Feedback — I. Introduction

Help Center

You submitted this quiz on **Fri 23 Jan 2015 2:14 PM CET**. You got a score of **5.00** out of **5.00**.

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 E?

Your Answer		Score	Explanation
The probability of it correctly predicting a future date's weather.			
The weather prediction task.			
None of these.			
The process of the algorithm examining a large amount of historical weather data.	✔	1.00	It is by examining the historical weather data that the learning algorithm improves it's performance, so this is the experience E.
Total		1.00 / 1.00	

Question 2

The amount of rain that falls in a day is usually measured in either millimeters (mm) or inches. Suppose you use a learning algorithm to predict how much rain will fall tomorrow. Would you treat this as a classification or a regression problem?

Your Answer		Score	Explanation
	✔	1.00	Regression is appropriate when we are trying to predict a

Regression		continuous-valued output, such as the amount of rainfall measured in inches or mm.
Classification		
Total	1.00 / 1.00	

Question 3

Suppose you are working on stock market prediction. You would like to predict whether the US Dollar will go up against the Euro tomorrow (i.e., whether a dollar will be worth more euros tomorrow than it is worth today). Would you treat this as a classification or a regression problem?

Your Answer		Score	Explanation
Regression			
Classification	✔	1.00	Classification is appropriate when we are trying to predict one of a small number of discrete-valued outputs. Here, there are two possible outcomes: That the US Dollar goes up (which we might designate as class 0, say) or that it does not (class 1).
Total		1.00 / 1.00	

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.

Your Answer		Score	Explanation
☑ In farming, given data on crop yields over the last 50 years, learn to predict next year's crop	✔	0.25	This can be addresses as a supervised learning problem, where we learn from historical data (labeled with historical crop yields) to predict future crop yields.

yields.			
☑ Given historical data of childrens' ages and heights, predict children's height as a function of their age.	✔	0.25	This is a supervised learning, regression problem, where we can learn from a training set to predict height.
☐ Examine a large collection of emails that are known to be spam email, to discover if there are sub-types of spam mail.	✔	0.25	This can addressed using a clustering (unsupervised learning) algorithm, to cluster spam mail into sub-types.
☐ 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.	✔	0.25	This can be addressed using an unsupervised learning, clustering, algorithm, in which we group the 1000 patients into different clusters based on their responses to the drug.
Total		1.00 / 1.00	

Question 5

Which of these is a reasonable definition of machine learning?

Your Answer		Score	Explanation
○ Machine learning means from labeled data.			
○ 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	✔	1.00	This was the definition given by Arthur

study that gives computers the ability to learn without being explicitly programmed.		Samuel (who had written the famous checkers playing, learning program).
Total	1.00 / 1.00	

Feedback — II. Linear regression with one variable

Help Center

You submitted this quiz on **Sat 24 Jan 2015 8:46 AM CET**. You got a score of **5.00** out of **5.00**.

Question 1

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

Questions 1 through 4 will use the following training set of a small sample of different students' performances. Here each row is one training example. Recall that in linear regression, our hypothesis is $h_\theta(x) = \theta_0 + \theta_1 x$, and we use m to denote the number of training examples.

x	y
3	2
1	2
0	1
4	3

For the training set given above, what is the value of m? In the box below, please enter your answer (which should be a number between 0 and 10).

You entered:

4

Your Answer		Score	Explanation
4	✔	1.00	
Total		1.00 / 1.00	

Question Explanation

m is the number of training examples. In this example, we have m=4 examples.

Question 2

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

x	y
1	0.5
2	1
4	2
0	0

Consider the linear regression model $h_\theta(x) = \theta_0 + \theta_1 x$. What are the values of θ_0 and 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.)

Your Answer		Score	Explanation
$\theta_0 = 0.5, \theta_1 = 0$			
$\theta_0 = 0.5, \theta_1 = 0.5$			
$\theta_0 = 0, \theta_1 = 0.5$	✔	1.00	
$\theta_0 = 1, \theta_1 = 1$			
Total		1.00 / 1.00	

Question Explanation

As $J(\theta_0, \theta_1) = 0, y = h_\theta(x) = \theta_0 + \theta_1 x$. Using any two values in the table, solve for θ_0, θ_1.

Question 3

Consider the training set below with only $m = 3$ training examples:

x	y
1	1
2	2
3	3

Recall the gradient descent algorithm used to update θ_0 and θ_1:

$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})$

$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x^{(i)}$

Now let's assume we choose θ_0 = 0 and θ_1 = 0.5 as our starting point, and we choose α to be 0.1. After you perform one iteration of gradient descent, what will be the new values for θ_0 and θ_1?

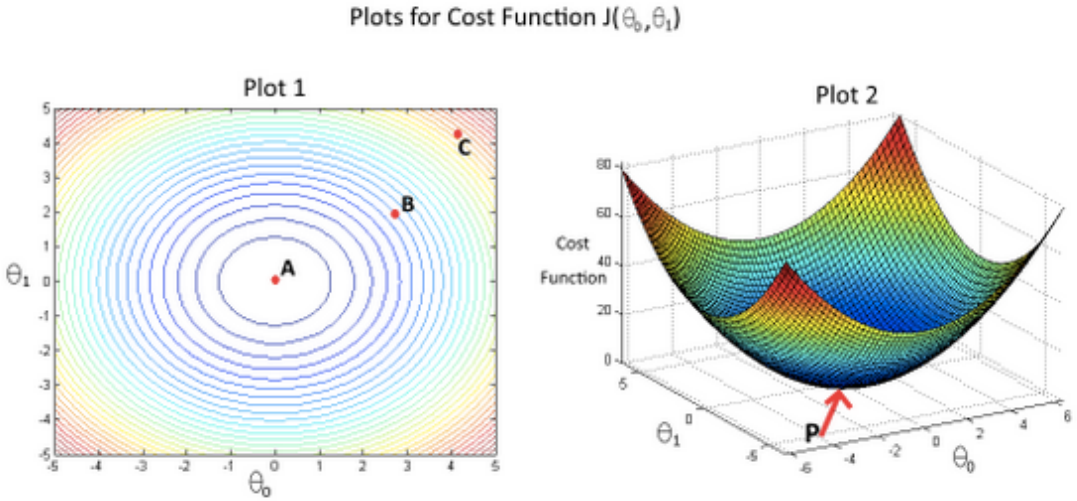
Your Answer		Score	Explanation
○ θ_0 = 0 and θ_1 = 1			
○ θ_0 = 0.1 and θ_1 = 0.713			
○ θ_0 = 0.15 and θ_1 = 0.632			
◉ θ_0 = 0.1 and θ_1 = 0.733	✔	1.00	
Total		1.00 / 1.00	

Question Explanation

Remember to update both θ_0 and θ_1 simultaneously. So,
$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})$ = 0 - 0.1 x (1/3) x ((0.5 x 1 - 1) + (0.5 x 2 - 2) + (0.5 x 3 - 3)) = 0.1 $\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x^{(i)}$ = 0.5 - 0.1 x (1/3) x ((0.5 x 1 - 1) x 1 + (0.5 x 2 - 2) x 2 + (0.5 x 3 - 3)) x 3) = 0.733

Question 4

In the given figure, the cost function $J(\theta_0, \theta_1)$ 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).



Your Answer		Score	Explanation
☑ Point P (the global minimum of plot 2) corresponds to point A of Plot 1.	✔	0.20	Correct. Plot 2 is a 3-D surface plot for cost function $J(\theta_0, \theta_1)$ against θ_0 and θ_1, whereas Plot 1 is the 2-D contour plot for the same cost function. Hence, the correspondence of the two plots can be understood.
☐ 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.	✔	0.20	
☐ Point P (The global minimum of plot 2) corresponds to point C of Plot 1.	✔	0.20	
☐ 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.	✔	0.20	
☑ 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.	✔	0.20	Correct. Correct implementation of Gradient Descent Algorithm will help us minimizing the cost function $J(\theta_0, \theta_1)$. Since point A represents the global minimum of the cost function, gradient descent should lead us to reach at or near point A.
Total		1.00 / 1.00	

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 θ_0, θ_1 such that $J(\theta_0, \theta_1) = 0$. Which of the statements below must then be true? (Check all that apply.)

Your Answer		Score	Explanation
☐ We can perfectly predict the value of y even for new examples that we have not yet seen. (e.g., we can perfectly predict prices of even new houses that we have not yet seen.)	✔	0.25	Even though we can fit our training set perfectly, this does not mean that we'll always make perfect predictions on houses in the future/on houses that we have not yet seen.
☑ Our training set can be fit perfectly by a straight line, i.e., all of our training examples lie perfectly on some straight line.	✔	0.25	If $J(\theta_0, \theta_1) = 0$, that means the line defined by the equation "$y = \theta_0 + \theta_1 x$" perfectly fits all of our data.
☐ For this to be true, we must have $y^{(i)} = 0$ for every value of $i = 1, 2, \ldots, m$.	✔	0.25	So long as all of our training examples lie on a straight line, we will be able to find θ_0 and θ_1 so that $J(\theta_0, \theta_1) = 0$. It is not necessary that $y^{(i)} = 0$ for all of our examples.
☐ For this to be true, we must have $\theta_0 = 0$ and $\theta_1 = 0$ so that $h_\theta(x) = 0$	✔	0.25	If $J(\theta_0, \theta_1) = 0$, that means the line defined by the equation "$y = \theta_0 + \theta_1 x$" perfectly fits all of our data. There's no particular reason to expect that the values of θ_0 and θ_1 that achieve this are both 0 (unless $y^{(i)} = 0$ for all of our training examples).
Total		1.00 / 1.00	

Feedback — III. Linear Algebra

Help Center

You submitted this quiz on **Sat 24 Jan 2015 8:27 AM CET**. You got a score of **5.00** out of **5.00**.

Question 1

Let two matrices be

$$A = \begin{bmatrix} 1 & -4 \\ -2 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 3 \\ 5 & 8 \end{bmatrix}$$

What is A + B?

Your Answer		Score	Explanation
$\begin{bmatrix} 1 & -1 \\ 7 & 9 \end{bmatrix}$			
$\begin{bmatrix} 1 & -7 \\ -7 & -7 \end{bmatrix}$			
$\begin{bmatrix} 1 & 7 \\ 7 & 9 \end{bmatrix}$			
$\begin{bmatrix} 1 & -1 \\ 3 & 9 \end{bmatrix}$	✔	1.00	To add two matrices, add them element-wise.
Total		1.00 / 1.00	

Question 2

Let $x = \begin{bmatrix} 2 \\ 7 \\ 4 \\ 1 \end{bmatrix}$

What is $3 * x$?

Your Answer	Score	Explanation

$\begin{bmatrix} 6 \\ 21 \\ 12 \\ 3 \end{bmatrix}$	✔	1.00	To multiply the vector x by 3, take each element of x and multiply that element by 3.
$\begin{bmatrix} \frac{2}{3} \\ \frac{7}{3} \\ \frac{4}{3} \\ \frac{1}{3} \end{bmatrix}$			
$\begin{bmatrix} \frac{2}{3} & \frac{7}{3} & \frac{4}{3} & \frac{1}{3} \end{bmatrix}$			
$\begin{bmatrix} 6 & 21 & 12 & 3 \end{bmatrix}$			
Total		1.00 / 1.00	

Question 3

Let u be a 3-dimensional vector, where specifically

$$u = \begin{bmatrix} 8 \\ 1 \\ 4 \end{bmatrix}$$

What is u^{T}?

Your Answer		Score	Explanation
$\begin{bmatrix} 4 & 1 & 8 \end{bmatrix}$			
$\begin{bmatrix} 8 & 1 & 4 \end{bmatrix}$	✔	1.00	
$\begin{bmatrix} 8 \\ 1 \\ 4 \end{bmatrix}$			
$\begin{bmatrix} 4 \\ 1 \\ 8 \end{bmatrix}$			
Total		1.00 / 1.00	

Question 4

Let u and v be 3-dimensional vectors, where specifically

$$u = \begin{bmatrix} 4 \\ -4 \\ -3 \end{bmatrix} \text{ and } v = \begin{bmatrix} 4 \\ 2 \\ 4 \end{bmatrix}$$

What is $u^T v$?

(Hint: u^T is a 1x3 dimensional matrix, and v can also be seen as a 3x1 matrix. The answer you want can be obtained by taking the matrix product of u^T and v.)

You entered:

-4

Your Answer		Score	Explanation
-4	✔	1.00	
Total		1.00 / 1.00	

Question 5

Let A and B be 3x3 (square) matrices. Which of the following must necessarily hold true?

