LINEAR REGRESSION

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam. What is the task T in this setting?



Classify emails as spam or not spam. -- T



Watching you label emails as spam or not spam. - E



The number (or fraction) of emails correctly classified as spam/not spam. - P



None of the above, this is not a machine learning algorithm.

: You’re running a company, and you want to develop learning algorithms to address each of two problems.

Problem 1:You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.

Problem 2: You’d like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised. Should you treat these as classification or as regression problems?

Treat both as classification problems.



Treat problem 1 as a classification problem, problem 2 as a regression problem.



Treat problem 1 as a regression problem, problem 2 as a classification problem.



Treat both as regression problems.

**Example 1:**

Given data about the size of houses on the real estate market, try to predict their price. Price as a function of size is a continuous output, so this is a regression problem.

We could turn this example into a classification problem by instead making our output about whether the house "sells for more or less than the asking price." Here we are classifying the houses based on price into two discrete categories.

**Example 2**:

(a) Regression - Given a picture of a person, we have to predict their age on the basis of the given picture

(b) Classification - Given a patient with a tumor, we have to predict whether the tumor is malignant or benign.

Google news is an example of clustering algorithm

Market segmentation

Social network analysis

Of the following examples, which would you address using an unsupervised learning algorithm? (Check all that apply.)



Given email labeled as spam/not spam, learn a spam filter.



Given a set of news articles found on the web, group them into sets of articles about the same stories.



Given a database of customer data, automatically discover market segments and group customers into different market segments.



Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.

**Example:**

Clustering: Take a collection of 1,000,000 different genes, and find a way to automatically group these genes into groups that are somehow similar or related by different variables, such as lifespan, location, roles, and so on.

Non-clustering: The "Cocktail Party Algorithm", allows you to find structure in a chaotic environment. (i.e. identifying individual voices and music from a mesh of sounds at a [cocktail party](https://en.wikipedia.org/wiki/Cocktail_party_effect)).

**1. Question 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 probability of it correctly predicting a future date's weather.



None of these.



The process of the algorithm examining a large amount of historical weather data.

Question 2

1

point

**2. Question 2**

Suppose you are working on weather prediction, and your weather

station makes one of three predictions for each day's weather:

Sunny, Cloudy or Rainy. You'd like to use a learning algorithm

to predict tomorrow's weather.

Would you treat this as a classification or a regression problem?



Classification



Regression

Question 3

1

point

**3. Question 3**

Suppose you are working on stock market prediction. You would like to predict whether or not a certain company will win a patent infringement lawsuit (by training on data of companies that had to defend against similar lawsuits). Would you treat this as a classification or a regression problem?



Regression



Classification

Question 4

1

point

**4. Question 4**

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 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.



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.

Question 5

1

point

**5. Question 5**

Which of these is a reasonable definition of machine learning?



Machine learning is the science of programming computers.



Machine learning is the field of allowing robots to act intelligently.



Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.



Machine learning learns from labeled data.

Consider the training set shown below. (x^{(i)}, y^{(i)})(*x*(*i*),*y*(*i*)) is the i^{th}*ith* training example. What is y^{(3)}*y*(3)?

|  |  |
| --- | --- |
| Size in feet^22 (x*x*) | Price ($) in 1000's (y*y*) |
| 2104 | 460 |
| 1416 | 232 |
| 1534 | 315 |
| 852 | 178 |
| ... | ... |



1416



1534



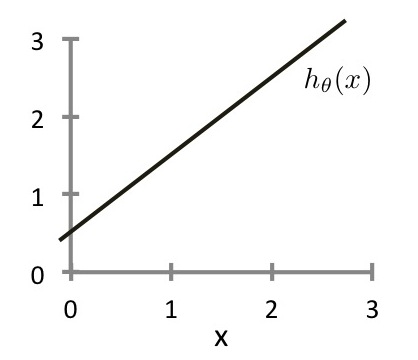
315



0

Hypothesis is a function that maps input variable with the output variable.

Consider the plot below of h\_\theta(x) = \theta\_0 + \theta\_1x*hθ*​(*x*)=*θ*0​+*θ*1​*x*. What are \theta\_0*θ*0​ and \theta\_1*θ*1​?





\theta\_0 = 0, \theta\_1 = 1*θ*0​=0,*θ*1​=1



\theta\_0 = 0.5, \theta\_1 = 1*θ*0​=0.5,*θ*1​=1

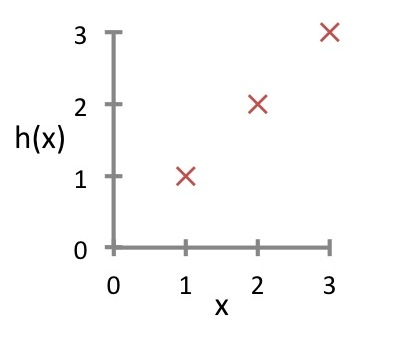


\theta\_0 = 1, \theta\_1 = 0.5*θ*0​=1,*θ*1​=0.5



\theta\_0 = 1, \theta\_1 = 1*θ*0​=1,*θ*1​=1

Suppose we have a training set with m=3 examples, plotted below. Our hypothesis representation is h\_\theta(x) = \theta\_1x*hθ*​(*x*)=*θ*1​*x*, with parameter \theta\_1*θ*1​. The cost function J(\theta\_1)*J*(*θ*1​) is J(\theta\_1) = \frac{1}{2m} \sum^m\_{i=1} (h\_\theta (x^{(i)}) - y^{(i)})^2*J*(*θ*1​)=2*m*1​∑*i*=1*m*​(*hθ*​(*x*(*i*))−*y*(*i*))2. What is J(0)*J*(0)?





0



1/6



1



14/6

Suppose \theta\_0= 1, \theta\_1= 2*θ*0​=1,*θ*1​=2, and we simultaneously update \theta\_0*θ*0​ and \theta\_1*θ*1​ using the rule:\theta\_j := \theta\_j + \sqrt{\theta\_0 \theta\_1}*θj*​:=*θj*​+*θ*0​*θ*1​​ (for j = 0 and j=1) What are the resulting values of \theta\_0*θ*0​ and \theta\_1*θ*1​?



\theta\_0 = 1, \theta\_1 =2*θ*0​=1,*θ*1​=2



\theta\_0 = 1+\sqrt{2}, \theta\_1 =2 + \sqrt{2}*θ*0​=1+2​,*θ*1​=2+2​



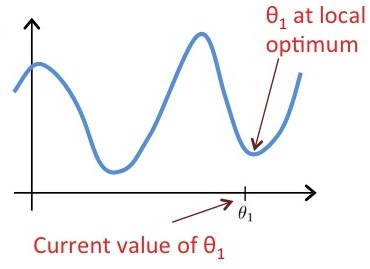
\theta\_0 = 2 + \sqrt{2}, \theta\_1 =1 + \sqrt{2}*θ*0​=2+2​,*θ*1​=1+2​



\theta\_0 = 1+\sqrt{2}, \theta\_1 =2 + \sqrt{(1 + \sqrt{2})\cdot 2}*θ*0​=1+2​,*θ*1​=2+(1+2​)⋅2​  
------------

Suppose \theta\_1*θ*1​ is at a local optimum of J(\theta\_1)*J*(*θ*1​), such as shown in the figure.

What will one step of gradient descent \theta\_1 := \theta\_1 -\alpha \frac{d}{d \theta\_1} J(\theta\_1)*θ*1​:=*θ*1​−*αdθ*1​*d*​*J*(*θ*1​) do?





Leave \theta\_1*θ*1​ unchanged



Change \theta\_1*θ*1​ in a random direction



Move \theta\_1*θ*1​ in the direction of the global minimum of J(\theta\_1)*J*(*θ*1​)



Decrease \theta\_1*θ*1​  
----------

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



To make gradient descent converge, we must slowly decrease \alpha*α* over time.



Gradient descent is guaranteed to find the global minimum for any function J(\theta\_0, \theta\_1)*J*(*θ*0​,*θ*1​).



Gradient descent can converge even if \alpha*α* is kept fixed. (But \alpha*α* cannot be too large, or else it may fail to converge.)



For the specific choice of cost function J(\theta\_0, \theta\_1)*J*(*θ*0​,*θ*1​) used in linear regression, there are no local optima (other than the global optimum).

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



To make gradient descent converge, we must slowly decrease \alpha*α* over time.



Gradient descent is guaranteed to find the global minimum for any function J(\theta\_0, \theta\_1)*J*(*θ*0​,*θ*1​).



Gradient descent can converge even if \alpha*α* is kept fixed. (But \alpha*α* cannot be too large, or else it may fail to converge.)



For the specific choice of cost function J(\theta\_0, \theta\_1)*J*(*θ*0​,*θ*1​) used in linear regression, there are no local optima (other than the global optimum).

1

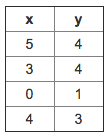
point

**1. Question 1**

Consider the problem of predicting how well a student does in her second year of college/university, given how well she did in her first year.

Specifically, let x be equal to the number of "A" grades (including A-. A and A+ grades) that a student receives in their first year of college (freshmen year). We would like to predict the value of y, which we define as the number of "A" grades they get in their second year (sophomore year).

Here each row is one training example. Recall that in linear regression, our hypothesis is h\_\theta(x) = \theta\_0 + \theta\_1x*hθ*​(*x*)=*θ*0​+*θ*1​*x*, and we use m*m* to denote the number of training examples.



For the training set given above (note that this training set may also be referenced in other questions in this quiz), what is the value of m*m*? In the box below, please enter your answer (which should be a number between 0 and 10).



Incorrect question 1

Question 2

1

point

**2. Question 2**

Consider the following training set of m=4*m*=4 training examples:

|  |  |
| --- | --- |
| x | y |
| 1 | 0.5 |
| 2 | 1 |
| 4 | 2 |
| 0 | 0 |

Consider the linear regression model h\_\theta(x) = \theta\_0 + \theta\_1x*hθ*​(*x*)=*θ*0​+*θ*1​*x*. What are the values of \theta\_0*θ*0​ and \theta\_1*θ*1​ that you would expect to obtain upon running gradient descent on this model? (Linear regression will be able to fit this data perfectly.)



\theta\_0 = 0.5, \theta\_1 = 0*θ*0​=0.5,*θ*1​=0



\theta\_0 = 1, \theta\_1 = 0.5*θ*0​=1,*θ*1​=0.5



\theta\_0 = 1, \theta\_1 = 1*θ*0​=1,*θ*1​=1



\theta\_0 = 0 , \theta\_1 = 0.5*θ*0​=0,*θ*1​=0.5



\theta\_0 = 0.5, \theta\_1 = 0.5*θ*0​=0.5,*θ*1​=0.5

1

point

**3. Question 3**

Suppose we set \theta\_0 = -1, \theta\_1 = 0.5*θ*0​=−1,*θ*1​=0.5. What is h\_{\theta}(4)*hθ*​(4)?



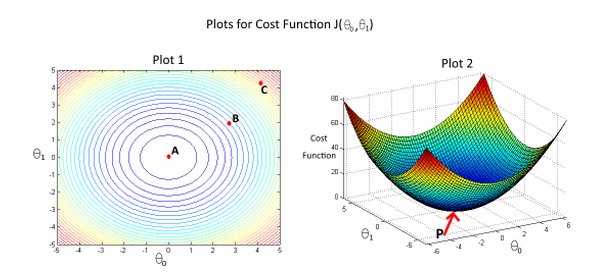
Question 4

1

point

**4. Question 4**

In the given figure, the cost function J(\theta\_0,\theta\_1)*J*(*θ*0​,*θ*1​) has been plotted against \theta\_0*θ*0​ and \theta\_1*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).





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)*J*(*θ*0​,*θ*1​) is minimum at point C.



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)*J*(*θ*0​,*θ*1​) is minimum at A.



