given a database of customer data, automatically discover market segments and group customers into different market segments
	Given the features, the prices and whether the house is sold or still not sold, try to find a house whether can be sold or cannot be sold at a given price.
	Given the height of a baby and its parents, find a height predictor of an unborn baby.

Q. Suppose $\theta_0 = 1$ and $\theta_1 = 2$ for a simple linear regression problem. We simultaneously update θ_0 and θ_1 using the rule $\theta_j = \theta_j + (\theta_1 + 2\theta_2)/100$. What are the new values of θ?

Q. Please write true or false. Each incorrect answer will **cost you -2** points:

	To make gradient descent converge, we have to decrease α slowly over time
	Gradient Descent is guaranteed to find a local or global minimum for any function f(θ_0,θ_1)
	Gradient descent can converge even if α is kept fixed (but α cannot be too large or else it may fail to converge)
	For the specific choice of cost function of the linear regression, the only minimum is the global minimum.
	Gradient descent uses the partial derivatives to find the direction of the step that you should take. If the derivative is negative you increase your estimate for the parameter θ.Otherwise you decrease the your estimate for the parameter θ.

Q Suppose you ran logistic regression twice, once with λ=0, and once with λ=1. One of the times, you got parameters θ=[74.81,45.05] and the other time you got θ=[1.37,0.51]. You forgot which λ values you used for each parameter set. Please make the associated connections. Which λ belongs to which theta values?

Q. Suppose *m*=4 students have taken some class, and the class had a midterm exam and a final exam. You have collected a dataset of their scores on the two exams, which is as follows:

Midterm	Midtem^2	Final Exam
89	7921	96
72	5184	74
94	8836	87
69	4761	78

You'd like to use polynomial regression to predict a student's final exam score from their midterm exam score. Concretely, suppose you want to fit a model of the form $h_\theta(x)=\theta_0 + \theta_1 x_1 + \theta_2 x^2$, where x_1 is the midterm score and x_2 is (midterm score)^2.

a) For $\theta_0 = -360$, $\theta_1 = 10.5$ and $\theta_2 = -0.06$ what is the error for the first data point in the data set?

b) Further, you plan to use both feature scaling (dividing by the "max-min", or range, of a feature) and mean normalization. What is the normalized feature $x_1^{(1)}$? (Hint: midterm = 89, final = 96 is training example 1.)

c) Please write true or false. Each incorrect answer will **cost you -2** points:

	If you don't use scaling here, then the optimization of the cost function will never converge.
	If you don't use scaling here, then the optimization will converge but it may be a local optima, i.e., the global optimal is not guaranteed.
	We can use regularization to make the algorithm faster.
	Regularization also reduces the total error in the training set
	If you add new features like gender of the student or (midterm)^3, then you can be sure that the total error in the training set will be reduced.
	Adding new features always results in overfitting, hence you should avoid adding new features.