Your Answer		Score	Explanation
☐ If $C = A * B$, then C is a 6x6 matrix.	✔	0.25	Since A and B are both 3x3 matrices, their product is 3x3. More generally, if A were an $m \times n$. matrix, and B a $n \times o$ matrix, then C would be $m \times o$. (In our example, $m = n = o = 3$.)
☑ If B is the 3x3 identity matrix, then $A * B = B * A$	✔	0.25	Even though matrix multiplication is not commutative in general ($A * B \neq B * A$ for general matrices A, B), for the special case where $B = I$, we have $A * B = A * I = A$, and also $B * A = I * A = A$. So, $A * B = B * A$.
☐ $A * B = B * A$	✔	0.25	We saw in the lecture that matrix multiplication is not commutative in general.
☑ If v is a 3	✔	0.25	Since A and B are both 3x3 matrices, $A * B$ is 3x3 matrix.

dimensional vector, then $A * B * v$ is a 3 dimensional vector.		Thus, $(A * B) * v$ is a 3x3 matrix times a 3×1 matrix (since v is a 3 dimensional vector, and thus also a 3x1 matrix), and the result gives a 3x1 vector.
Total	1.00 / 1.00	

Feedback — IV. Linear Regression with Multiple Variables

Help Center

You submitted this quiz on **Mon 26 Jan 2015 10:00 AM CET**. You got a score of **5.00** out of **5.00**.

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)^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. 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 $x_2^{(4)}$? (Hint: midterm = 89, final = 96 is training example 1.) Please enter your answer in the text box below. If applicable, please provide at least two digits after the decimal place.

You entered:

-0.46982

Your Answer		Score	Explanation
-0.46982	✔	1.00	
Total		1.00 / 1.00	

Question Explanation

The mean of x_2 is 6675.5 and the range is $8836 - 4761 = 4075$ So $x_1^{(4)}$ is $\frac{4761-6675.5}{4075} = -0.47.$

Question 2

You run gradient descent for 15 iterations with $\alpha = 0.3$ and compute $J(\theta)$ after each iteration. You find that the value of $J(\theta)$ **decreases** quickly then levels off. Based on this, which of the following conclusions seems most plausible?

Your Answer		Score	Explanation
Rather than use the current value of α, it'd be more promising to try a smaller value of α (say $\alpha = 0.1$).			
Rather than use the current value of α, it'd be more promising to try a larger value of α (say $\alpha = 1.0$).			
$\alpha = 0.3$ is an effective choice of learning rate.	✔	1.00	We want gradient descent to quickly converge to the minimum, so the current setting of α seems to be good.
Total		1.00 / 1.00	

Question 3

Suppose you have $m = 14$ training examples with $n = 3$ features (excluding the additional all-ones feature for the intercept term, which you should add). The normal equation is $\theta = (X^T X)^{-1} X^T y$. For the given values of m and n, what are the dimensions of θ, X, and y in this equation?

Your Answer		Score	Explanation
X is 14×4, y is 14×1, θ is 4×1	✔	1.00	
X is 14×3, y is 14×1, θ is 3×3			
X is 14×4, y is 14×4, θ is 4×4			
X is 14×3, y is 14×1, θ is 3×1			
Total		1.00 / 1.00	

Question Explanation

X has m rows and $n+1$ columns (+1 because of the $x_0 = 1$ term). y is an m-vector. θ is an ($n+1$)-vector.

Question 4

Suppose you have a dataset with $m = 1000000$ examples and $n = 200000$ 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?

Your Answer		Score	Explanation
○ Gradient descent, since it will always converge to the optimal θ.			
◉ Gradient descent, since $(X^T X)^{-1}$ will be very slow to compute in the normal equation.	✔	1.00	With $n = 200000$ features, you will have to invert a 200001×200001 matrix to compute the normal equation. Inverting such a large matrix is computationally expensive, so gradient descent is a good choice.
○ The normal equation, since gradient descent might be unable to find the optimal θ.			
○ The normal equation, since it provides an efficient way to directly find the solution.			
Total		1.00 / 1.00	

Question 5

Which of the following are reasons for using feature scaling?

Your Answer	Score	Explanation

☑ It speeds up gradient descent by making it require fewer iterations to get to a good solution.	✔	0.25	Feature scaling speeds up gradient descent by avoiding many extra iterations that are required when one or more features take on much larger values than the rest.
☐ It prevents the matrix $X^T X$ (used in the normal equation) from being non-invertable (singular/degenerate).	✔	0.25	$X^T X$ can be singular when features are redundant or there are too few examples. Feature scaling does not solve these problems.
☐ It speeds up gradient descent by making each iteration of gradient descent less expensive to compute.	✔	0.25	The magnitude of the feature values are insignificant in terms of computational cost.
☐ It speeds up solving for θ using the normal equation.	✔	0.25	The magnitude of the feature values are insignificant in terms of computational cost.
Total		1.00 / 1.00	

Feedback — VI. Logistic Regression

Help Center

You submitted this quiz on **Sat 14 Feb 2015 1:29 PM CET**. You got a score of **5.00** out of **5.00**.

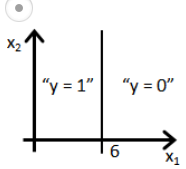
Question 1

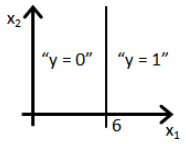
Suppose that you have trained a logistic regression classifier, and it outputs on a new example x a prediction $h_\theta(x) = 0.7$. This means (check all that apply):

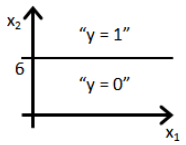
Your Answer		Score	Explanation
☐ Our estimate for $P(y = 0\|x; \theta)$ is 0.7.	✔	0.25	$h_\theta(x)$ is $P(y = 1\|x; \theta)$, not $P(y = 0\|x; \theta)$.
☑ Our estimate for $P(y = 1\|x; \theta)$ is 0.7.	✔	0.25	$h_\theta(x)$ is precisely $P(y = 1\|x; \theta)$, so each is 0.7.
☑ Our estimate for $P(y = 0\|x; \theta)$ is 0.3.	✔	0.25	Since we must have $P(y = 0\|x; \theta) = 1 - P(y = 1\|x; \theta)$, the former is 1 - 0.7 = 0.3.
☐ Our estimate for $P(y = 1\|x; \theta)$ is 0.3.	✔	0.25	$h_\theta(x)$ gives $P(y = 1\|x; \theta)$, not $1 - P(y = 1\|x; \theta)$.
Total		1.00 / 1.00	

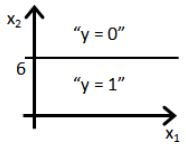
Question 2

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

Your Answer		Score	Explanation
◉ 	✔	1.00	In this figure, we transition from negative to positive when x_1 goes from above 6 to below 6 which is true for the given values of θ.





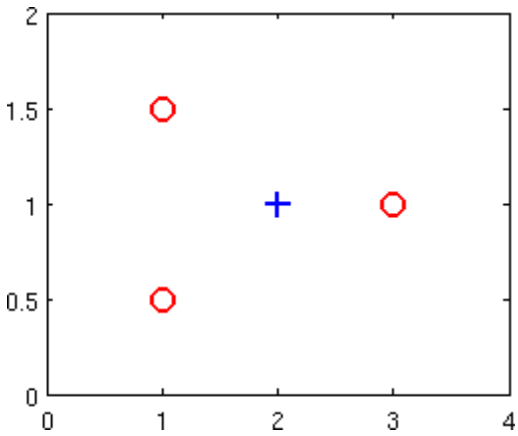


Total	1.00 / 1.00

Question 3

Suppose you have the following training set, and fit a logistic regression classifier $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$.

x_1	x_2	y
1	0.5	0
1	1.5	0
2	1	1
3	1	0



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

Your Answer		Score	Explanation
☑ $J(\theta)$ will be a convex function, so gradient descent should converge to the global minimum.	✔	0.25	The cost function $J(\theta)$ is guaranteed to be convex for logistic regression.
☑ Adding polynomial features (e.g., instead using $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1 x_2 + \theta_5 x_2^2)$) could increase how well we can fit the training data.	✔	0.25	Adding new features can only improve the fit on the training set:

			since setting $\theta_3 = \theta_4 = \theta_5 = 0$ makes the hypothesis the same as the original one, gradient descent will use those features (by making the corresponding θ_j non-zero) only if doing so improves the training set fit.
☐ The positive and negative examples cannot be separated using a straight line. So, gradient descent will fail to converge.	✔	0.25	While it is true they cannot be separated, gradient descent will still converge to the optimal fit. Some examples will remain misclassified at the optimum.
☐ Because the positive and negative examples cannot be separated using a straight line, linear regression will perform as well as logistic regression on this data.	✔	0.25	While it is true they cannot be separated, logistic regression will outperform linear regression since its cost function focuses on classification, not prediction.
Total		1.00 / 1.00	

Question 4

For logistic regression, the gradient is given by $\frac{\partial}{\partial \theta_j} J(\theta) = \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$. Which of these is a correct gradient descent update for logistic regression with a learning rate of α? Check all that apply.

Your Answer		Score	Explanation
☐ $\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(\theta^T x - y^{(i)}\right) x_j^{(i)}$	✔	0.25	This uses the linear regression

(simultaneously update for all j).			hypothesis $\theta^T x$ instead of that for logistic regression.
☑ $\theta := \theta - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{1+e^{-\theta^T x^{(i)}}} - y^{(i)} \right) x^{(i)}$.	✔	0.25	This is a vectorized version of gradient descent that substitues in the exact form of $h_\theta(x^{(i)})$ used by logistic regression.
☐ $\theta := \theta - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(\theta^T x - y^{(i)} \right) x^{(i)}$.	✔	0.25	This vectorized version uses the linear regression hypothesis $\theta^T x$ instead of that for logistic regression.
☑ $\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{1+e^{-\theta^T x^{(i)}}} - y^{(i)} \right) x_j^{(i)}$ (simultaneously update for all j).	✔	0.25	This substitutes the exact form of $h_\theta(x^{(i)})$ used by logistic regression into the gradient descent update.
Total		1.00 / 1.00	

Question 5

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

Your Answer		Score	Explanation
☑ The cost function $J(\theta)$ for logistic regression trained with $m \geq 1$ examples is always greater than or equal to zero.	✔	0.25	The cost for any example $x^{(i)}$ is always ≥ 0 since it is the negative log of a quantity less than one. The cost function $J(\theta)$ is a summation over the cost for each eample, so the cost function itself must be greater than or equal to zero.
☐ For logistic regression, sometimes gradient descent will converge to a local minimum (and fail to find the global minimum). This is the reason we prefer more advanced optimization	✔	0.25	The cost function for logistic regression is convex, so gradient descent will always converge to the global minimum. We still might use a more advanded optimization algorithm since they can be faster and don't require you to select a learning rate.

algorithms such as fminunc (conjugate gradient/BFGS/L-BFGS/etc).			
☐ Linear regression always works well for classification if you classify by using a threshold on the prediction made by linear regression.	✔	0.25	As demonstrated in the lecture, linear regression often classifies poorly since its training prodcedure focuses on predicting real-valued outputs, not classification.
☑ The sigmoid function $g(z) = \frac{1}{1+e^{-z}}$ is never greater than one (> 1).	✔	0.25	The denomiator ranges from ∞ to 1 as z grows, so the result is always in $(0, 1)$.
Total		1.00 / 1.00	

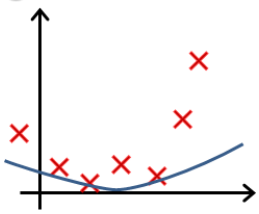
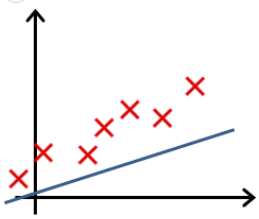
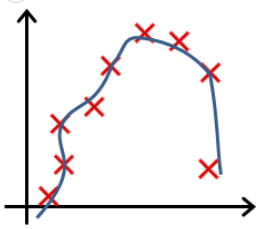
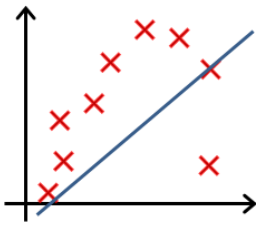
Feedback — VII. Regularization

Help Center

You submitted this quiz on **Sat 14 Feb 2015 12:57 PM CET**. You got a score of **5.00** out of **5.00**.

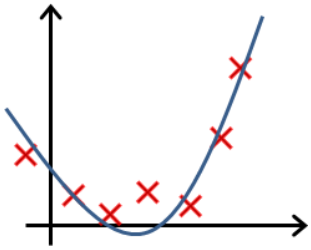
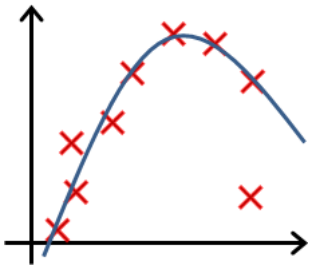
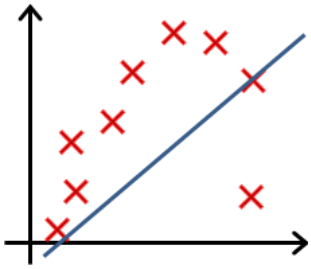
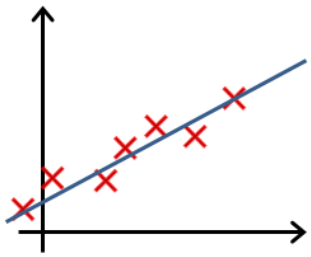
Question 1

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

Your Answer	Score	Explanation
		
		
	✔ 1.00	The hypothesis follows the data points very closely and is highly complicated, indicating that it is overfitting the training set.
		