Point P (The global minimum of plot 2) corresponds to point C of Plot 1.



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 A, as the value of cost function J(\theta\_0,\theta\_1)*J*(*θ*0​,*θ*1​) is maximum at point A.

Question 5

1

point

**5. Question 5**

Suppose that for some linear regression problem (say, predicting housing prices as in the lecture), we have some training set, and for our training set we managed to find some \theta\_0*θ*0​, \theta\_1*θ*1​ such that J(\theta\_0, \theta\_1)=0*J*(*θ*0​,*θ*1​)=0.

Which of the statements below must then be true? (Check all that apply.)



For this to be true, we must have y^{(i)} = 0*y*(*i*)=0 for every value of i = 1, 2, \ldots, m*i*=1,2,…,*m*.



Our training set can be fit perfectly by a straight line,

i.e., all of our training examples lie perfectly on some straight line.



For this to be true, we must have \theta\_0 = 0*θ*0​=0 and \theta\_1 = 0*θ*1​=0

so that h\_\theta(x) = 0*hθ*​(*x*)=0



Gradient descent is likely to get stuck at a local minimum and fail to find the global minimum.

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



⎡⎣140201⎤⎦ is a 3\times23×2 matrix.



[03144029] is a 4\times24×2 matrix.



⎡⎣03544−1290⎤⎦ is a 3\times33×3 matrix.



[12] is a 1\times21×2 matrix.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Size (feet)^22 | Number of bedrooms | Number of floors | Age of home (years) | Price ($1000) |
| 2104 | 5 | 1 | 45 | 460 |
| 1416 | 3 | 2 | 40 | 232 |
| 1534 | 3 | 2 | 30 | 315 |
| 852 | 2 | 1 | 36 | 178 |
| ... | ... | ... | ... | ... |

In the training set above, what is x\_1^{(4)}*x*1(4)​?



The size (in feet^22) of the 1st home in the training set



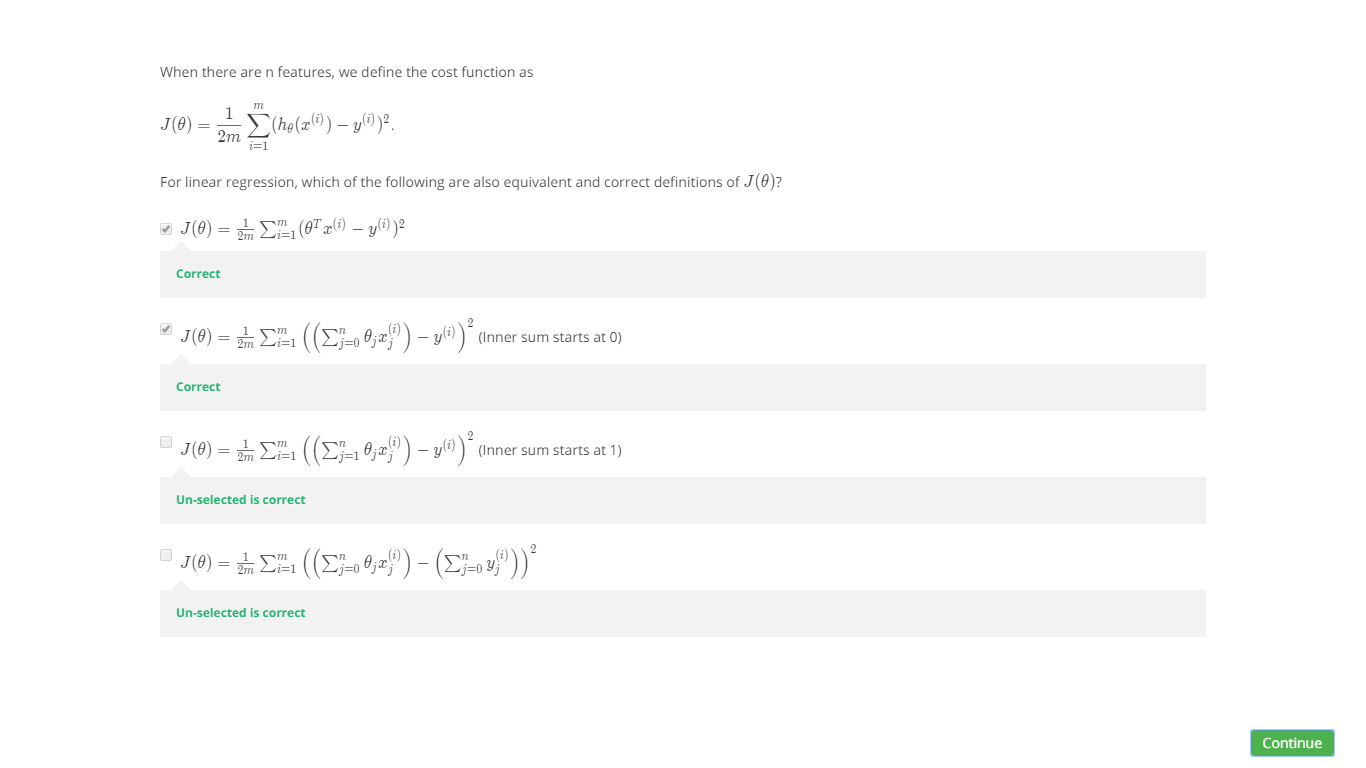
The age (in years) of the 1st home in the training set



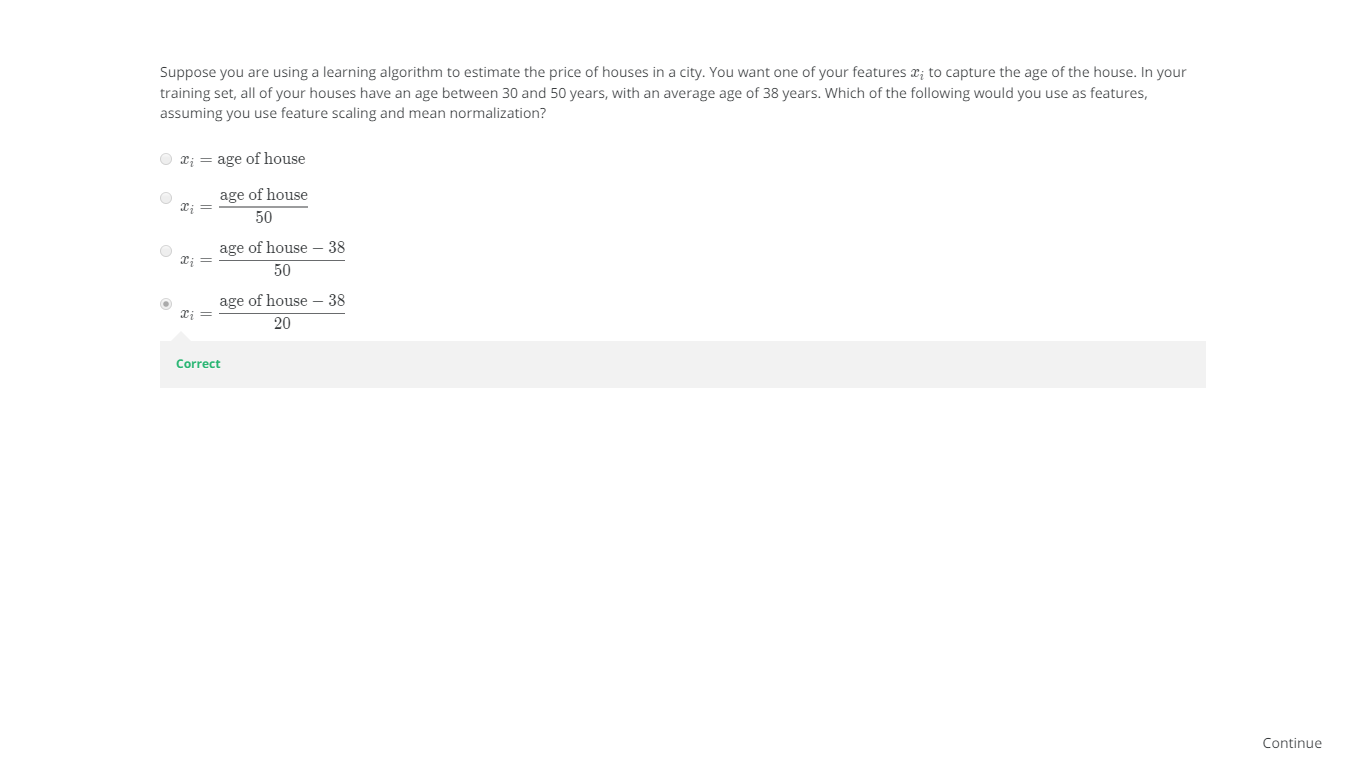
The size (in feet^22) of the 4th home in the training set