Total	1.00 / 1.00	

Question 2

In which one of the following figures do you think the hypothesis has underfit the training set?

Your Answer	Score	Explanation
		
		
	✔ 1.00	The hypothesis does not predict many data points well, and is thus underfitting the training set.
		
Total	1.00 / 1.00	

Question 3

You are training a classification model with logistic regression. Which of the following statements are true? Check all that apply.

Your Answer		Score	Explanation
☐ Adding many new features to the model helps prevent overfitting on the training set.	✔	0.25	Adding many new features gives us more expressive models which are able to better fit our training set. If too many new features are added, this can lead to overfitting of the training set.
☐ Introducing regularization to the model always results in equal or better performance on examples not in the training set.	✔	0.25	If we introduce too much regularization, we can underfit the training set and this can lead to worse performance even for examples not in the training set.
☐ Introducing regularization to the model always results in equal or better performance on the training set.	✔	0.25	If we introduce too much regularization, we can underfit the training set and have worse performance on the training set.
☑ Adding many new features to the model makes it more likely to overfit the training set.	✔	0.25	Adding many new features gives us more expressive models which are able to better fit our training set. If too many new features are added, this can lead to overfitting of the training set.
Total		1.00 / 1.00	

Question 4

Suppose you ran logistic regression twice, once with $\lambda = 0$, and once with $\lambda = 1$. One of the 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 value of λ corresponds to which value of θ. Which one do you think corresponds to $\lambda = 1$?

Your Answer	Score	Explanation

○ $\theta = \begin{bmatrix} 26.29 \\ 65.41 \end{bmatrix}$			
◉ $\theta = \begin{bmatrix} 2.75 \\ 1.32 \end{bmatrix}$	✔	1.00	When λ is set to 1, we use regularization to penalize large values of θ. Thus, the parameters, θ, obtained will in general have smaller values.
Total		1.00 / 1.00	

Question 5

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

Your Answer		Score	Explanation
☐ Using a very large value of λ cannot hurt the performance of your hypothesis; the only reason we do not set λ to be too large is to avoid numerical problems.	✔	0.25	Using a very large value of λ can lead to underfitting of the training set.
☐ Using too large a value of λ can cause your hypothesis to overfit the data; this can be avoided by reducing λ.	✔	0.25	Using a very large value of λ can lead to underfitting of the training set.
☐ Because regularization causes $J(\theta)$ to no longer be convex, gradient descent may not always converge to the global minimum (when $\lambda > 0$, and when using an appropriate learning rate α).	✔	0.25	Regularized logistic regression and regularized linear regression are both convex, and thus gradient descent will still converge to the global minimum.
☑ Using too large a value of λ can cause your hypothesis to underfit the data.	✔	0.25	A large value of λ results in a large regularization penalty and thus a strong preference for simpler models which can underfit the data.
Total		1.00 / 1.00	

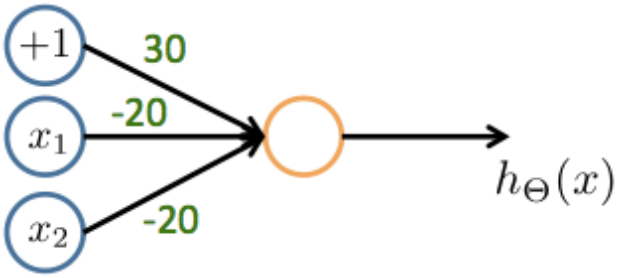
Feedback — VIII. Neural Networks: Representation

Help Center

You submitted this quiz on **Tue 7 Apr 2015 10:24 AM CEST**. You got a score of **5.00** out of **5.00**.

Question 1

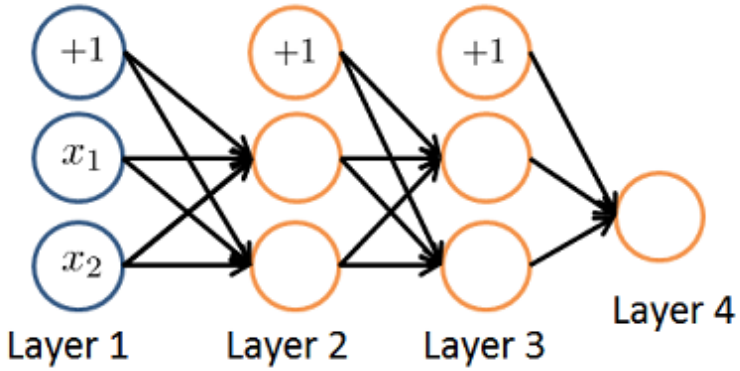
Consider the following neural network which takes two binary-valued inputs $x_1, x_2 \in \{0, 1\}$ and outputs $h_\Theta(x)$. Which of the following logical functions does it (approximately) compute?



Your Answer		Score	Explanation
XOR (exclusive OR)			
AND			
OR			
NAND (meaning "NOT AND")	✔	1.00	This network outputs approximately 1 as long as one of the two inputs is 0.
Total		1.00 / 1.00	

Question 2

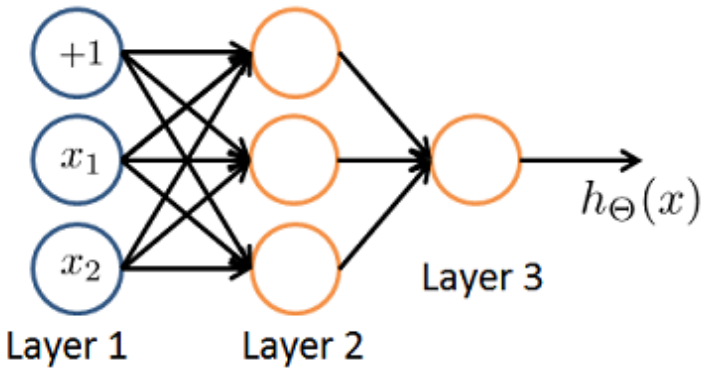
Consider the neural network given below. Which of the following equations correctly computes the activation $a_1^{(3)}$? Note: $g(z)$ is the sigmoid activation function.



Your Answer		Score	Explanation
○ $a_1^{(3)} = g(\Theta_{1,0}^{(1)} a_0^{(2)} + \Theta_{1,1}^{(1)} a_1^{(2)} + \Theta_{1,2}^{(1)} a_2^{(2)})$			
◉ $a_1^{(3)} = g(\Theta_{1,0}^{(2)} a_0^{(2)} + \Theta_{1,1}^{(2)} a_1^{(2)} + \Theta_{1,2}^{(2)} a_2^{(2)})$	✔	1.00	This correctly uses the first row of $\Theta^{(2)}$ and includes the "+1" term of $a_0^{(2)}$.
○ The activation $a_1^{(3)}$ is not present in this network.			
○ $a_1^{(3)} = g(\Theta_{1,0}^{(1)} a_0^{(1)} + \Theta_{1,1}^{(1)} a_1^{(1)} + \Theta_{1,2}^{(1)} a_2^{(1)})$			
Total		1.00 / 1.00	

Question 3

You have the following neural network:



You'd like to compute the activations of the hidden layer $a^{(2)} \in \mathbb{R}^3$. One way to do so is the following Octave code:

```
% Theta1 is Theta with superscript "(1)" from lecture
% ie, the matrix of parameters for the mapping from layer 1 (input) to laye
r 2
% Theta1 has size 3x3
% Assume 'sigmoid' is a built-in function to compute 1 / (1 + exp(-z))

a2 = zeros (3, 1);
for i = 1:3
  for j = 1:3
    a2(i) = a2(i) + x(j) * Theta1(i, j);
  end
  a2(i) = sigmoid (a2(i));
end
```

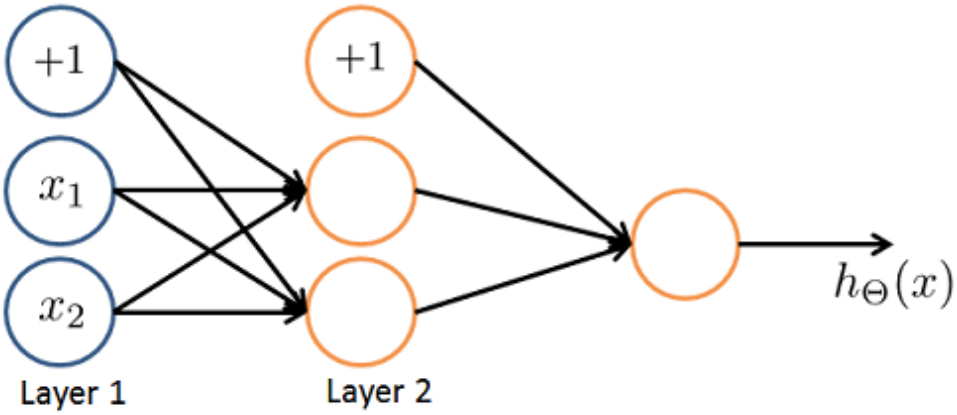
You want to have a vectorized implementation of this (i.e., one that does not use for loops). Which of the following implementations correctly compute $a^{(2)}$? Check all that apply.

Your Answer		Score	Explanation
☐ a2 = sigmoid (Thet a2 * x);	✔	0.25	$\Theta^{(2)}$ specifies the parameters from the second to third layers, not first to second.
☐ a2 = sigmoid (x * T heta1);	✔	0.25	The order of the multiplication is important, this will not work as x is a vector of size 3×1 while Theta1 is a matrix of size $3x3$.
☐ z = sigmoid(x); a2 = sigmoid (Theta1 * z);	✔	0.25	You do not need to apply the sigmoid function to the inputs.
☑ z = Theta1 * x; a2 = sigmoid (z);	✔	0.25	This version computes $a^{(2)} = g(\Theta^{(1)}x)$ correctly in two steps, first the multiplication and then the sigmoid activation.
Total		1.00 / 1.00	

Question 4

You are using the neural network pictured below and have learned the parameters $\Theta^{(1)} = \begin{bmatrix} 1 & 1 & 2.4 \\ 1 & 1.7 & 3.2 \end{bmatrix}$ (used to compute $a^{(2)}$) and $\Theta^{(2)} = \begin{bmatrix} 1 & 0.3 & -1.2 \end{bmatrix}$ (used to compute $a^{(3)}$} as a function of $a^{(2)}$). Suppose you swap the parameters for the first hidden layer

between its two units so $\Theta^{(1)} = \begin{bmatrix} 1 & 1.7 & 3.2 \\ 1 & 1 & 2.4 \end{bmatrix}$ and also swap the output layer so $\Theta^{(2)} = \begin{bmatrix} 1 & -1.2 & 0.3 \end{bmatrix}$. How will this change the value of the output $h_\Theta(x)$?



Your Answer		Score	Explanation
◉ It will stay the same.	✔	1.00	Swapping $\Theta^{(1)}$ swaps the hidden layers output $a^{(2)}$. But the swap of $\Theta^{(2)}$ cancels out the change, so the output will remain unchanged.
○ Insufficient information to tell: it may increase or decrease.			
○ It will increase.			
○ It will decrease			
Total		1.00 / 1.00	

Question 5

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

Your Answer		Score	Explanation
☐ Suppose you have a multi-class classification problem with three classes, trained with a 3 layer network. Let $a_1^{(3)} = (h_\Theta(x))_1$ be the activation of the first output unit, and	✔	0.25	The outputs of a neural network are not probabilities, so their sum need not be 1.

similarly $a_2^{(3)} = (h_\Theta(x))_2$ and $a_3^{(3)} = (h_\Theta(x))_3$. Then for any input x, it must be the case that $a_1^{(3)} + a_2^{(3)} + a_3^{(3)} = 1$.			
☐ A two layer (one input layer, one output layer; no hidden layer) neural network can represent the XOR function.	✔	0.25	We must compose multiple logical operations by using a hidden layer to represent the XOR function.
☑ Any logical function over binary-valued (0 or 1) inputs x_1 and x_2 can be (approximately) represented using some neural network.	✔	0.25	Since we can build the basic AND, OR, and NOT functions with a two layer network, we can (approximately) represent any logical function by composing these basic functions over multiple layers.
☑ If a neural network is overfitting the data, one solution would be to increase the regularization parameter λ.	✔	0.25	A larger value of λ will shrink the magnitude of the parameters Θ, thereby reducing the chance of overfitting the data.
Total		1.00 / 1.00	

Feedback — IX. Neural Networks: Learning

Help Center

You submitted this quiz on **Mon 9 Mar 2015 9:50 AM CET**. You got a score of **5.00** out of **5.00**.

Question 1

You are training a three layer neural network and would like to use backpropagation to compute the gradient of the cost function. In the backpropagation algorithm, one of the steps is to update $\Delta_{ij}^{(2)} := \Delta_{ij}^{(2)} + \delta_i^{(3)} * (a^{(2)})_j$ for every i, j. Which of the following is a correct vectorization of this step?

Your Answer		Score	Explanation
$\Delta^{(2)} := \Delta^{(2)} + \delta^{(3)} * (a^{(2)})^T$	✔	1.00	This version is correct, as it takes the "outer product" of the two vectors $\delta^{(3)}$ and $a^{(2)}$ which is a matrix such that the (i, j)-th entry is $\delta_i^{(3)} * (a^{(2)})_j$ as desired.
$\Delta^{(2)} := \Delta^{(2)} + (a^{(3)})^T * \delta^{(3)}$			
$\Delta^{(2)} := \Delta^{(2)} + \delta^{(2)} * (a^{(2)})^T$			
$\Delta^{(2)} := \Delta^{(2)} + (a^{(2)})^T * \delta^{(3)}$			
Total		1.00 / 1.00	

Question 2

Suppose `Theta1` is a 2x5 matrix, and `Theta2` is a 3x6 matrix. You set `thetaVec = [Theta1(:) ; Theta2(:)]`. Which of the following correctly recovers `Theta2`?

Your Answer	Score	Explanation

`reshape(thetaVec(11:20), 3, 6)`			
`reshape(thetaVec(9:26), 3, 6)`			
`reshape(thetaVec(11:28), 3, 6)`	✔	1.00	This choice is correct, since `Theta1` has 10 elements, so `Theta2` begins at index 11 and ends at index 11 + 18 - 1 = 28.
`reshape(thetaVec(1:18), 3, 6)`			
Total		1.00 / 1.00	

Question 3

Let $J(\theta) = 3\theta^4 + 4$. Let $\theta = 1$, and $\epsilon = 0.01$. Use the formula $\frac{J(\theta+\epsilon)-J(\theta-\epsilon)}{2\epsilon}$ to numerically compute an approximation to the derivative at $\theta = 1$. What value do you get? (When $\theta = 1$, the true/exact derivati ve is $\frac{dJ(\theta)}{d\theta} = 12$.)

Your Answer		Score	Explanation
6			
12.0012	✔	1.00	We compute $\frac{(3(1.01)^4+4)-(3(0.99)^4+4)}{2(0.01)} = 12.0012$.
12			
11.9988			
Total		1.00 / 1.00	

Question 4

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

Your Answer		Score	Explanation
For computational	✔	0.25	Checking the gradient numerically is a debugging tool:

efficiency, after we have performed gradient checking to verify that our backpropagation code is correct, we usually disable gradient checking before using backpropagation to train the network.			it helps ensure a corre ct implementation, but it is too slow to use as a method for actually computing gradients.
☐ Using a large value of λ cannot hurt the performance of your neural network; the only reason we do not set λ to be too large is to avoid numerical problems.	✔	0.25	A large value of λ can be quite detrimental. If you set it too high, then the network will be underfit to the training data and give poor predictions on both training data and new, unseen test data.
☑ Using gradient checking can help verify if one's implementation of backpropagation is bug-free.	✔	0.25	If the gradient computed by backpropagation is the same as one computed numerically with gradient checking, this is very strong evidence that you have a correct implementation of backpropagation.
☐ Gradient checking is useful if we are using one of the advanced optimization methods (such as in fminunc) as our optimization algorithm. However, it serves little purpose if we are using gradient descent.	✔	0.25	Gradient descent depends on the computation of correct gradient values at different parameter settings. Gradient checking ensures the computed values are correct.
Total		1.00 / 1.00	

Question 5

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

Your Answer	Score	Explanation

☐ Suppose you have a three layer network with parameters $\Theta^{(1)}$ (controlling the function mapping from the inputs to the hidden units) and $\Theta^{(2)}$ (controlling the mapping from the hidden units to the outputs). If we set all the elements of $\Theta^{(1)}$ to be 0, and all the elements of $\Theta^{(2)}$ to be 1, then this suffices for symmetry breaking, since the neurons are no longer all computing the same function of the input.	✔	0.25	Since the parameters are the same within layers, every unit in each layer will receive the same update during backpropagation. The result is that such an initialization does not break symmetry.
☐ Suppose that the parameter $\Theta^{(1)}$ is a square matrix (meaning the number of rows equals the number of columns). If we replace $\Theta^{(1)}$ with its transpose $(\Theta^{(1)})^T$, then we have not changed the function that the network is computing.	✔	0.25	$\Theta^{(1)}$ can be an arbitrary matrix, so when you compute $a^{(2)} = g(\Theta^{(1)}a^{(1)})$, replacing $\Theta^{(1)}$ with its transpose will compute a different value.
☑ If we are training a neural network using gradient descent, one reasonable "debugging" step to make sure it is working is to plot $J(\Theta)$ as a function of the number of	✔	0.25	Since gradient descent uses the gradient to take a step toward parameters with lower cost (ie, lower $J(\Theta)$), the value of $J(\Theta)$ should be equal or less at each iteration if the gradient computation is correct and the learning rate is set properly.

iterations, and make sure it is decreasing (or at least non-increasing) after each iteration.			
☑ Suppose you are training a neural network using gradient descent. Depending on your random initialization, your algorithm may converge to different local optima (i.e., if you run the algorithm twice with different random initializations, gradient descent may converge to two different solutions).	✔	0.25	The cost function for a neural network is non-convex, so it may have multiple minima. Which minimum you find with gradient descent depends on the initialization.
Total		1.00 / 1.00	

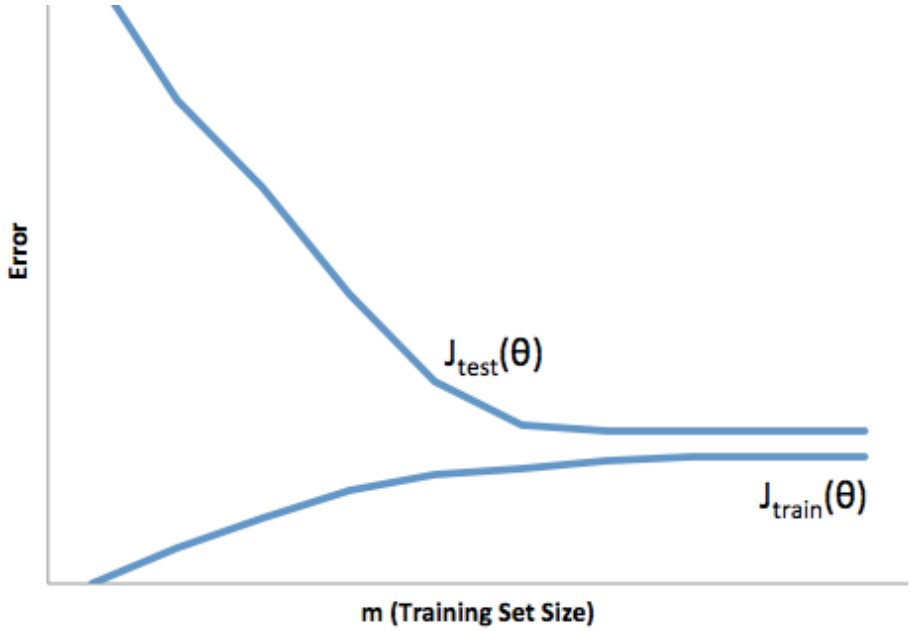
Feedback — X. Advice for Applying Machine Learning

Help Center

You submitted this quiz on **Tue 7 Apr 2015 10:41 AM CEST**. You got a score of **5.00** out of **5.00**.

Question 1

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?



Your Answer		Score	Explanation
○ Neither			
◉ High bias	✔	1.00	This learning curve shows high error on both the training and test sets, so the algorithm is suffering from high bias.
○ High variance			
Total		1.00 / 1.00	

Question 2

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.

Your Answer		Score	Explanation
☑ Get more training examples.	✔	0.25	The gap in errors between training and test suggests a high variance problem in which the algorithm has overfit the training set. Adding more training data will increase the complexity of the training set and help with the variance problem.
☐ Try decreasing the regularization parameter λ.	✔	0.25	The gap in errors between training and test suggests a high variance problem in which the algorithm has overfit the training set. Decreasing the regularization parameter will increase the overfitting, not decrease it.
☐ Use fewer training examples.	✔	0.25	The gap in errors between training and test suggests a high variance problem in which the algorithm has overfit the training set. Using fewer training examples will only exacerbate the overfitting.
☑ Try increasing the regularization parameter λ.	✔	0.25	The gap in errors between training and test suggests a high variance problem in which the algorithm has overfit the training set. Increasing the regularization parameter will reduce overfitting and help with the variance problem.
Total		1.00 / 1.00	

Question 3

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.

Your Answer		Score	Explanation
☐ Try increasing the regularization parameter λ.	✔	0.25	The poor performance on both the training and test sets suggests a high bias problem. Increasing regularization will decrease the fit of the hypothesis to the data, exacerbating the high bias problem.
☐ Use fewer training examples.	✔	0.25	Using fewer training examples should never improve test set performance, as the model has fewer data points from which to learn.
☑ Try decreasing the regularization parameter λ.	✔	0.25	The poor performance on both the training and test sets suggests a high bias problem. Decreasing the regularization parameter will allow the hypothesis to fit the data more closely, improving both training and test set performance.
☑ Try to obtain and use additional features.	✔	0.25	The poor performance on both the training and test sets suggests a high bias problem. Using additional features will increase the complexity of the hypothesis, thereby improving the fit to both the train and test data.
Total		1.00 / 1.00	

Question 4

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

Your Answer		Score	Explanation
☑ 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.	✔	0.25	The cross validation lets us find the "just right" setting of the regularization parameter given the fixed model parameters learned from the training set.
☑ The performance of a learning algorithm on the	✔	0.25	The learning algorithm finds parameters to minimize training set error, so the performance should be better on the training set than the test set.

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.	✔	0.25	You should not use training error to choose the regularization parameter, as you can always improve training error by using less regularization (a smaller value of λ). But too small of a value will not generalize well on the test set.
☐ It is okay to use data from the test set to choose the regularization parameter λ, but not the model parameters (θ).	✔	0.25	You should not use test set data in choosing the regularization parameter, as it means the test error will not be a good estimate of generalization error.
Total		1.00 / 1.00	

Question 5

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

Your Answer		Score	Explanation
☑ If a learning algorithm is suffering from high bias, only adding more training examples may **not** improve the test error significantly.	✔	0.25	With high bias, the model is not fitting the training data currently present, so adding more data is unlikely to help.
☑ A model with more parameters is more prone to overfitting and typically has higher variance.	✔	0.25	More model parameters increases the model's complexity, so it can more tightly fit data in training, increasing the chances of overfitting.
☑ When debugging learning	✔	0.25	The shape of a learning curve is a good indicator

algorithms, it is useful to plot a learning curve to understand if there is a high bias or high variance problem.			of bias or variance problems with your learning algorithm.
☐ If the training and test errors are about the same, adding more features will **not** help improve the results.	✔	0.25	If the two errors are the same, then the model has high bias, so adding more features will be helpful.
Total		1.00 / 1.00	

Feedback — XI. Machine Learning System Design Help Center

You submitted this quiz on **Tue 7 Apr 2015 10:16 AM CEST**. You got a score of **5.00** out of **5.00**.

Question 1

You are working on a spam classification system using regularized logistic regression. "Spam" is the 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	**0**
Predicted Class	**1**	85	890
	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 F_1 score (as a value from 0 to 1)? Enter your answer in the box below. If necessary, provide at least two values after the decimal point.

You entered:

0.15814

Your Answer		Score	Explanation
0.15814	✔	1.00	Precision is 0.087 and recall is 0.85, so F_1 score is (2 * precision * recall) / (precision + recall) = 0.158.
Total		1.00 / 1.00	

Question 2

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?

Your Answer		Score	Explanation
☐ When we are willing to include high order polynomial features of x (such as x_1^2, x_2^2, $x_1 x_2$, etc.).	✔	0.25	As we saw with neural networks, polynomial features can still be insufficient to capture the complexity of the data, especially if the features are very high-dimensional. Instead, you should use a complex model with many parameters to fit to the large training set.
☑ 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).	✔	0.25	It is important that the features contain sufficient information, as otherwise no amount of data can solve a learning problem in which the features do not contain enough information to make an accurate prediction.
☑ 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).	✔	0.25	You should use a complex, "low bias" algorithm, as it will be able to make use of the large dataset provided. If the model is too simple, it will underfit the large training set.
☐ We train a learning algorithm with a small number of parameters (that is thus unlikely to overfit).	✔	0.25	If the model has a small number of parameters, then it will underfit the large training set and not make good use of all the data.
Total		1.00 / 1.00	

Question 3

Suppose you have trained a logistic regression classifier which is outputing $h_\theta(x)$. Currently, you predict 1 if $h_\theta(x) \geq \text{threshold}$, and predict 0 if $h_\theta(x) < \text{threshold}$, where currently the threshold is set to 0.5. Suppose you **decrease** the threshold to 0.1. Which of the following are true? Check all that apply.