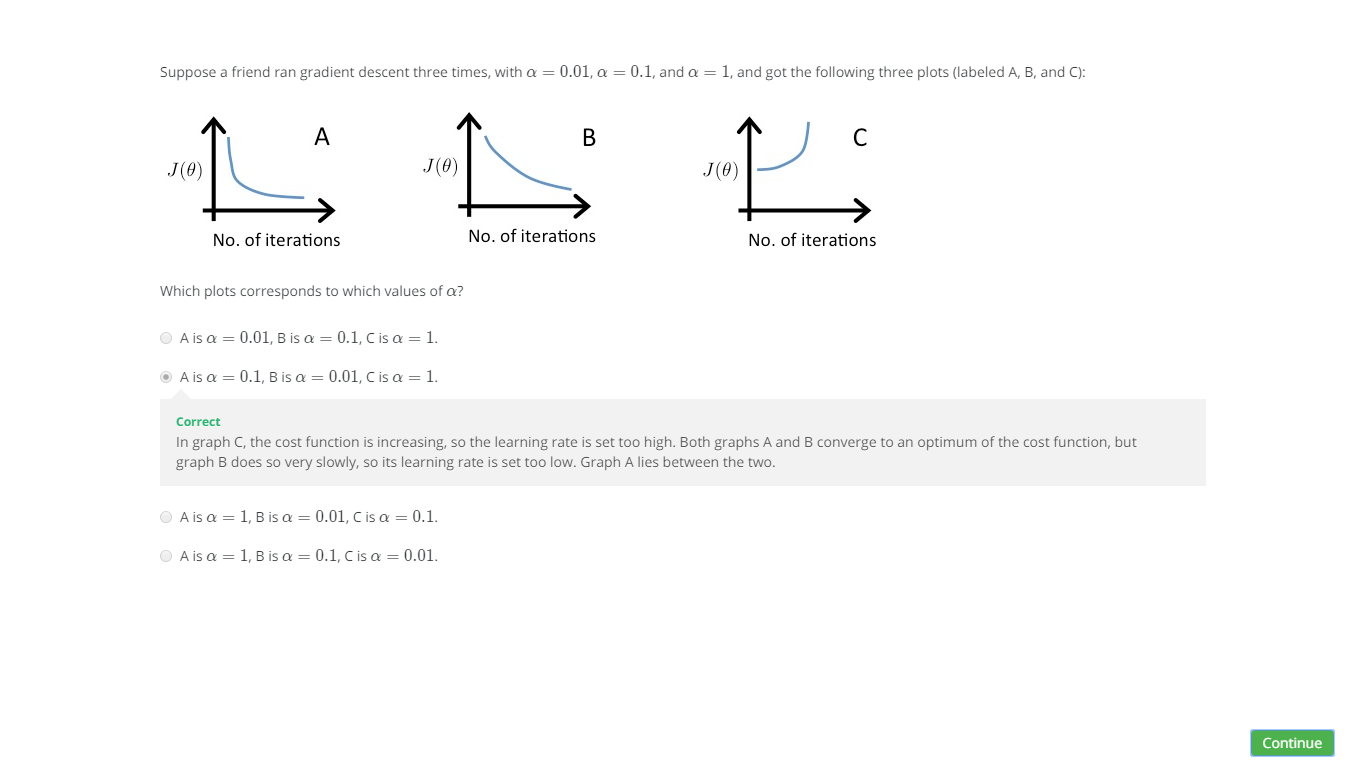


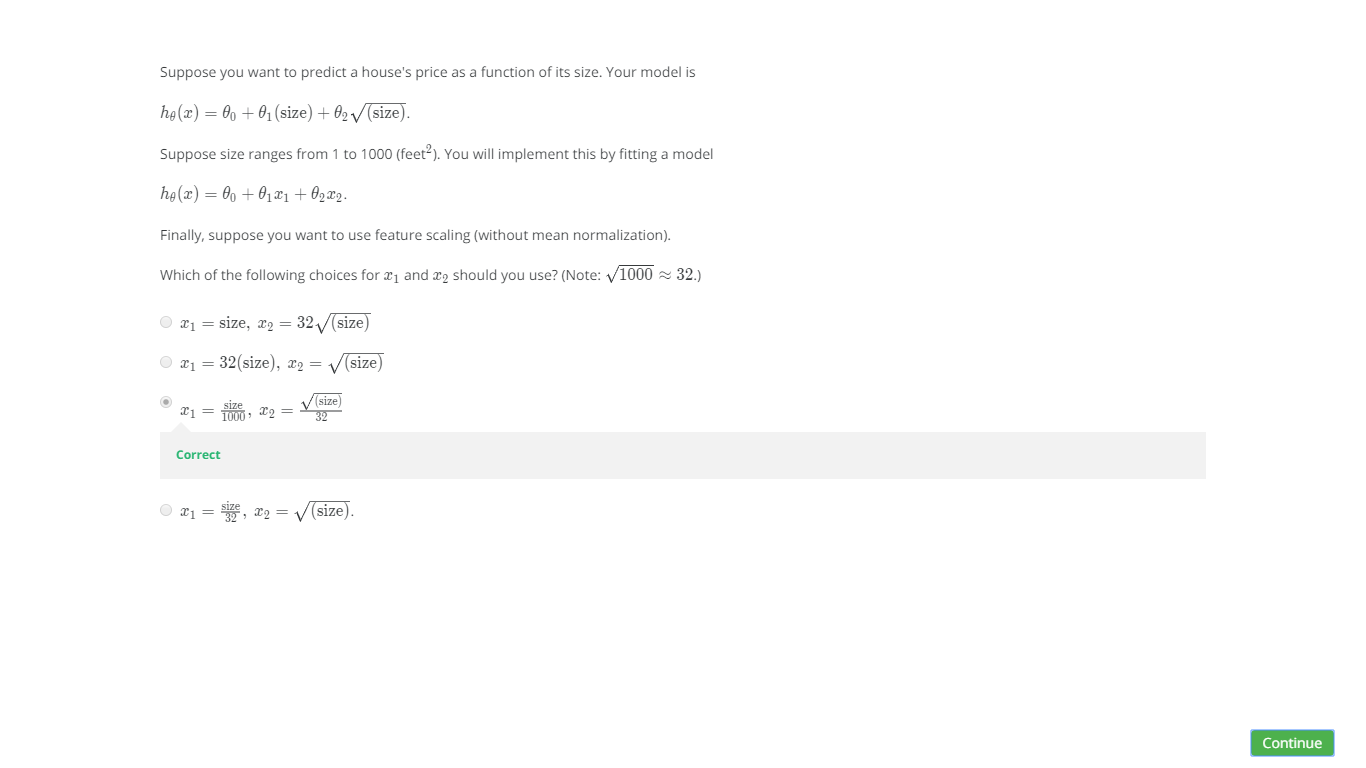
The age (in years) of the 4th home in the training set

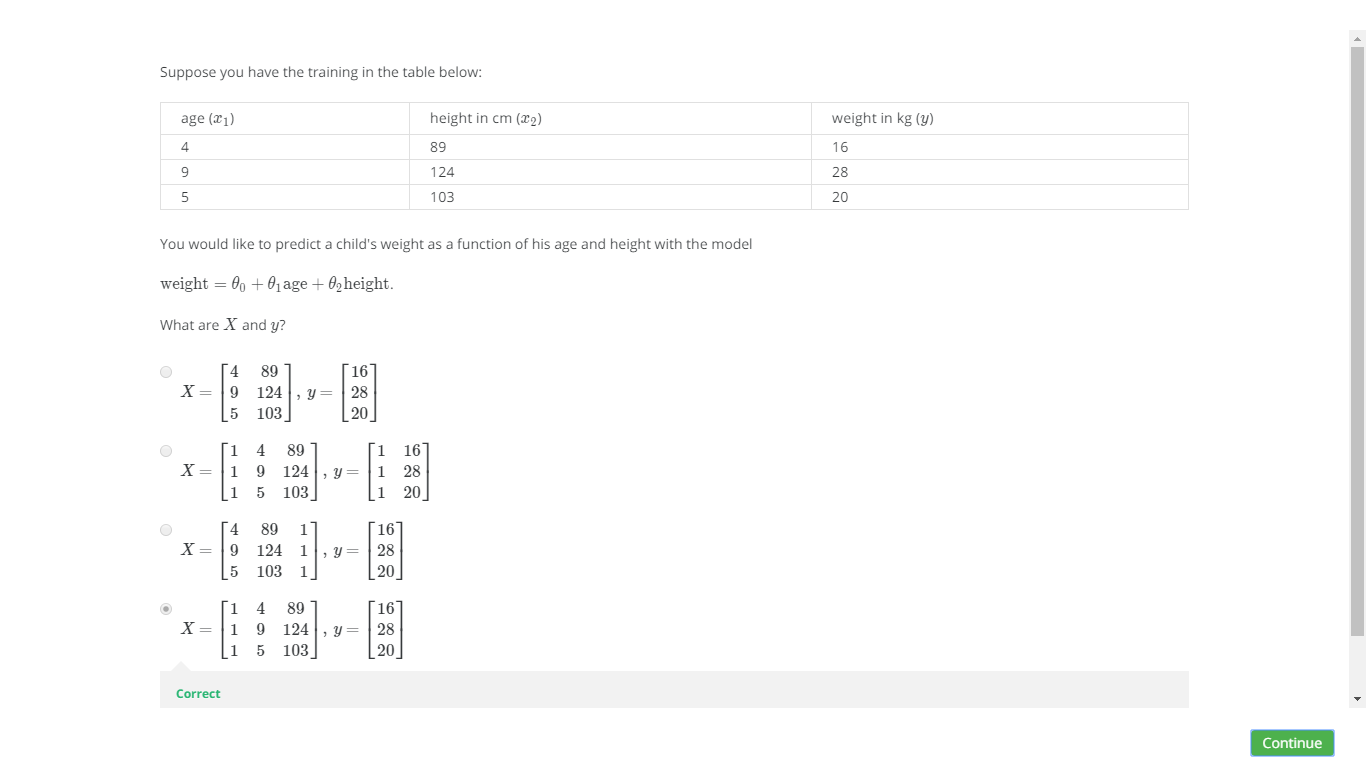


Suppose you are using a learning algorithm to estimate the price of houses in a city. You want one of your features x\_i*xi*​ to capture the age of the house. In your training set, all of your houses have an age between 30 and 50 years, with an average age of 38 years. Which of the following would you use as features, assuming you use feature scaling and mean normalization?









## 1. Question 1

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 exam | (midterm exam)^22 | 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*hθ*​(*x*)=*θ*0​+*θ*1​*x*1​+*θ*2​*x*2​, where x\_1*x*1​ is the midterm score and x\_2*x*2​ is (midterm score)^22. 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\_2^{(2)}*x*2(2)​? (Hint: midterm = 72, final = 74 is training example 2.) Please round off your answer to two decimal places and enter in the text box below.



Question 2

1

point

## 2. Question 2

You run gradient descent for 15 iterations

with \alpha = 0.3*α*=0.3 and compute J(\theta)*J*(*θ*) after each

iteration. You find that the value of J(\theta)*J*(*θ*) **increases** over

time. Based on this, which of the following conclusions seems

most plausible?



\alpha = 0.3*α*=0.3 is an effective choice of learning rate.



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



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

Question 3

1

point

## 3. Question 3

Suppose you have m = 23*m*=23 training examples with n = 5*n*=5 features (excluding the additional all-ones feature for the intercept term, which you should add). The normal equation is \theta = (X^TX)^{-1}X^Ty*θ*=(*XTX*)−1*XTy*. For the given values of m*m* and n*n*, what are the dimensions of \theta*θ*, X*X*, and y*y* in this equation?



X*X* is 23\times623×6, y*y* is 23\times623×6, \theta*θ* is 6\times66×6



X*X* is 23\times623×6, y*y* is 23\times123×1, \theta*θ* is 6\times16×1



X*X* is 23\times523×5, y*y* is 23\times123×1, \theta*θ* is 5\times55×5



X*X* is 23\times523×5, y*y* is 23\times123×1, \theta*θ* is 5\times15×1

Question 4

1

point

## 4. Question 4

Suppose you have a dataset with m = 50*m*=50 examples and n = 200000*n*=200000 features for each example. You want to use multivariate linear regression to fit the parameters \theta*θ* to our data. Should you prefer gradient descent or the normal equation?



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



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



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



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

Question 5

1

point

## 5. Question 5

Which of the following are reasons for using feature scaling?



It prevents the matrix X^TX*XTX* (used in the normal equation) from being non-invertable (singular/degenerate).



It speeds up gradient descent by making it require fewer iterations to get to a good solution.