Your Answer		Score	Explanation
☐ The classifier is likely to have unchanged precision and recall, and thus the same F_1 score.	✔	0.25	By making more y = 1 predictions, we increase true and false positives and decrease true and false negatives. Thus, precision and recall will certainly change.
☐ The classifier is likely to now have higher precision.	✔	0.25	Lowering the threshold means more y = 1 predictions. This will increase both true and false positives, so precision will decrease, not increase.
☑ The classifier is likely to now have lower precision.	✔	0.25	Lowering the threshold means more y = 1 predictions. This will increase both true and false positives, so precision will decrease.
☐ The classifier is likely to have unchanged precision and recall, but higher accuracy.	✔	0.25	By making more y = 1 predictions, we increase true and false positives and decrease true and false negatives. Thus, precision and recall will certainly change. We cannot say whether accuracy will increase or decrease.
Total		1.00 / 1.00	

Question 4

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.

Your Answer	Score	Explanation

☑ If you always predict non-spam (output $y = 0$), your classifier will have a recall of 0%.	✔	0.25	Since every prediction is y = 0, there will be no true positives, so recall is 0%.
☐ If you always predict spam (output $y = 1$), your classifier will have a recall of 0% and precision of 99%.	✔	0.25	Every prediction is y = 1, so recall is 100% and precision is only 1%.
☑ If you always predict spam (output $y = 1$), your classifier will have a recall of 100% and precision of 1%.	✔	0.25	Since every prediction is y = 1, there are no false negatives, so recall is 100%. Furthermore, the precision will be the fraction of examples with are positive, which is 1%.
☑ 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.	✔	0.25	The classifier achieves 99% accuracy on the training set because of how skewed the classes are. We can expect that the cross-validation set will be skewed in the same fashion, so the classifier will have approximately the same accuracy.
Total		1.00 / 1.00	

Question 5

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

Your Answer		Score	Explanation
☐ 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.	✔	0.20	You cannot know whether a huge dataset will be important until you have built a first version and find that the algorithm has high variance.

☐ After training a logistic regression classifier, you **must** use 0.5 as your threshold for predicting whether an example is positive or negative.	✔	0.20	You can and should adjust the threshold in logistic regression using cross validation data.
☑ 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.	✔	0.20	This process of error analysis is crucial in developing high performance learning systems, as the space of possible improvements to your system is very large, and it gives you direction about what to work on next.
☐ If your model is underfitting the training set, then obtaining more data is likely to help.	✔	0.20	If the model is underfitting the training data, it has not captured the information in the examples you already have. Adding further examples will not help any more.
☑ On skewed datasets (e.g., when there are more positive examples than negative examples), accuracy is not a good measure of performance and you should instead use F_1 score based on the precision and recall.	✔	0.20	You can always achieve high accuracy on skewed datasets by predicting the most the same output (the most common one) for every input. Thus the F_1 score is a better way to measure performance.
Total		1.00 / 1.00	

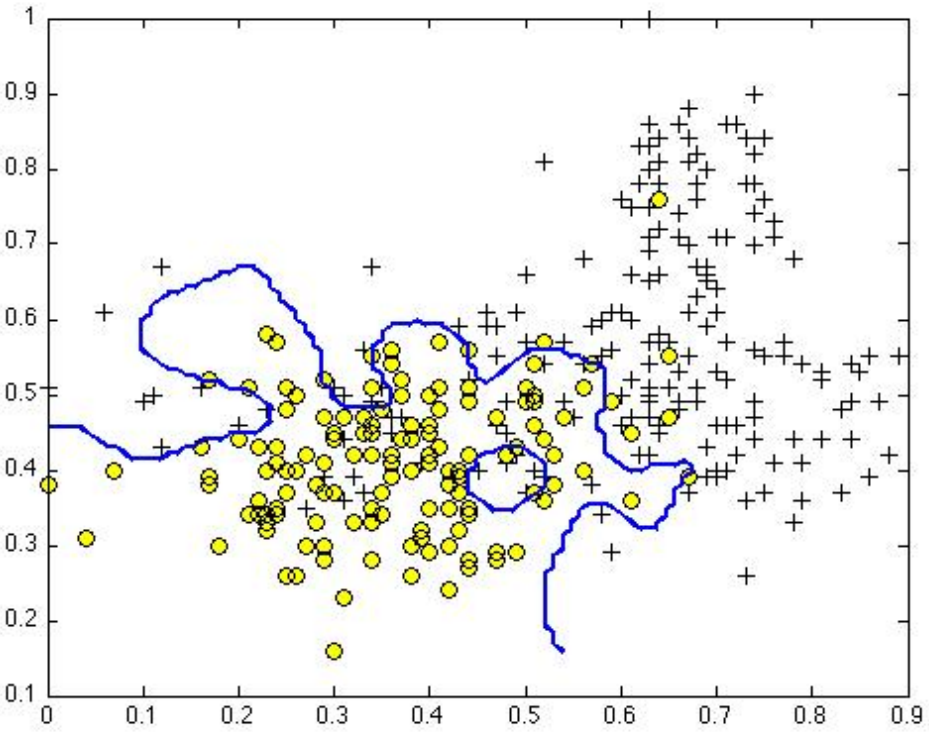
Feedback — XII. Support Vector Machines

Help Center

You submitted this quiz on **Tue 7 Apr 2015 11:03 AM CEST**. You got a score of **5.00** out of **5.00**.

Question 1

Suppose you have trained an SVM classifier with a Gaussian kernel, and it learned the following decision boundary on the training set:



When you measure the SVM's performance on a cross validation set, it does poorly. Should you try increasing or decreasing C? Increasing or decreasing σ^2?

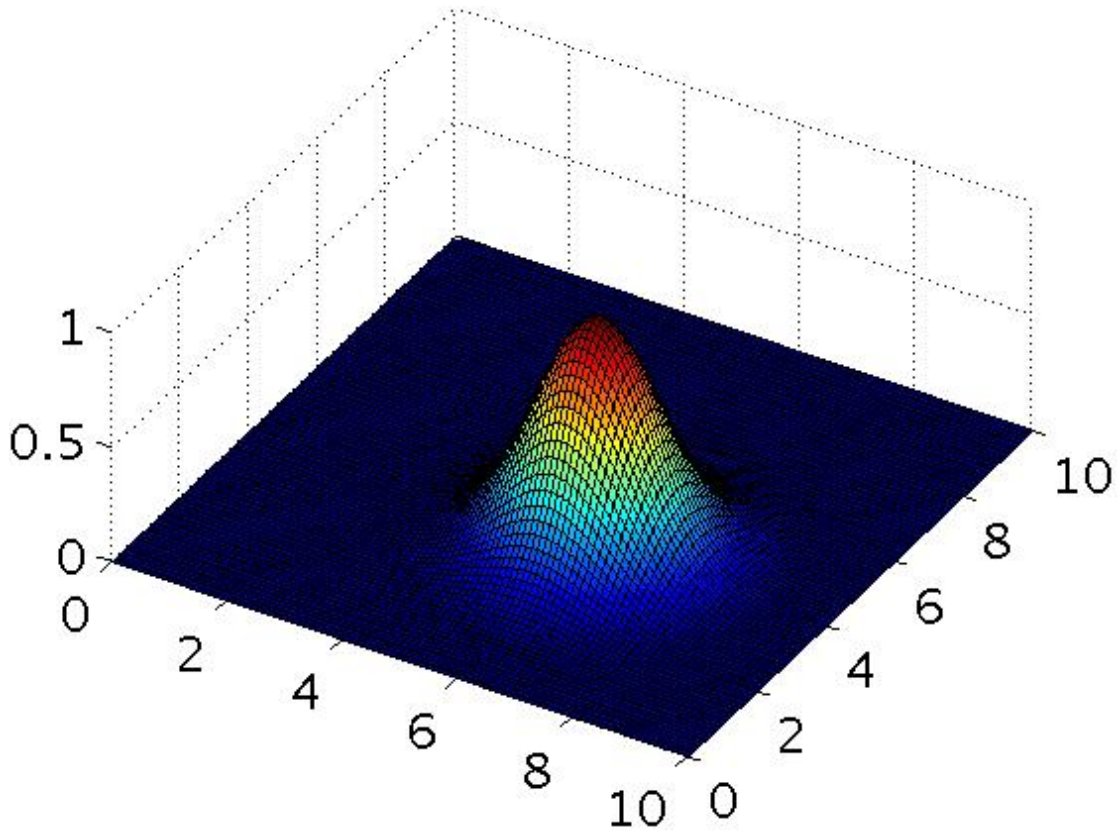
Your Answer		Score	Explanation
It would be reasonable to try **decreasing** C. It would also be reasonable to try **increasing** σ^2.	✔	1.00	The figure shows a decision boundary that is overfit to the training set, so we'd like to increase the bias / lower the variance of the SVM. We can do so by either decreasing the parameter C or increasing σ^2.

- It would be reasonable to try **increasing** C. It would also be reasonable to try **decreasing** σ^2.
- It would be reasonable to try **decreasing** C. It would also be reasonable to try **decreasing** σ^2.
- It would be reasonable to try **increasing** C. It would also be reasonable to try **increasing** σ^2.

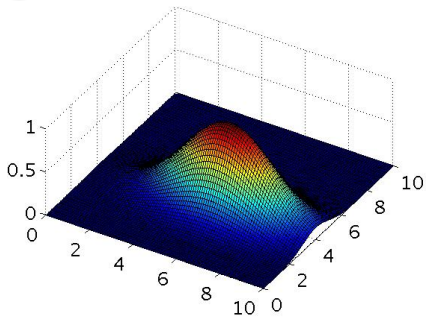
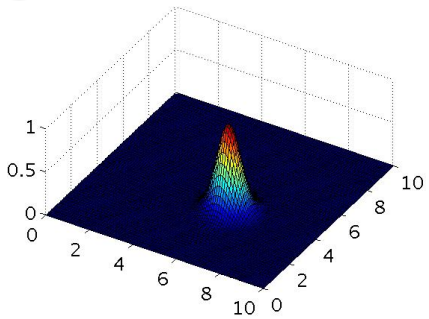
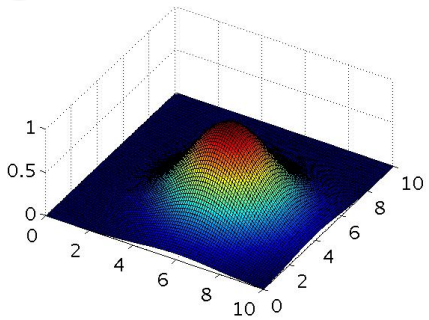
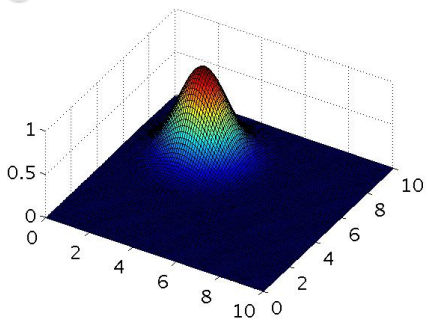
Total	1.00 / 1.00

Question 2

The formula for the Gaussian kernel is given by $\text{similarity}(x, l^{(1)}) = \exp\left(-\frac{\|x - l^{(1)}\|^2}{2\sigma^2}\right)$. The figure below shows a plot of $f_1 = \text{similarity}(x, l^{(1)})$ when $\sigma^2 = 1$.

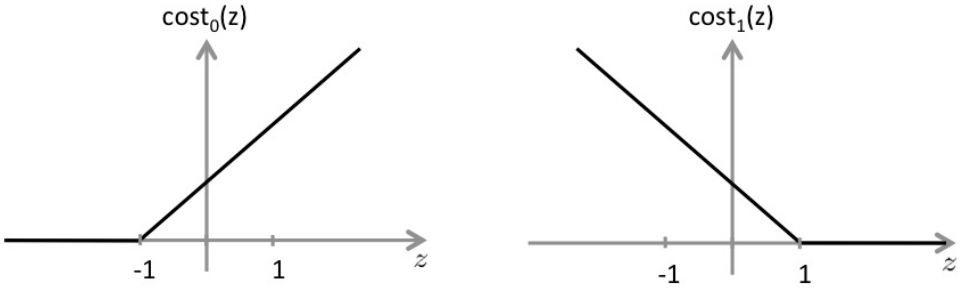


Which of the following is a plot of f_1 when $\sigma^2 = 0.25$?

Your Answer	Score	Explanation
		
	✔ 1.00	This figure shows a "narrower" Gaussian kernel centered at the same location which is the effect of decreasing σ^2.
		
		
Total	1.00 / 1.00	

Question 3

The SVM solves $\min_\theta \ C\sum_{i=1}^{m} y^{(i)}\text{cost}_1(\theta^T x^{(i)}) + (1-y^{(i)})\text{cost}_0(\theta^T x^{(i)}) + \sum_{j=1}^{n}\theta_j^2$ where the functions $\text{cost}_0(z)$ and $\text{cost}_1(z)$ look like this:



The first term in the objective is: $C\sum_{i=1}^{m} y^{(i)}\text{cost}_1(\theta^T x^{(i)}) + (1-y^{(i)})\text{cost}_0(\theta^T x^{(i)})$. This first term will be zero if two of the following four conditions hold true. Which are the two conditions that would guarantee that this term equals zero?

Your Answer		Score	Explanation
☑ For every example with $y^{(i)} = 1$, we have that $\theta^T x^{(i)} \geq 1$.	✔	0.25	For examples with $y^{(i)} = 1$, only the $\text{cost}_1(\theta^T x^{(i)})$ term is present. As you can see in the graph, this will be zero for all inputs greater than or equal to 1.
☑ For every example with $y^{(i)} = 0$, we have that $\theta^T x^{(i)} \leq -1$.	✔	0.25	For examples with $y^{(i)} = 0$, only the $\text{cost}_0(\theta^T x^{(i)})$ term is present. As you can see in the graph, this will be zero for all inputs less than or equal to -1.
☐ For every example with $y^{(i)} = 1$, we have that $\theta^T x^{(i)} \geq 0$.	✔	0.25	$\text{cost}_1(\theta^T x^{(i)})$ is still non-zero for inputs between 0 and 1, so being greater than or equal to 0 is insufficient.
☐ For every example with $y^{(i)} = 0$, we have that $\theta^T x^{(i)} \leq 0$.	✔	0.25	$\text{cost}_0(\theta^T x^{(i)})$ is still non-zero for inputs between -1 and 0, so being less than or equal to 0 is insufficient.
Total		1.00 / 1.00	

Question 4

Suppose you have a dataset with n = 10 features and m = 5000 examples. After training your logistic regression classifier with gradient descent, you find that it has underfit the training set and does not achieve the desired performance on the training or cross validation sets. Which of the

following might be promising steps to take? Check all that apply.

Your Answer		Score	Explanation
☑ Use an SVM with a Gaussian Kernel.	✔	0.25	By using a Gaussian kernel, your model will have greater complexity and can avoid underfitting the data.
☐ Use a different optimization method since using gradient descent to train logistic regression might result in a local minimum.	✔	0.25	The logistic regression cost function is convex, so gradient descent will always find the global minimum.
☑ Create / add new polynomial features.	✔	0.25	When you add more features, you increase the variance of your model, reducing the chances of underfitting.
☐ Reduce the number of examples in the training set.	✔	0.25	While you can improve accuracy on the training set by removing examples, doing so results in a worse model that will not generalize as well.
Total		1.00 / 1.00	

Question 5

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

Your Answer		Score	Explanation
☑ Suppose you have 2D input examples (ie, $x^{(i)} \in \mathbb{R}^2$). The decision boundary of the SVM (with the linear kernel) is a straight line.	✔	0.25	The SVM without any kernel (ie, the linear kernel) predicts output based only on $\theta^T x$, so it gives a linear / straight-line decision boundary, just as logistic regression does.
☐ Suppose you are using SVMs to do multi-class classification and would like to use the one-vs-all approach. If you have K different classes, you will train K - 1 different SVMs.	✔	0.25	The one-vs-all method requires that we have a separate classifier for every class, so you will train K different SVMs.
☑ It is important to	✔	0.25	The similarity measure used by the Gaussian kernel

perform feature normalization before using the Gaussian kernel.			expects that the data lie in approximately the same range.
☐ If you are training multi-class SVMs with the one-vs-all method, it is not possible to use a kernel.	✔	0.25	Each SVM you train in the one-vs-all method is a standard SVM, so you are free to use a kernel.
Total		1.00 / 1.00	

Feedback — XIII. Clustering

Help Center

You submitted this quiz on **Wed 8 Apr 2015 9:00 AM CEST**. You got a score of **5.00** out of **5.00**.

Question 1

For which of the following tasks might K-means clustering be a suitable algorithm? Select all that apply.

Your Answer		Score	Explanation
☐ Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.	✔	0.25	Such a prediction is a regression problem, and K-means does not use labels on the data, so it cannot perform regression.
☑ Given sales data from a large number of products in a supermarket, figure out which products tend to form coherent groups (say are frequently purchased together) and thus should be put on the same shelf.	✔	0.25	If you cluster the sales data with K-means, each cluster should correspond to coherent groups of items.
☐ Given many emails, you want to determine if they are Spam or Non-Spam emails.	✔	0.25	Classifying input as spam / non-spam requires labels for the data, which K-means does not use.
☑ Given a database of information about your users, automatically group them into different market segments.	✔	0.25	You can use K-means to cluster the database entries, and each cluster will correspond to a different market segment.
Total		1.00 / 1.00	

Question 2

Suppose we have three cluster centroids $\mu_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $\mu_2 = \begin{bmatrix} -3 \\ 0 \end{bmatrix}$ and $\mu_3 = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$. Furthermore, we have a training example $x^{(i)} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$. After a cluster assignment step, what will $c^{(i)}$ be?

Your Answer		Score	Explanation
○ $c^{(i)} = 1$			
◉ $c^{(i)} = 2$	✔	1.00	$x^{(i)}$ is closest to μ_2, so $c^{(i)} = 2$
○ $c^{(i)}$ is not assigned			
○ $c^{(i)} = 3$			
Total		1.00 / 1.00	

Question 3

K-means is an iterative algorithm, and two of the following steps are repeatedly carried out in its inner-loop. Which two?

Your Answer		Score	Explanation
☑ Move the cluster centroids, where the centroids μ_k are updated.	✔	0.25	The cluster update is the second step of the K-means loop.
☐ Test on the cross-validation set.	✔	0.25	Any sort of testing is outside the scope of the K-means algorithm itself.
☐ The cluster centroid assignment step, where each cluster centroid μ_i is assigned (by setting $c^{(i)}$) to the closest training example $x^{(i)}$.	✔	0.25	This is not a correct description of the cluster assignment step.
☑ The cluster assignment step, where the parameters $c^{(i)}$ are updated.	✔	0.25	This is the correst first step of the K-means loop.
Total		1.00 / 1.00	

Question 4

Suppose you have an unlabeled dataset $\{x^{(1)}, \ldots, x^{(m)}\}$. You run K-means with 50 different random initializations, and obtain 50 different clusterings of the data. What is the recommended way for choosing which one of these 50 clusterings to use?

Your Answer		Score	Explanation
○ Manually examine the clusterings, and pick the best one.			
○ Plot the data and the cluster centroids, and pick the clustering that gives the most "coherent" cluster centroids.			
◉ For each of the clusterings, compute $\frac{1}{m}\sum_{i=1}^{m} \|x^{(i)} - \mu_{c^{(i)}}\|^2$, and pick the one that minimizes this.	✔	1.00	This function is the distortion function. Since a lower value for the distortion function implies a better clustering, you should choose the clustering with the smallest value for the distortion function.
○ Use the elbow method.			
Total		1.00 / 1.00	

Question 5

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

Your Answer		Score	Explanation
☑ For some datasets, the "right" or "correct" value of K (the number of clusters) can be ambiguous, and hard even for a human expert looking carefully at the data to decide.	✔	0.25	In many datasets, different choices of K will give different clusterings which appear quite reasonable. With no labels on the data, we cannot say one is better than the other.
☑ If we are worried	✔	0.25	Since each run of K-means is independent, multiple

about K-means getting stuck in bad local optima, one way to ameliorate (reduce) this problem is if we try using multiple random initializations.			runs can find different optima, and some should avoid bad local optima.
☐ The standard way of initializing K-means is setting $\mu_1 = \cdots = \mu_k$ to be equal to a vector of zeros.	✔	0.25	This is a poor initialization, since every centroid needs to start in a different location. Otherwise, each will be updated in the same way at each iteration and they will never spread out into different clusters.
☐ Once an example has been assigned to a particular centroid, it will never be reassigned to another different centroid	✔	0.25	Each iteration of K-means performs a cluster assignment step in which each example may be assigned to a different centroid.
Total		1.00 / 1.00	

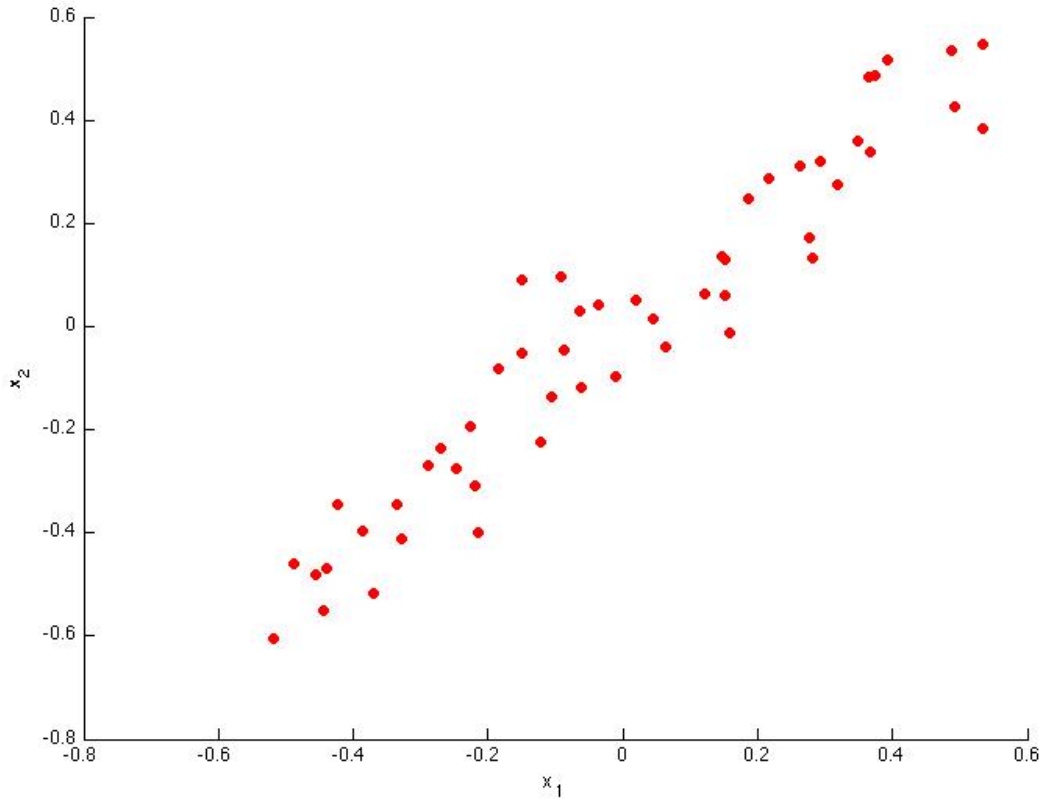
Feedback — XIV. Principal Component Analysis

Help Center

You submitted this quiz on **Wed 8 Apr 2015 9:18 AM CEST**. You got a score of **5.00** out of **5.00**.

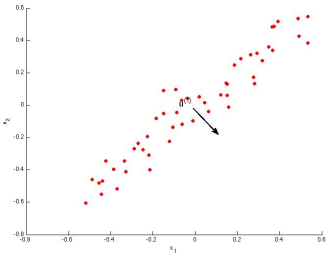
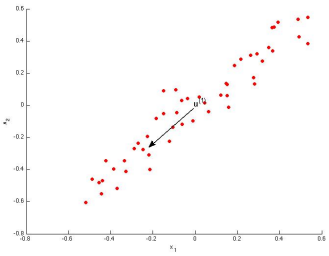
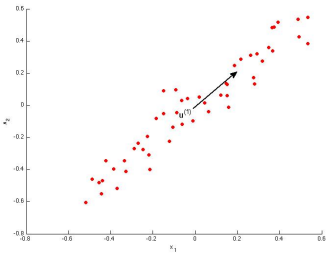
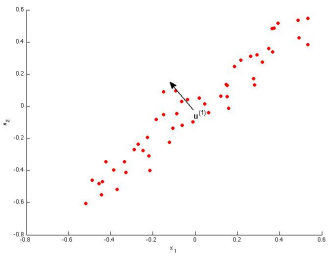
Question 1

Consider the following 2D dataset:



Which of the following figures correspond to possible values that PCA may return for $u^{(1)}$ (the first eigenvector / first principal component)? Check all that apply (you may have to check more than one figure).

Your Answer		Score	Explanation
☐	✔	0.25	The first principal component is aligned with the direction of maximal variance, but this is aligned with the direction of minimal variance.