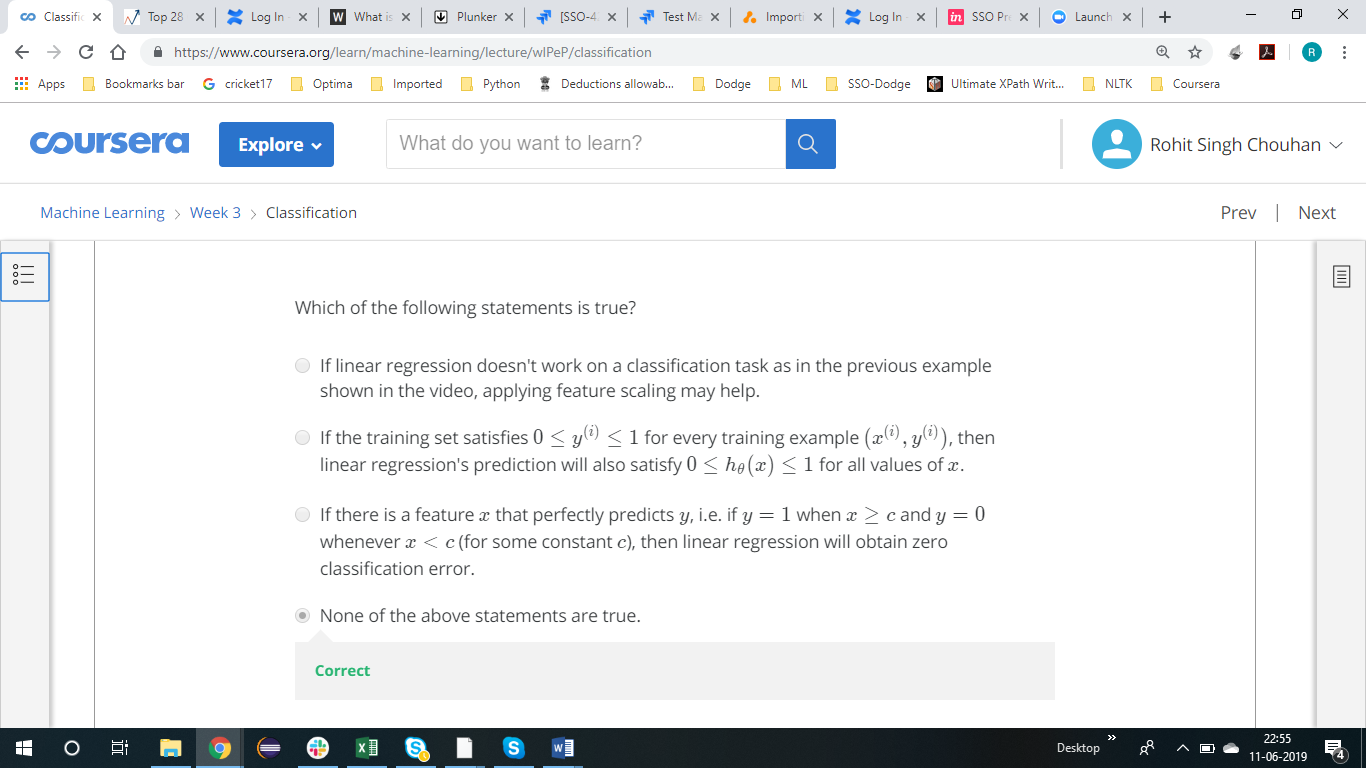


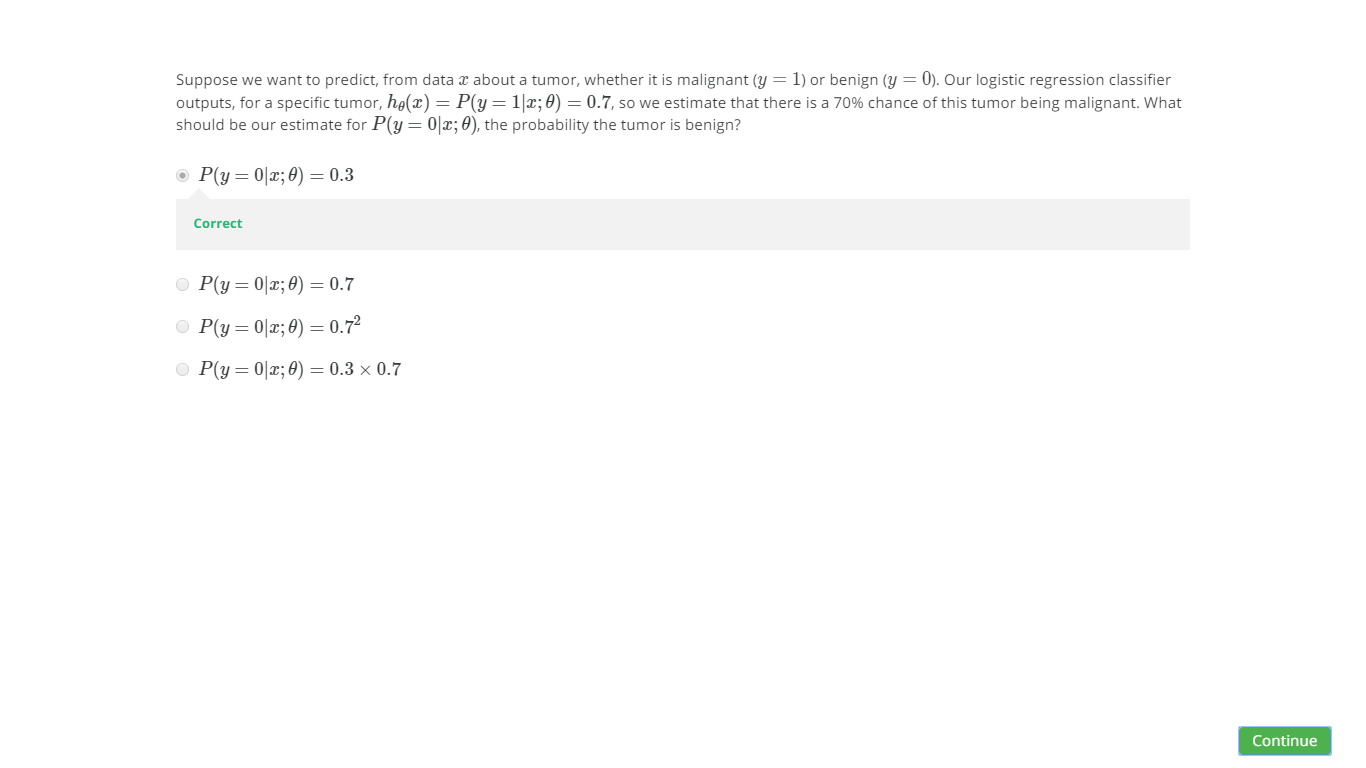
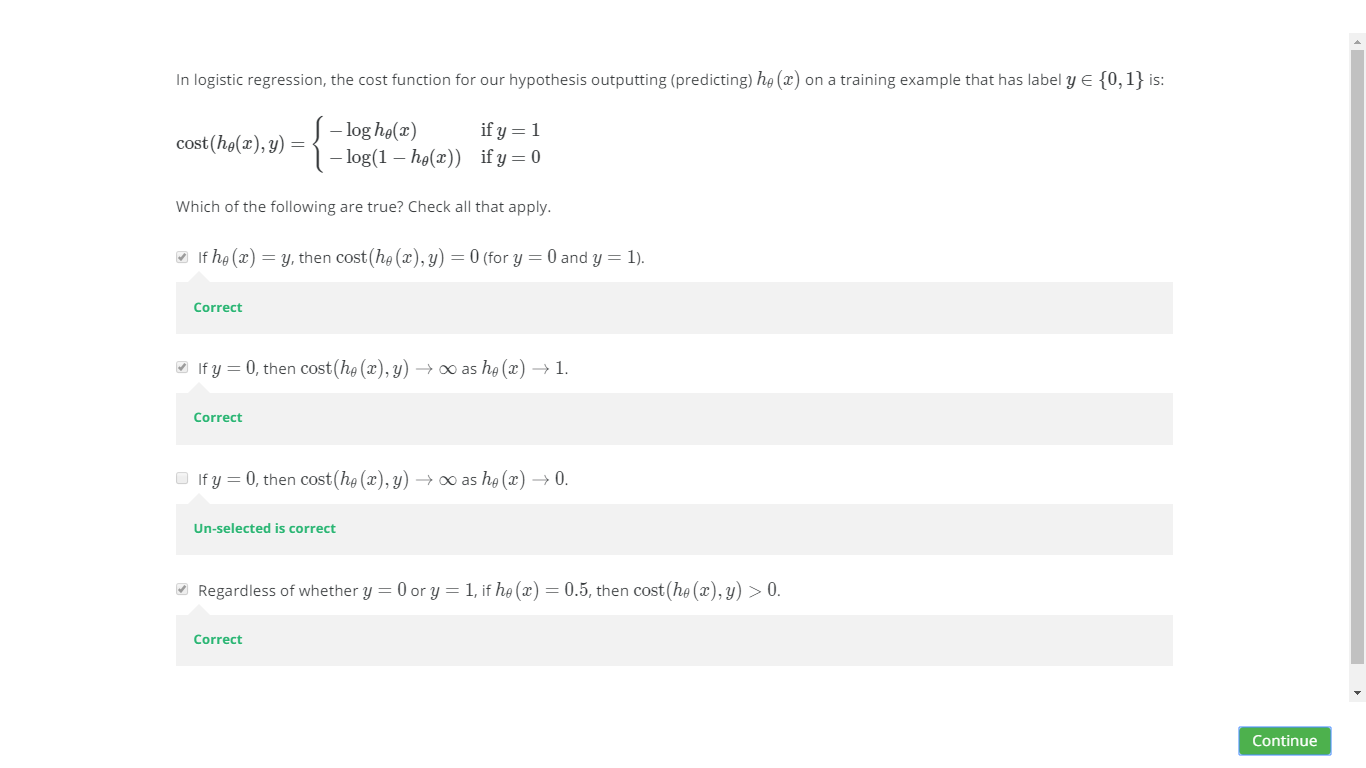
It speeds up solving for \theta*θ* using the normal equation.

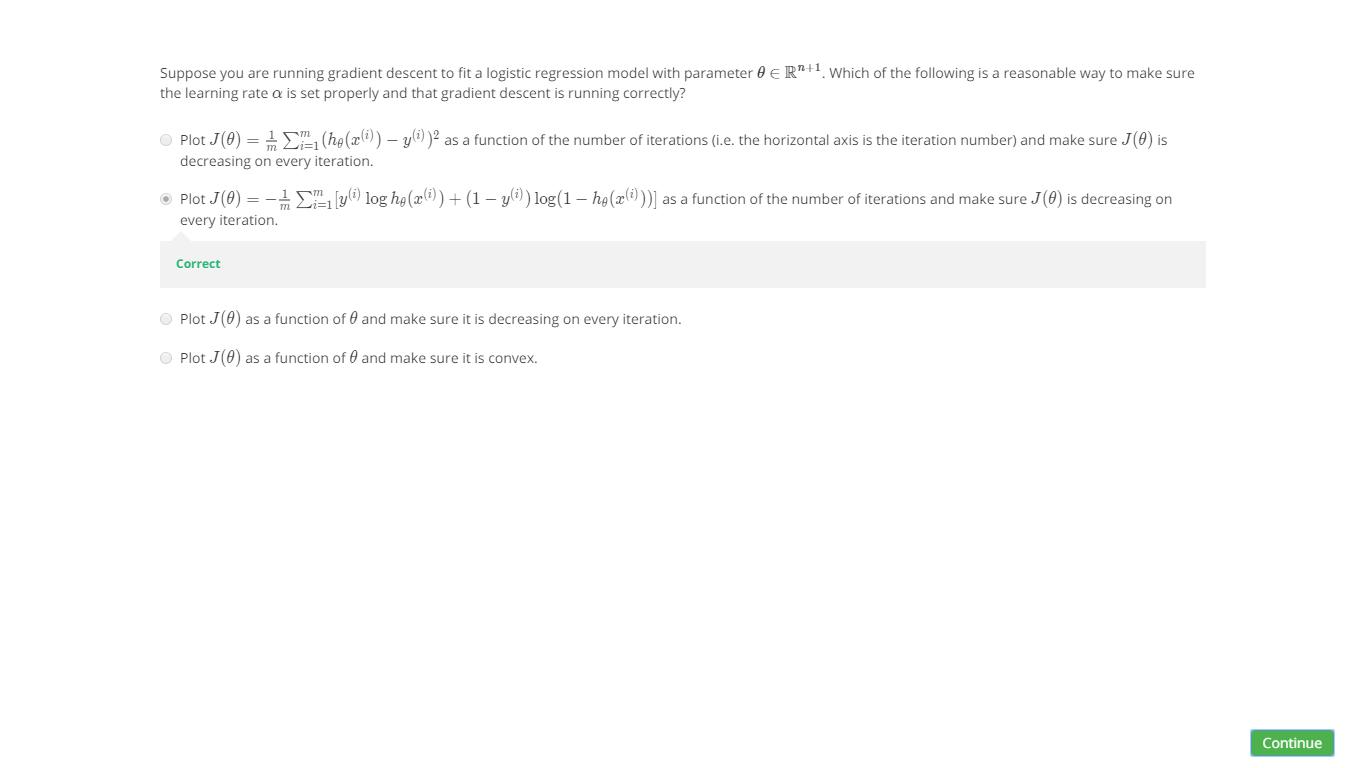


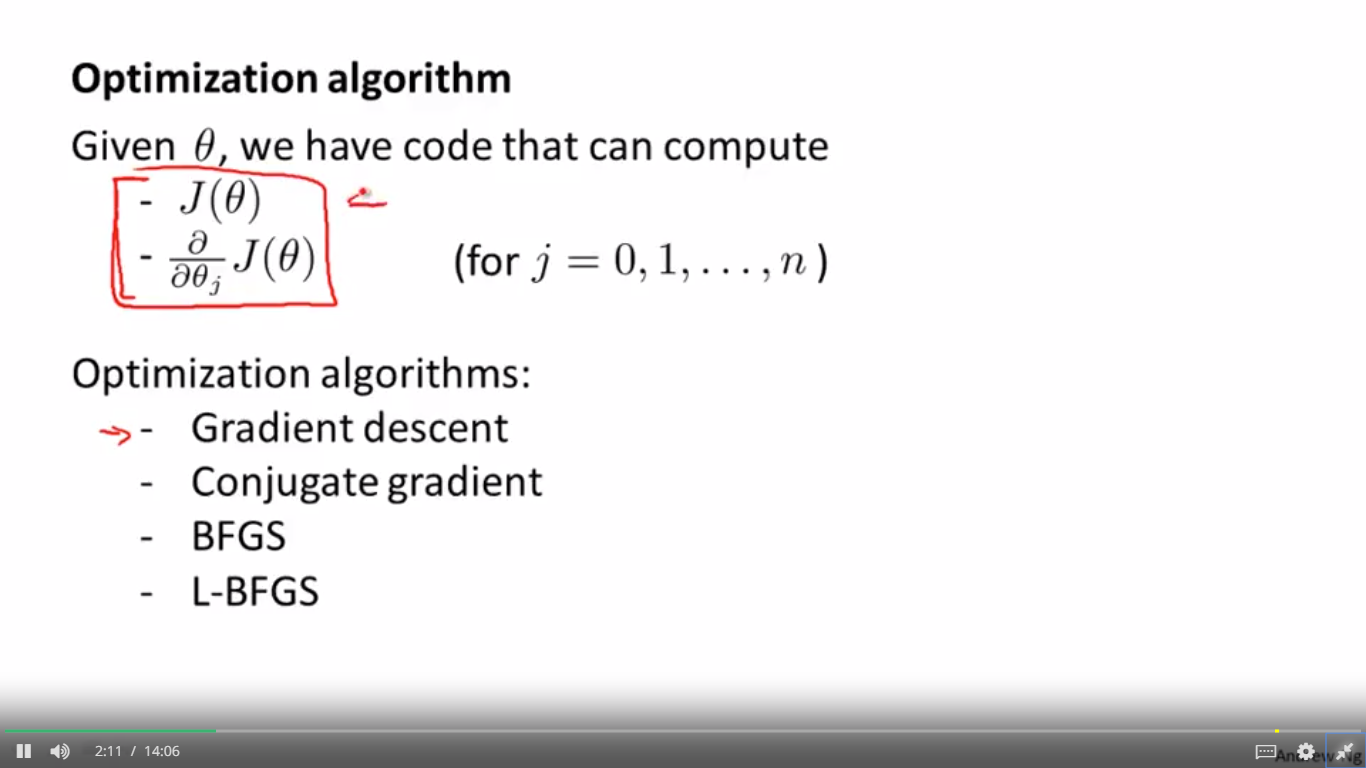
It is necessary to prevent gradient descent from getting stuck in local optima.