			
☑ 	✔	0.25	The maximal variance is along the y = x line, so the negative vector along that line is correct for the first principal component.
☑ 	✔	0.25	The maximal variance is along the y = x line, so this option is correct.
☐ 	✔	0.25	The first principal component is aligned with the direction of maximal variance, but this is aligned with the direction of minimal variance.
Total		1.00 / 1.00	

Question 2

Which of the following is a reasonable way to select the number of principal components k?

(Recall that n is the dimensionality of the input data and m is the number of input examples.)

Your Answer		Score	Explanation
◉ Choose k to be the smallest value so that at least 99% of the variance is retained.	✔	1.00	This is correct, as it maintains the structure of the data while maximally reducing its dimension.

○ Choose k to be 99% of n (i.e., $k = 0.99 * n$, rounded to the nearest integer).		
○ Use the elbow method.		
○ Choose the value of k that minimizes the approximation error $\frac{1}{m}\sum_{i=1}^{m} \|\|x^{(i)} - x^{(i)}_{\text{approx}}\|\|^2$.		
Total		1.00 / 1.00

Question 3

Suppose someone tells you that they ran PCA in such a way that "95% of the variance was retained." What is an equivalent statement to this?

Your Answer		Score	Explanation
○ $\frac{\frac{1}{m}\sum_{i=1}^{m} \|\|x^{(i)}-x^{(i)}_{\text{approx}}\|\|^2}{\frac{1}{m}\sum_{i=1}^{m} \|\|x^{(i)}\|\|^2} \geq 0.05$			
○ $\frac{\frac{1}{m}\sum_{i=1}^{m} \|\|x^{(i)}-x^{(i)}_{\text{approx}}\|\|^2}{\frac{1}{m}\sum_{i=1}^{m} \|\|x^{(i)}\|\|^2} \leq 0.95$			
○ $\frac{\frac{1}{m}\sum_{i=1}^{m} \|\|x^{(i)}\|\|^2}{\frac{1}{m}\sum_{i=1}^{m} \|\|x^{(i)}-x^{(i)}_{\text{approx}}\|\|^2} \leq 0.95$			
◉ $\frac{\frac{1}{m}\sum_{i=1}^{m} \|\|x^{(i)}-x^{(i)}_{\text{approx}}\|\|^2}{\frac{1}{m}\sum_{i=1}^{m} \|\|x^{(i)}\|\|^2} \leq 0.05$	✔	1.00	This is the correct formula.
Total		1.00 / 1.00	

Question 4

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

Your Answer		Score	Explanation
☑ Given an input $x \in \mathbb{R}^n$, PCA compresses it to a lower-dimensional vector $z \in \mathbb{R}^k$.	✔	0.25	PCA compresses it to a lower dimensional vector by projecting it onto the learned principal components.

☑ Even if all the input features are on very similar scales, we should still perform mean normalization (so that each feature has zero mean) before running PCA.	✔	0.25	If you do not perform mean normalization, PCA will rotate the data in a possibly undesired way.
☐ PCA can be used only to reduce the dimensionality of data by 1 (such as 3D to 2D, or 2D to 1D).	✔	0.25	PCA can reduce data of dimension n to any dimension $k < n$.
☐ PCA is susceptible to local optima; trying multiple random initializations may help.	✔	0.25	PCA is a deterministic algorithm: there is no initialization and there are no local optima.
Total		1.00 / 1.00	

Question 5

Which of the following are recommended applications of PCA? Select all that apply.

Your Answer		Score	Explanation
☐ Preventing overfitting: Reduce the number of features (in a supervised learning problem), so that there are fewer parameters to learn.	✔	0.25	You should use regularization to prevent overfitting, not PCA.
☐ As a replacement for (or alternative to) linear regression: For most learning applications, PCA and linear regression give substantially similar results.	✔	0.25	PCA is not linear regression. They have different goals (and cost functions), so they give different results.
☑ Data compression: Reduce the dimension of your input data $x^{(i)}$, which will be used in a supervised learning algorithm (i.e., use PCA so that your supervised learning algorithm runs faster).	✔	0.25	If your learning algorithm is too slow because the input dimension is too high, then using PCA to speed it up is a reasonable choice.
☑ Data visualization: Reduce data to 2D (or 3D) so that it can be plotted.	✔	0.25	This is a good use of PCA, as it can give you intuition about your data that would otherwise be impossible to see.

Total	1.00 / 1.00

Feedback — XV. Anomaly Detection

Help Center

You submitted this quiz on **Fri 10 Apr 2015 1:17 PM CEST**. You got a score of **5.00** out of **5.00**.

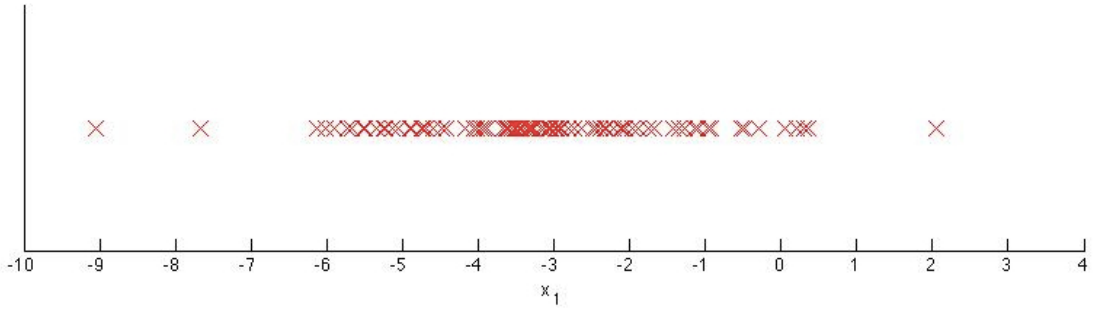
Question 1

For which of the following problems would anomaly detection be a suitable algorithm?

Your Answer		Score	Explanation
☐ Given data from credit card transactions, classify each transaction according to type of purchase (for example: food, transportation, clothing).	✔	0.25	Anomaly detection is not appropriate for a traditional classification problem.
☑ In a computer chip fabrication plant, identify microchips that might be defective.	✔	0.25	The defective chips are the anomalies you are looking for by modeling the properties of non-defective chips.
☐ Given an image of a face, determine whether or not it is the face of a particular famous individual.	✔	0.25	This problem is more suited to traditional supervised learning, as you want both famous and non-famous images in the training set.
☑ From a large set of primary care patient records, identify individuals who might have unusual health conditions.	✔	0.25	Since you are just looking for unusual conditions instead of a particular disease, this is a good appliation of anomaly detection.
Total		1.00 / 1.00	

Question 2

You have a 1-D dataset $\{x^{(1)}, \ldots, x^{(m)}\}$ and you want to detect outliers in the dataset. You first plot the dataset and it looks like this:



Suppose you fit the gaussian distribution parameters μ_1 and σ_1^2 to this dataset. Which of the following values for μ_1 and σ_1^2 might you get?

Your Answer		Score	Explanation
$\mu_1 = -3, \sigma_1^2 = 2$			
$\mu_1 = -3, \sigma_1^2 = 4$	✔	1.00	This is correct, as the data are centered around -3 and tail most of the points lie in [-5, -1].
$\mu_1 = -6, \sigma_1^2 = 2$			
$\mu_1 = -6, \sigma_1^2 = 4$			
Total		1.00 / 1.00	

Question 3

Suppose you have trained an anomaly detection system for fraud detection, and your system that flags anomalies when $p(x)$ is less than ε, and you find on the cross-validation set that it mis-flagging far too many good transactions as fradulent. What should you do?

Your Answer		Score	Explanation
Increase ε			
Decrease ε	✔	1.00	By decreasing ε, you will flag fewer anomalies, as desired.
Total		1.00 / 1.00	

Question 4

Suppose you are developing an anomaly detection system to catch manufacturing defects in airplane engines. You model uses $p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2)$. You have two features x_1 = vibration intensity, and x_2 = heat generated. Both x_1 and x_2 take on values between 0 and 1 (and are strictly greater than 0), and for most "normal" engines you expect that $x_1 \approx x_2$. One of the suspected anomalies is that a flawed engine may vibrate very intensely even without generating much heat (large x_1, small x_2), even though the particular values of x_1 and x_2 may not fall outside their typical ranges of values. What additional feature x_3 should you create to capture these types of anomalies:

Your Answer		Score	Explanation
○ $x_3 = x_1^2 \times x_2$			
○ $x_3 = x_1 \times x_2$			
○ $x_3 = x_1 + x_2$			
◉ $x_3 = \frac{x_1}{x_2}$	✔	1.00	This is correct, as it will take on large values for anomalous examples and smaller values for normal examples.
Total		1.00 / 1.00	

Question 5

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

Your Answer		Score	Explanation
☑ In anomaly detection, we fit a model $p(x)$ to a set of negative ($y = 0$) examples, without using any positive examples we may have collected of previously observed anomalies.	✔	0.25	We want to model "normal" examples, so we only use negative examples in training.
☐ In a typical anomaly	✔	0.25	It is the reverse: we have many normal

detection setting, we have a large number of anomalous examples, and a relatively small number of normal/non-anomalous examples.			examples and few anomalous examples.
☑ When developing an anomaly detection system, it is often useful to select an appropriate numerical performance metric to evaluate the effectiveness of the learning algorithm.	✔	0.25	You should have a good evaluation metric, so you can evaluate changes to the model such as new features.
☐ When evaluating an anomaly detection algorithm on the cross validation set (containing some positive and some negative examples), classification accuracy is usually a good evaluation metric to use.	✔	0.25	Classification accuracy is a poor metric because of the skewed classes in the cross-validation set (almost all examples are negative).
Total		1.00 / 1.00	

Feedback — XVI. Recommender Systems

Help Center

You submitted this quiz on **Sun 12 Apr 2015 4:48 PM CEST**. You got a score of **5.00** out of **5.00**.

Question 1

Suppose you run a bookstore, and have ratings (1 to 5 stars) of books. Your collaborative filtering algorithm has learned a parameter vector $\theta^{(j)}$ for user j, and a feature vector $x^{(i)}$ for each book. You would like to compute the "training error", meaning the average squared error of your system's predictions on all the ratings that you have gotten from your users. Which of these are correct ways of doing so (check all that apply)? For this problem, let m be the total number of ratings you have gotten from your users. (Another way of saying this is that $m = \sum_{i=1}^{n_m} \sum_{j=1}^{n_u} r(i,j)$). [Hint: Two of the four options below are correct.]

Your Answer		Score	Explanation
☐ $\frac{1}{m} \sum_{j=1}^{n_u} \sum_{i:r(i,j)=1} (\sum_{k=1}^{n} (\theta^{(k)})_j x_i^{(k)} - y^{(i,j)})^2$	✔	0.25	This incorrectly indexes into $\theta^{(j)}$ and $x^{(i)}$.
☐ $\frac{1}{m} \sum_{(i,j):r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - r(i,j))^2$	✔	0.25	This incorrectly used $r(i,j)$ as the actual rating.
☑ $\frac{1}{m} \sum_{(i,j):r(i,j)=1} (\sum_{k=1}^{n} (\theta^{(j)})_k x_k^{(i)} - y^{(i,j)})^2$	✔	0.25	This correctly sums over all ratings and computes the predicted rating with the explicit sum $\sum_{k=1}^{n} \theta^{(j)})_k x_k^{(i)}$.
☑ $\frac{1}{m} \sum_{(i,j):r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)})^2$	✔	0.25	This is correct, as it sums over all ratings the square difference between the predicted ratings $\theta^{(j)})^T x^{(i)}$ and the actual rating $y^{(i,j)}$.
Total		1.00 / 1.00	

Question 2

In which of the following situations will a collaborative filtering system be the most appropriate learning algorithm (compared to linear or logistic regression)?

Your Answer		Score	Explanation
☑ You manage an online bookstore and you have the book ratings from many users. For each user, you want to recommend other books she will enjoy, based on her own ratings and the ratings of other users.	✔	0.25	Collaborative filtering makes sense here, as you can use the ratings of all users to both learn features for books and recommend other books to each user.
☐ You manage an online bookstore and you have the book ratings from many users. You want to learn to predict the expected sales volume (number of books sold) as a function of the average rating of a book.	✔	0.25	This is a regression problem of predicting sales volume from ratings data, so collaborative filtering is not applicable.
☑ You run an online bookstore and collect the ratings of many users. You want to use this to identify what books are "similar" to each other (i.e., if one user likes a certain book, what are other books that she might also like?)	✔	0.25	You can find "similar" books by learning feature values using collaborative filtering.
☐ You're an artist and hand-paint portraits for your clients. Each client gets a different portrait (of themselves) and gives you 1-5 star rating feedback, and each client purchases at most 1 portrait. You'd like to predict what rating your next customer will give you.	✔	0.25	Since there is no overlap in the items reviewed by different clients, you cannot get good results using collaborative filtering.
Total		1.00 / 1.00	

Question 3

Suppose you have two matrices A and B, where A is 5x3 and B is 3x5. Their product is $C = AB$, a 5x5 matrix. Furthermore, you have a 5x5 matrix R where every entry is 0 or 1. You want to find the sum of all elements $C(i,j)$ for which the corresponding $R(i,j)$ is 1, and ignore all elements $C(i,j)$ where $R(i,j) = 0$. One way to do so is the following code:

```
C = A * B;
total = 0;
for i = 1:5
  for j = 1:5
    if (R(i,j) == 1)
      total = total + C(i,j);
    end
  end
end
```

Which of the following pieces of Octave code will also correctly compute this total? Check all that apply.