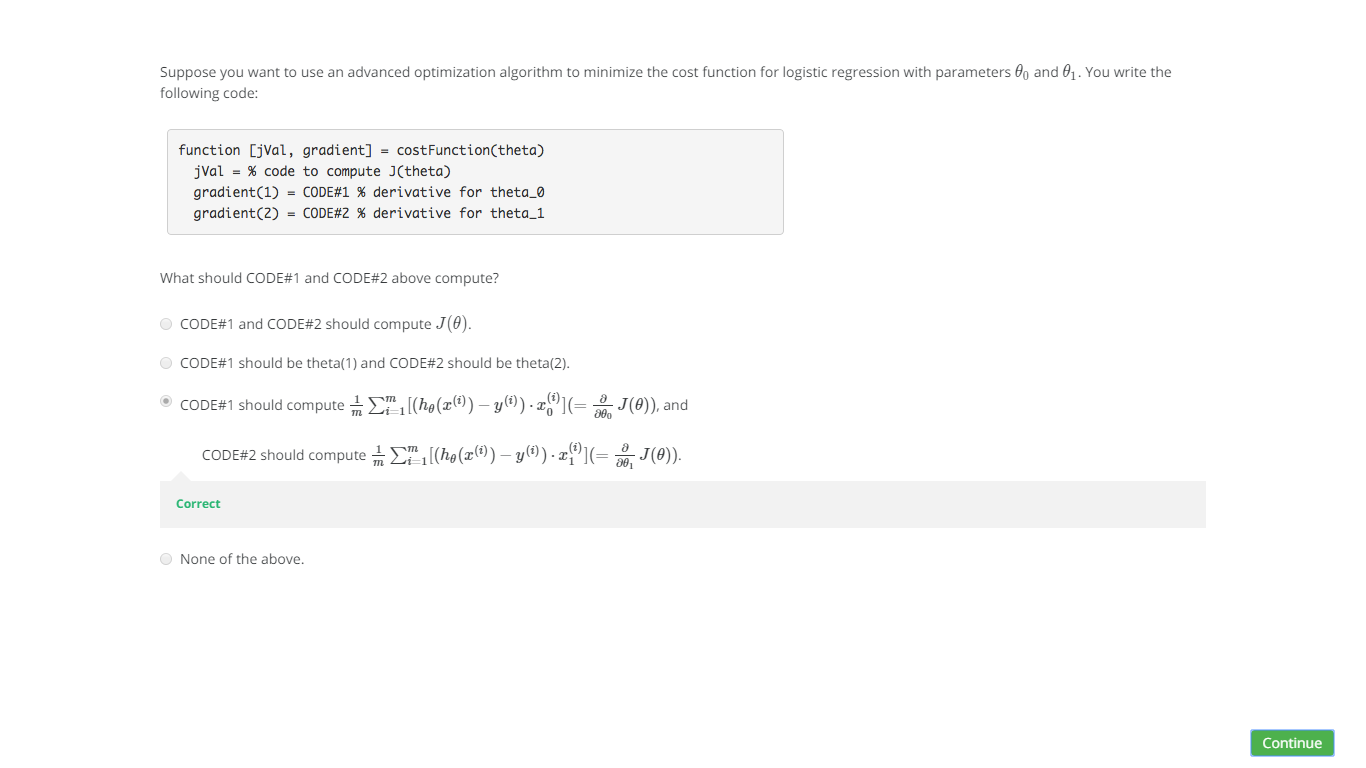
LOGISTIC REGRESSION



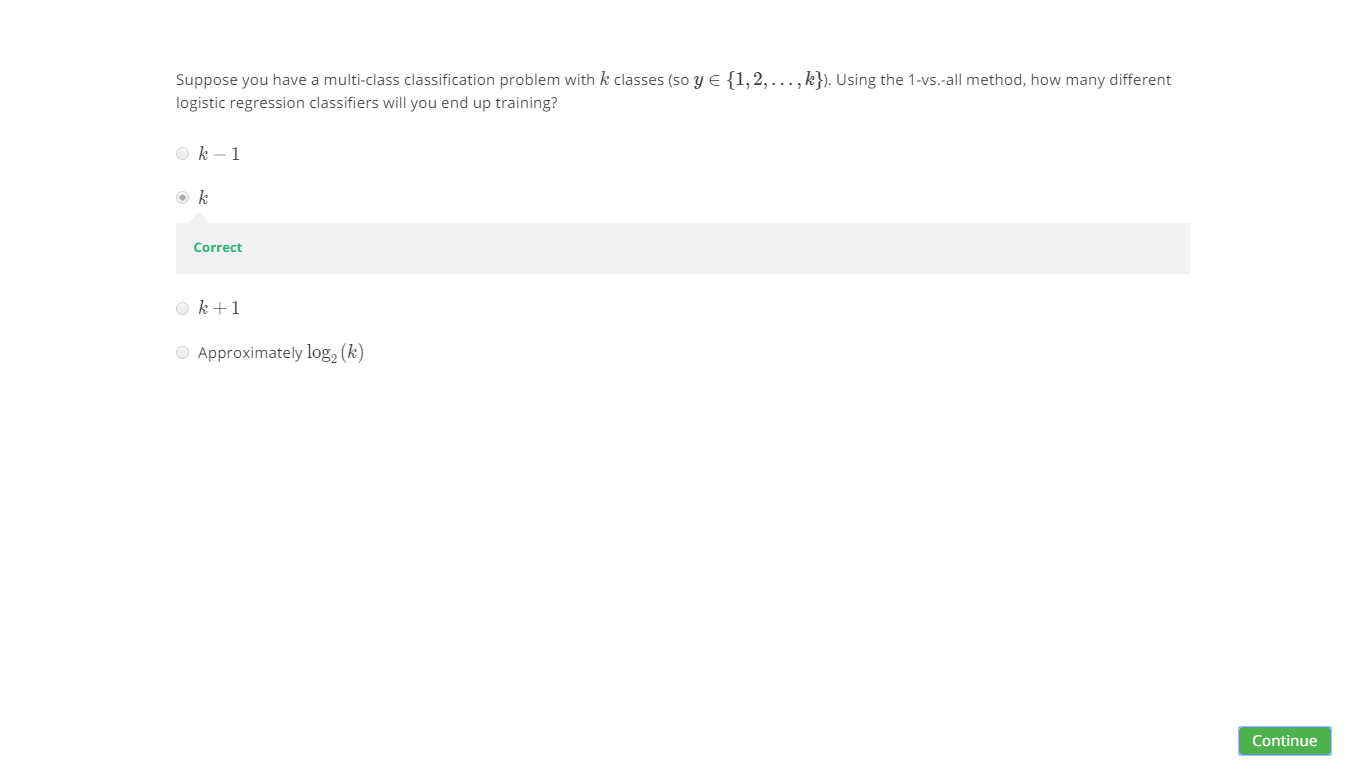
 

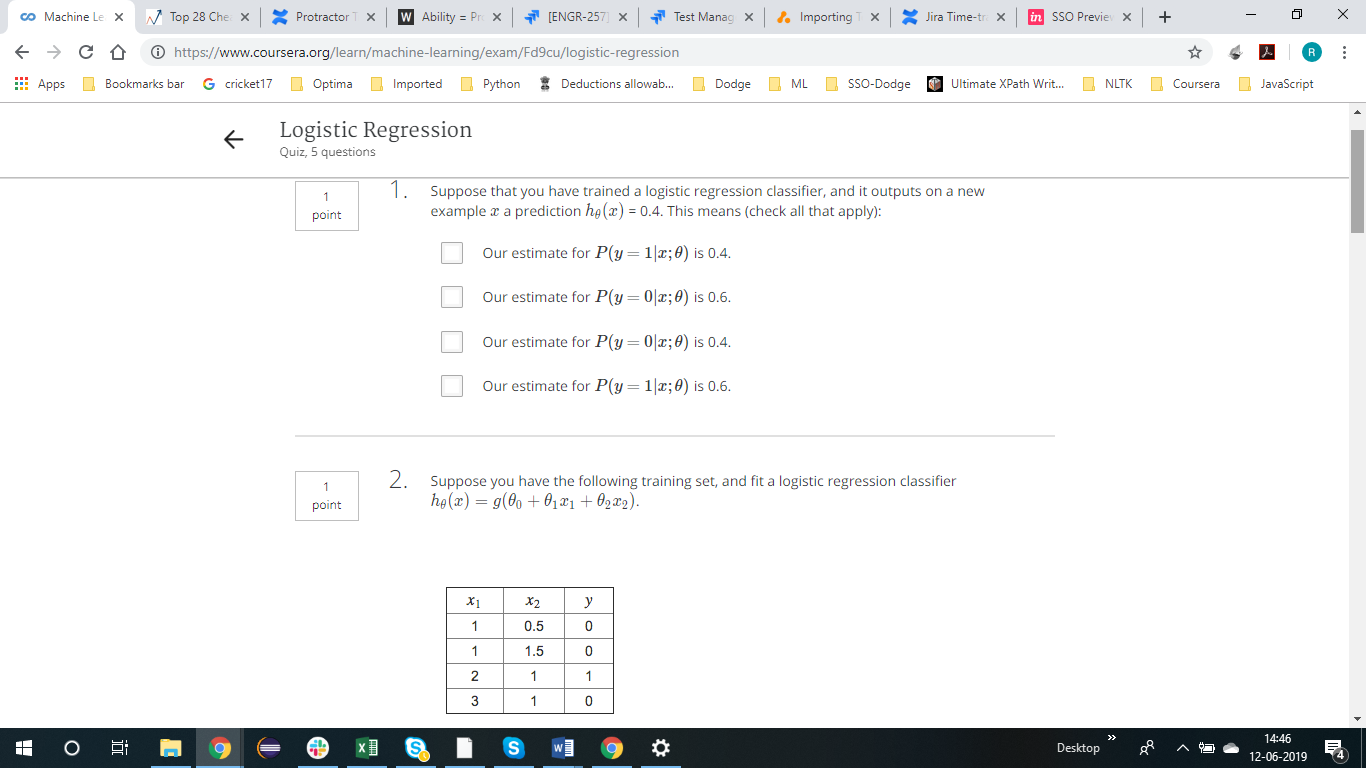
 

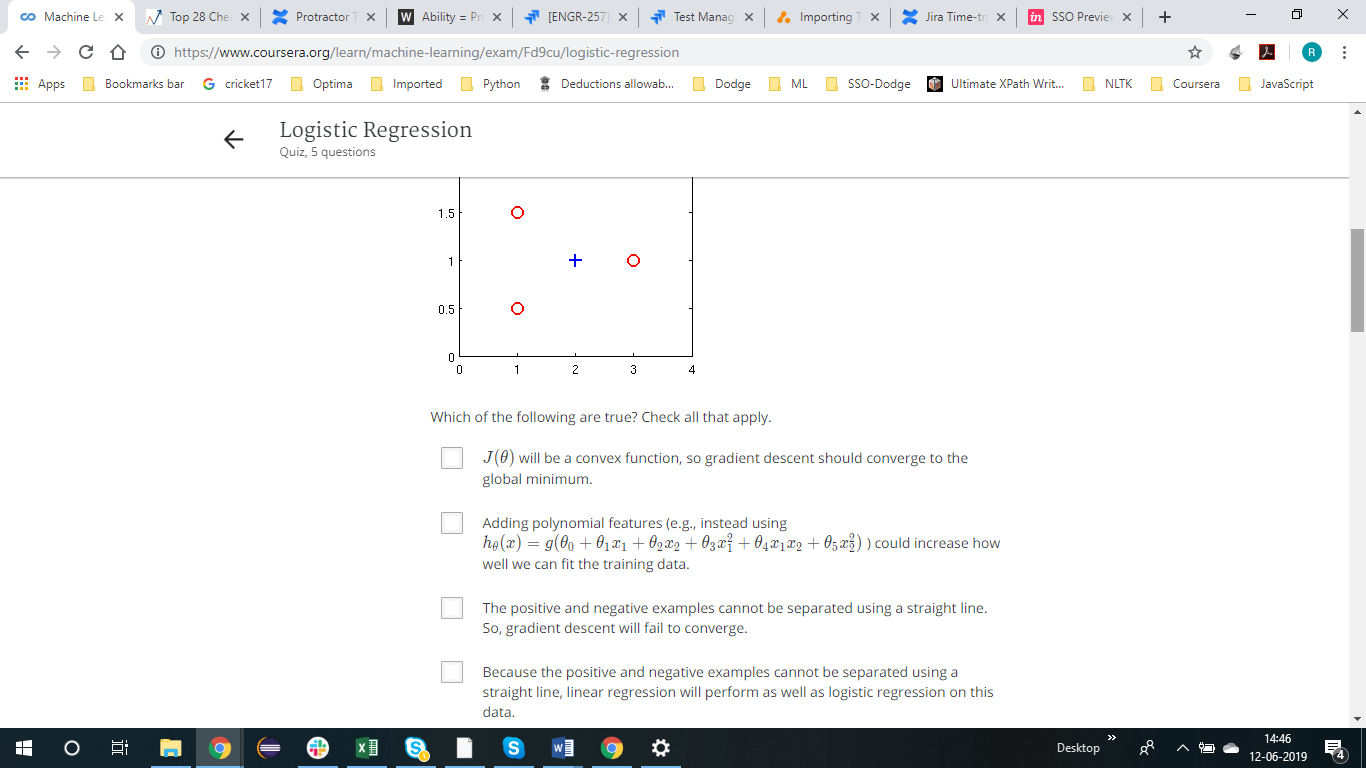


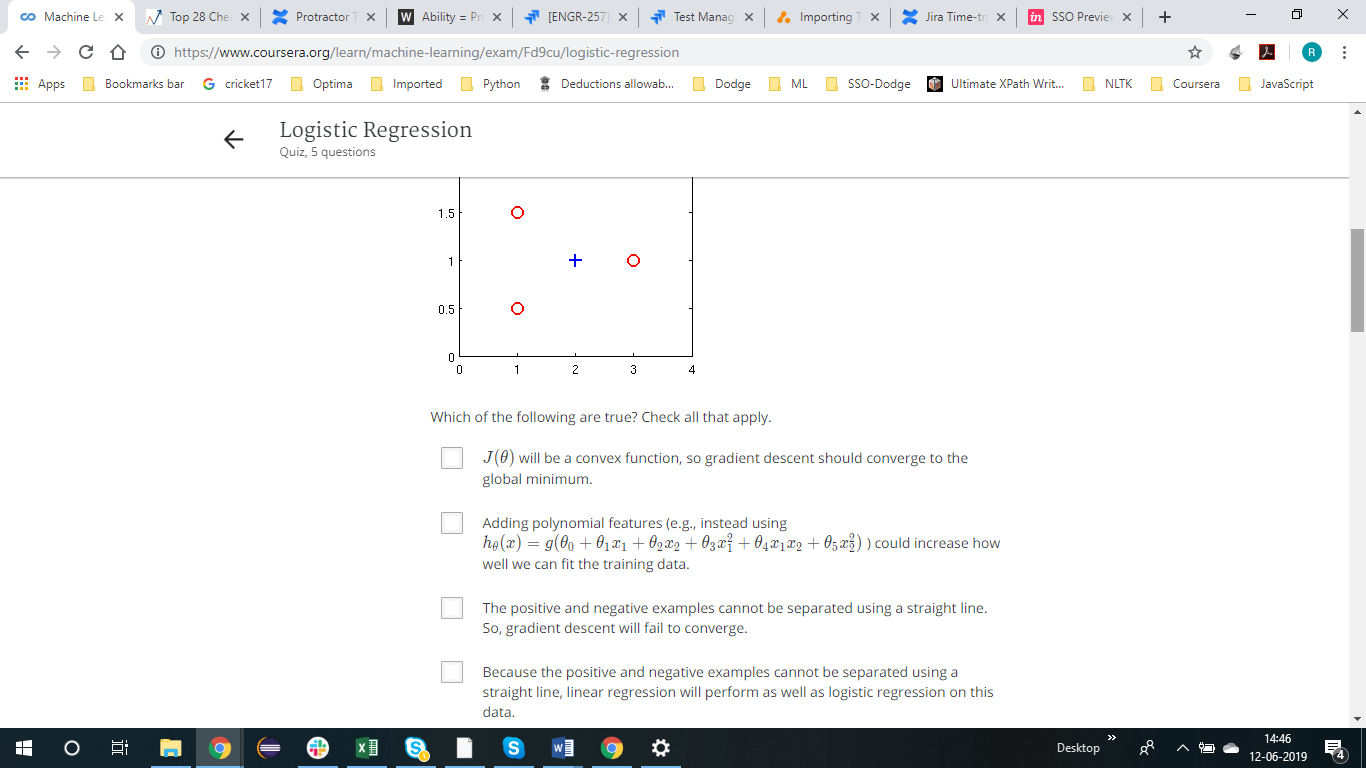


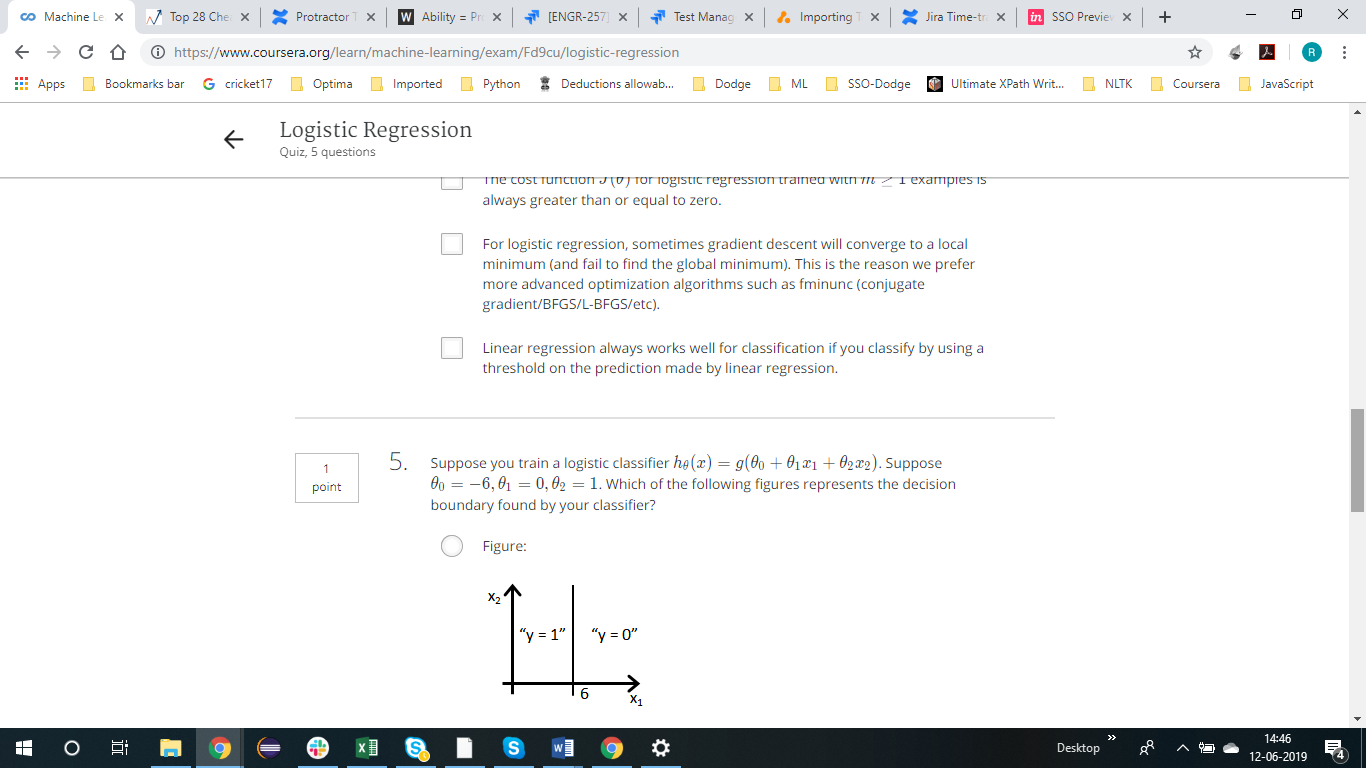
MULTI CLASS CLASSIFICATION

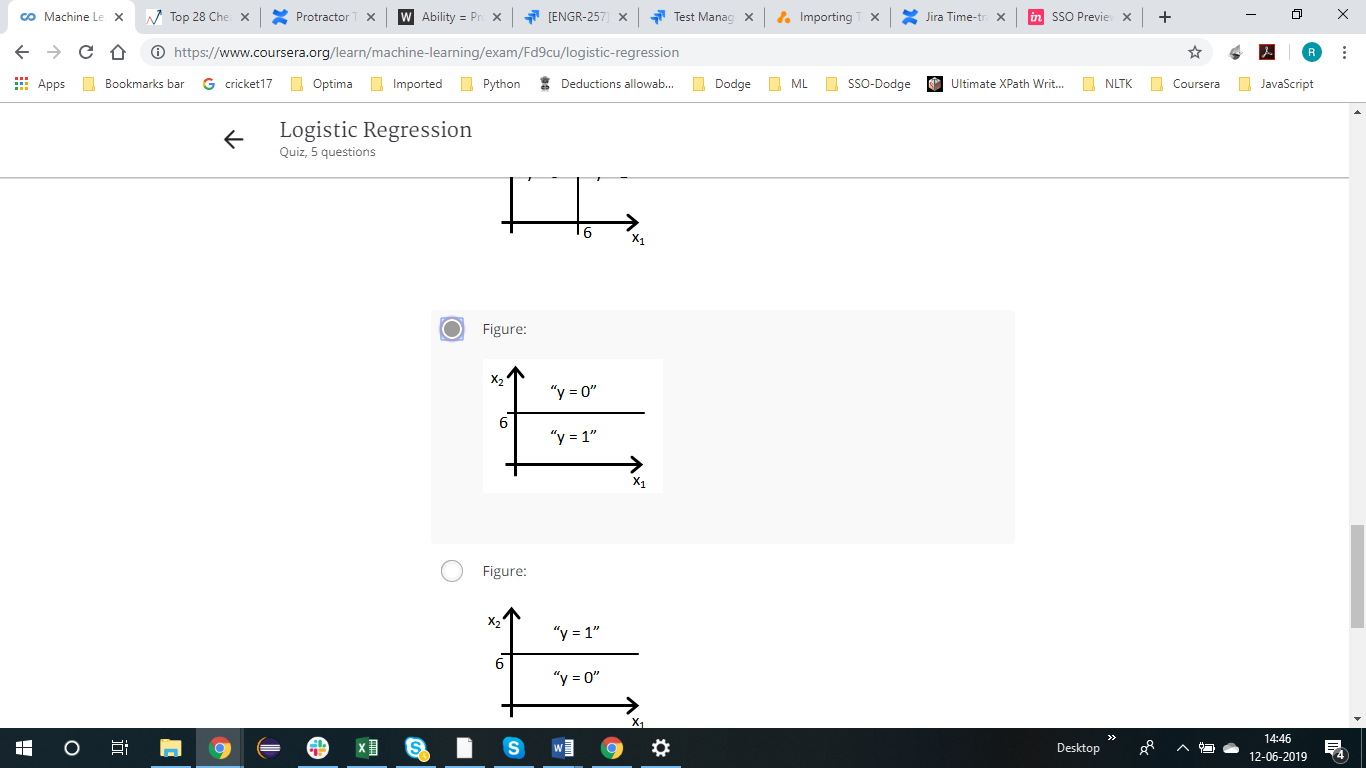


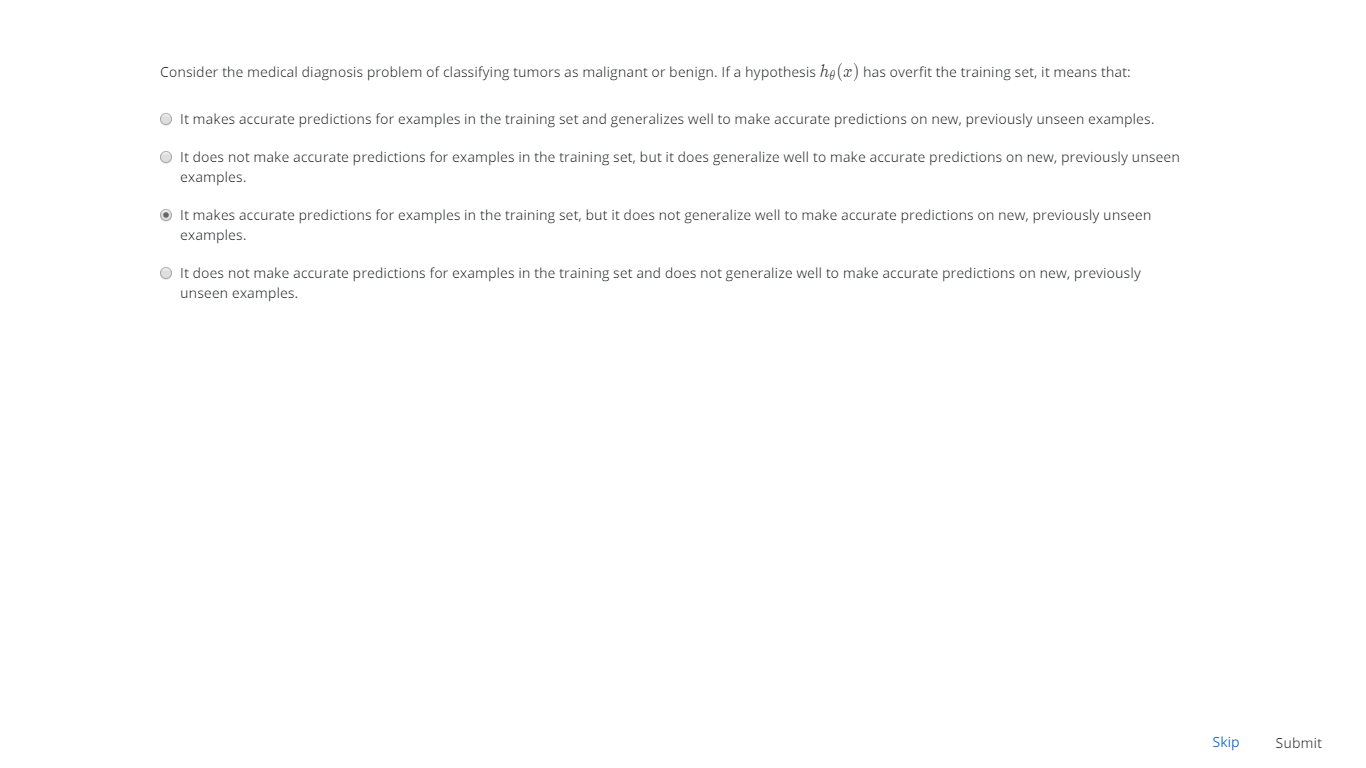


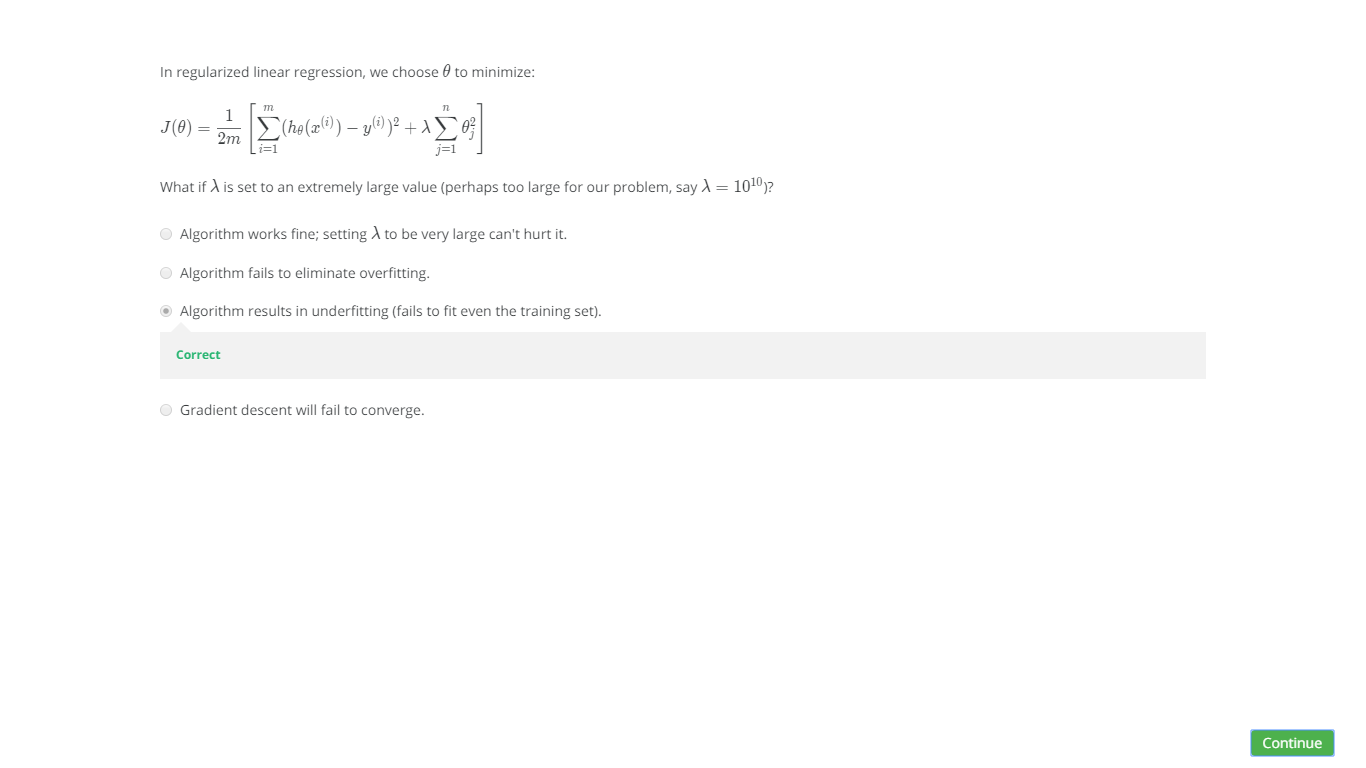


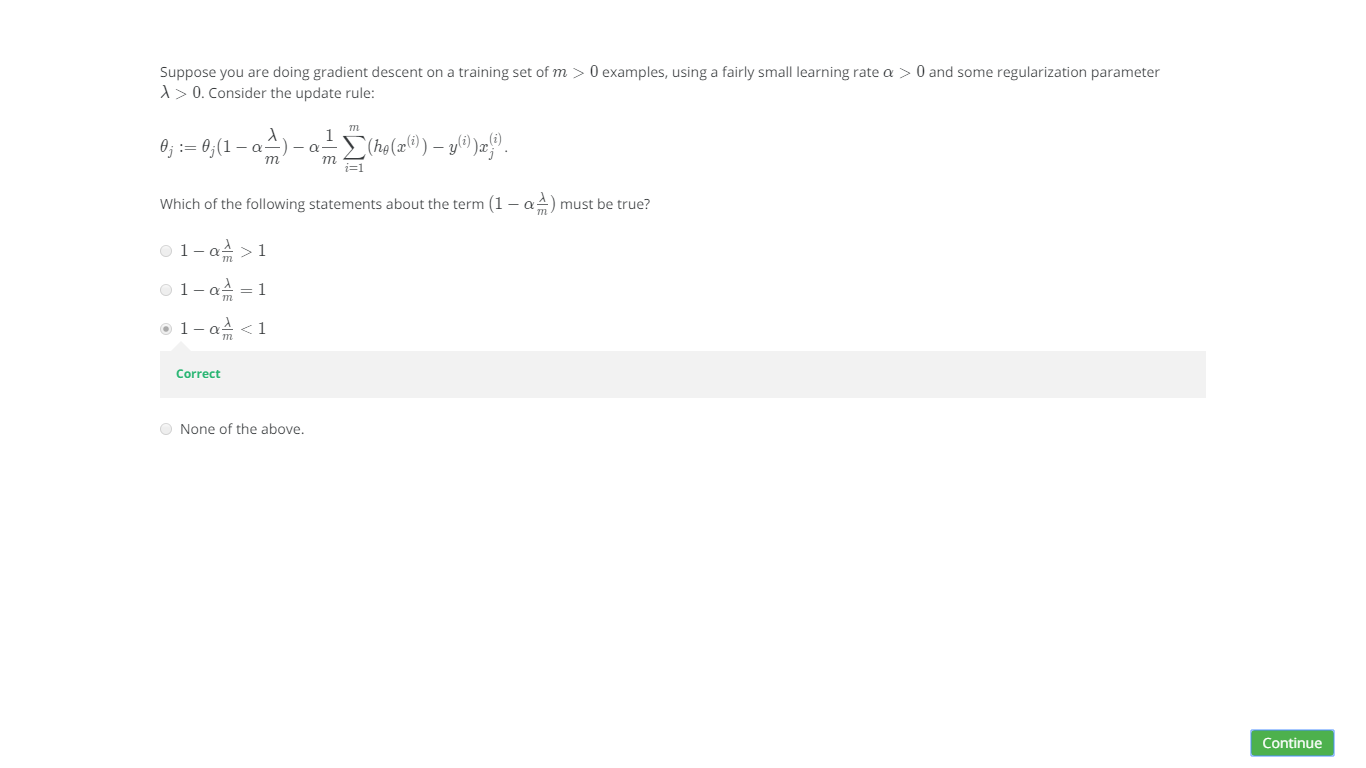


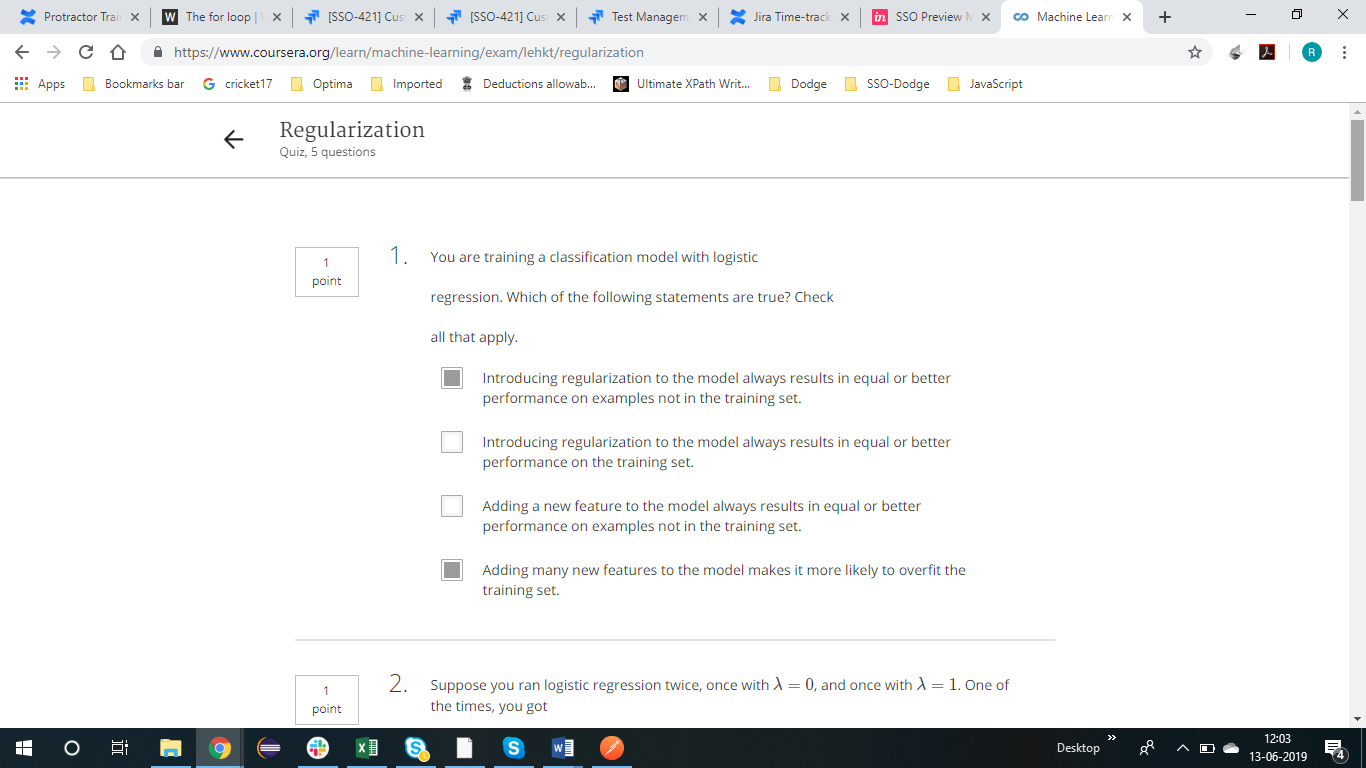


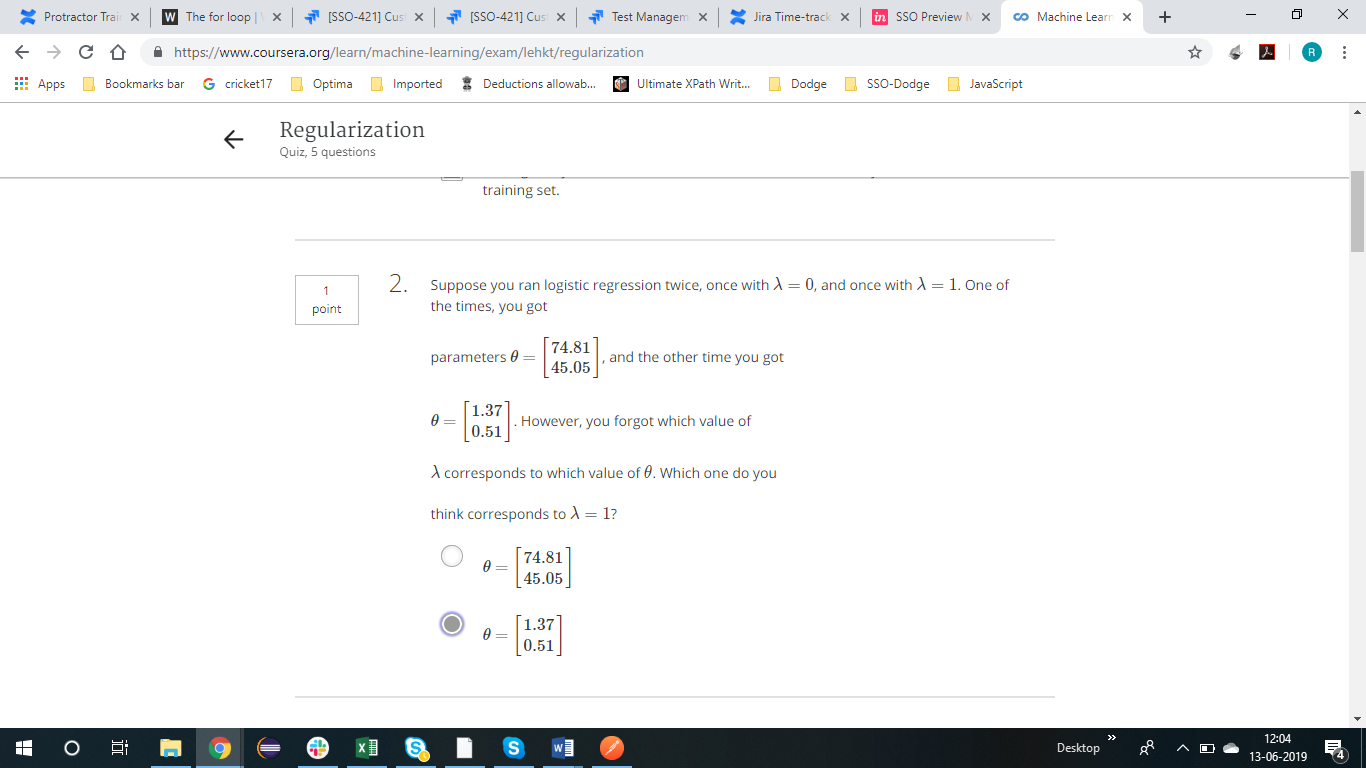


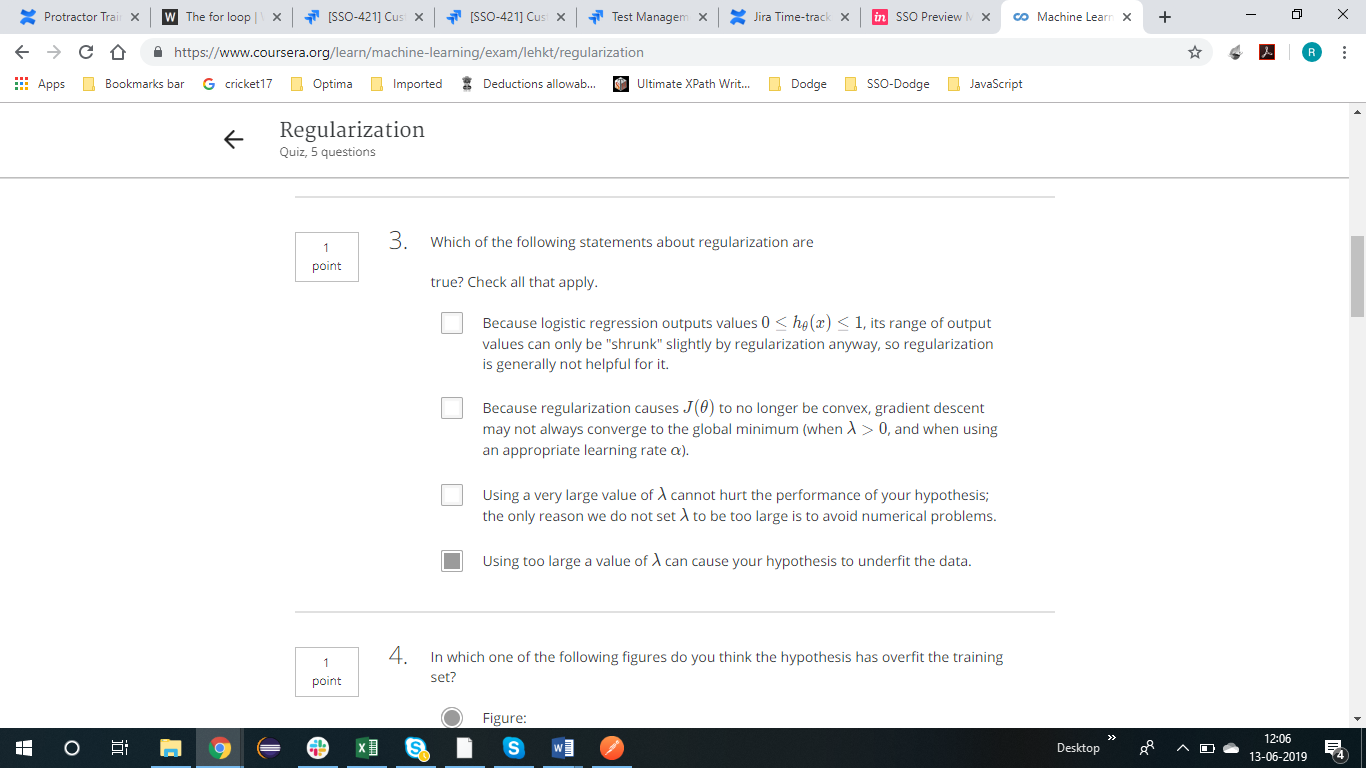










NEURAL NETWORK