Your Answer		Score	Explanation
☐ C = (A * B) * R; total = sum(C(:));	✔	0.25	Multiplying `(A * B) * R` will perform regular matrix multiplication and won't "mask out" entries.
☐ total = sum(sum((A * B) * R));	✔	0.25	Multiplying `(A * B) * R` will perform regular matrix multiplication and won't "mask out" entries.
☑ C = A * B; total = sum(sum(C(R == 1)));	✔	0.25	This sums up all the elements in `C(R == 1)`, where the "logical indexing" expression selects only elements of C whose index matches an index in R for 1 elements.
☑ C = (A * B) .* R; total = sum(C(:));	✔	0.25	This sums up all elements of `(A * B) .* R`, where the .* operator performs element-wise multiplication, setting the elements of `A * B` to zero that correspond to zero entries in `R`.
Total		1.00 / 1.00	

Question 4

You run a movie empire, and want to build a movie recommendation system based on collaborative filtering. There were three popular review websites (which we'll call A, B and C) which users to go to rate movies, and you have just acquired all three companies that run these websites. You'd like to merge the three companies' datasets together to build a single/unified system. On website A, users rank a movie as having 1 through 5 stars. On website B, users rank on a scale of 1 - 10, and decimal values (e.g., 7.5) are allowed. On website C, the ratings are from 1 to 100. You also have enough information to identify users/movies on one website with users/movies on a different website. Which of the following statements is true?

Your Answer		Score	Explanation
○ It is not possible to combine these websites' data. You must build three separate recommendation systems.			
○ You can combine all three training sets into one as long as your perform mean normalization and feature scaling **after** you merge the data.			
◉ You can merge the three datasets into one, but you should first normalize each dataset's ratings (say rescale each dataset's ratings to a 0-1 range).	✔	1.00	By normalizing each dataset, you ensure that all ratings are on the same scale, so they are comparable during training.
○ Assuming that there is at least one movie/user in one database that doesn't also appear in a second database, there is no sound way to merge the datasets, because of the missing data.			
Total		1.00 / 1.00	

Question 5

Which of the following are true of collaborative filtering systems? Check all that apply.

Your Answer		Score	Explanation
☑ If you have a dataset of user ratings on some products, you can uses these to predict one user's preferences on products he has not rated.	✔	0.25	This is exactly the job of the collaborative filtering algorithm.
☐ When using gradient descent to train a collaborative filtering system, it is okay to initialize all the parameters ($x^{(i)}$ and $\theta^{(j)}$) to zero.	✔	0.25	You need to initialize them to different values so that you learn different features and parameters (i.e., perform symmetry breaking).
☑ For collaborative filtering, it is possible to use one of the advanced optimization algoirthms (L-BFGS/conjugate gradient/etc.) to solve for both the $x^{(i)}$'s and $\theta^{(j)}$'s simultaneously.	✔	0.25	You can compute the cost function and gradient, so any of these algorithms will work fine.
☐ For collaborative filtering, the optimization algorithm you should use is gradient descent. In particular, you cannot use more advanced optimization algorithms (L-BFGS/conjugate gradient/etc.) for collaborative filtering, since you have to solve for both the $x^{(i)}$'s and $\theta^{(j)}$'s simultaneously.	✔	0.25	You can compute the cost function and gradient, so any of the advanced optimization algorithms will also work.
Total		1.00 / 1.00	

Feedback — XVII. Large Scale Machine Learning

Help Center

You submitted this quiz on **Sun 12 Apr 2015 11:00 PM CEST**. You got a score of **5.00** out of **5.00**.

Question 1

Suppose you are training a logistic regression classifier using stochastic gradient descent. You find that the cost (say, $cost(\theta, (x^{(i)}, y^{(i)}))$, averaged over the last 500 examples), plotted as a function of the number of iterations, is slowly increasing over time. Which of the following changes are likely to help?

Your Answer		Score	Explanation
Try averaging the cost over a smaller number of examples (say 250 examples instead of 500) in the plot.			
Try averaging the cost over a larger number of examples (say 1000 examples instead of 500) in the plot.			
Try halving (decreasing) the learning rate α, and see if that causes the cost to now consistently go down; and if not, keep halving it until it does.	✔	1.00	Such a plot indicates that the algorithm is diverging. Decreasing the learning rate α means that each iteration of stochastic gradient descent will take a smaller step, thus it will likely converge instead of diverging.
Use fewer examples from your training set.			
Total		1.00 / 1.00	

Question 2

Which of the following statements about stochastic gradient descent are true? Check all that apply.

Your Answer		Score	Explanation
☑ If you have a huge training set, then stochastic gradient descent may be much faster than batch gradient descent.	✔	0.25	Because stochastic gradient descent can make progress after only a few examples, it can converge much more quickly than batch gradient descent.
☐ One of the advantages of stochastic gradient descent is that it uses parallelization and thus runs much faster than batch gradient descent.	✔	0.25	Stochastic gradient descent still runs in series, one example at a time.
☑ One of the advantages of stochastic gradient descent is that it can start progress in improving the parameters θ after looking at just a single training example; in contrast, batch gradient descent needs to take a pass over the entire training set before it starts to make progress in improving the parameters' values.	✔	0.25	This is true, since stochastic gradient descent updates the parameters for every training example, but batch gradient descent updates them based on an average over the entire training set.
☐ In order to make sure stochastic gradient descent is converging, we typically compute $J_{\text{train}}(\theta)$ after each iteration (and plot it) in order to make sure that the cost function is	✔	0.25	We want to plot $cost(\theta, (x^{(i)}, y^{(i)}))$ at each iteration, as computing the full summation $J_{\text{train}}(\theta)$ is too expensive.

generally decreasing.		
Total	1.00 / 1.00	

Question 3

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

Your Answer		Score	Explanation
☑ Online learning algorithms are usually best suited to problems were we have a continuous/non-stop stream of data that we want to learn from.	✔	0.25	Such a stream of data is well-suited to online learning because online learning does not save old training examples, but instead uses them once and then throws them out.
☐ One of the advantages of online learning is that there is no need to pick a learning rate α.	✔	0.25	One still must choose a learning rate to use online learning.
☐ One of the disadvantages of online learning is that it requires a large amount of computer memory/disk space to store all the training examples we have seen.	✔	0.25	Since online learning algorithms do not save old examples, they can be very efficent in terms of computer memory and disk space.
☑ In the approach to online learning discussed in the lecture video, we repeatedly get a single training example, take one step of stochastic gradient descent using that example, and then move on to the next example.	✔	0.25	This is one good approach to online learning discussed in the lecture video.
Total		1.00 / 1.00	

Question 4

Assuming that you have a very large training set, which of the following algorithms do you think can be parallelized using map-reduce and splitting the training set across different machines? Check all that apply.

Your Answer		Score	Explanation
☐ Linear regression trained using stochastic gradient descent.	✔	0.25	Since stochastic gradient descent processes one example at a time and updates the parameter values after each, it cannot be easily parallelized.
☑ Computing the average of all the features in your training set $\mu = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}$ (say in order to perform mean normalization).	✔	0.25	You can split the dataset into N smaller batches, compute the feature average of each smaller batch on one of N separate computers, and then average those results on a central computer to get the final result.
☐ A neural network trained using stochastic gradient descent.	✔	0.25	Since stochastic gradient descent processes one example at a time and updates the parameter values after each, it cannot be easily parallelized.
☑ A neural network trained using batch gradient descent.	✔	0.25	You can split the dataset into N smaller batches, compute the gradient for each smaller batch on one of N separate computers, and then average those gradients on a central computer to use for the gradient update.
Total		1.00 / 1.00	

Question 5

Which of the following statements about map-reduce are true? Check all that apply.

Your Answer		Score	Explanation
☑ Because of network latency and other overhead	✔	0.25	The maximum speedup possible is N-fold, and it is unlikely you will get an N-fold speedup because of

associated with map-reduce, if we run map-reduce using N computers, we might get less than an N-fold speedup compared to using 1 computer.			the overhead.
☑ When using map-reduce with gradient descent, we usually use a single machine that accumulates the gradients from each of the map-reduce machines, in order to compute the parameter update for that iteration.	✔	0.25	Such a setup allows us to use many computers to do the hard work of gradient computation while making the parameter update simple, as it occurs in one place.
☑ If you are have just 1 computer, but your computer has multiple CPUs or multiple cores, then map-reduce might be a viable way to parallelize your learning algorithm.	✔	0.25	Treating each core as a separate computer makes map-reduce just as useful with multiple cores as with multiple computers.
☐ If we run map-reduce using N computers, then we will always get at least an N-fold speedup compared to using 1 computer.	✔	0.25	The maximum speedup possible is N-fold, and it is unlikely you will get an N-fold speedup because of the overhead.
Total		1.00 / 1.00	

Feedback — XVIII. Application: Photo OCR

Help Center

You submitted this quiz on **Mon 13 Apr 2015 10:00 PM CEST**. You got a score of **5.00** out of **5.00**.

Question 1

Suppose you are running a sliding window detector to find text in images. Your input images are 1000x1000 pixels. You will run your sliding windows detector at two scales, 10x10 and 20x20 (i.e., you will run your classifier on lots of 10x10 patches to decide if they contain text or not; and also on lots of 20x20 patches), and you will "step" your detector by 2 pixels each time. About how many times will you end up running your classifier on a single 1000x1000 test set image?

Your Answer		Score	Explanation
(•) 500,000	✔	1.00	With a stride of 2, you will run your classifier approximately 500 times for each dimension. Since you run the classifier twice (at two scales), you will run it 2 * 500 * 500 = 500,000 times.
( ) 1,000,000			
( ) 100,000			
( ) 250,000			
Total		1.00 / 1.00	

Question 2

Suppose that you just joined a product team that has been developing a machine learning application, using $m = 1,000$ training examples. You discover that you have the option of hiring additional personnel to help collect and label data. You estimate that you would have to pay each of the labellers $10 per hour, and that each labeller can label 4 examples per minute. About how

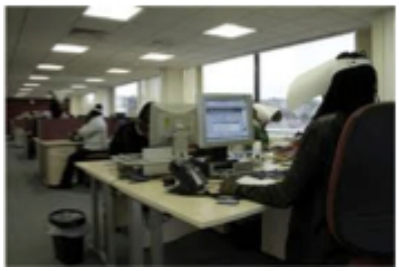
much will it cost to hire labellers to label 10,000 new training examples?

Your Answer		Score	Explanation
$600			
$10,000			
$250			
$400	✔	1.00	On labeller can label $4 \times 60 = 240$ examples in one hour. It will thus take him $10,000/240 \approx 40$ hours to complete 10,000 examples. At $10 an hour, this is $400.
Total		1.00 / 1.00	

Question 3

Suppose you are building an object classifier, that takes as input an image, and recognizes that image as either containing a car ($y = 1$) or not ($y = 0$). For example, here are a positive example and a negative example:

Positive example ($y = 1$)



Negative example ($y = 0$)

After carefully analyzing the performance of your algorithm, you conclude that you need more

positive ($y = 1$) training examples. Which of the following might be a good way to get additional positive examples?

Your Answer		Score	Explanation
Make two copies of each image in the training set; this immediately doubles your training set size.			
Take a training example and set a random subset of its pixel to 0 to generate a new example.			
Apply translations, distortions, and rotations to the images already in your training set.	✔	1.00	These geometric distortions are likely to occur in real-world images, so they are a good way to generate additional data.
Select two car images and average them to make a third example.			
Total		1.00 / 1.00	

Question 4

Suppose you have a PhotoOCR system, where you have the following pipeline:

You have decided to perform a ceiling analysis on this system, and find the following:

Component	Accuracy
Overall System	70%
Text Detection	72%
Character Segmentation	82%
Character Recognition	100%

Which of the following statements are true?

Your Answer	Score	Explanation

☐ The least promising component to work on is the character recognition system, since it is already obtaining 100% accuracy.	✔	0.25	The character recognition component is the most promising, as ground truth character recognition improves performance by 18% over feeding the current character recognition system ground truth character segmentation.
☐ We should dedicate significant effort to collecting additional training data for the text detection system.	✔	0.25	A perfect text detection system improves overall performance by only 2%, so collecting additional data for that system is not a good investment of time.
☑ If we conclude that the character recognition's errors are mostly due to the character recognition system having high variance, then it may be worth significant effort obtaining additional training data for character recognition.	✔	0.25	Since the biggest improvement comes from character recognition ground truth, we would like to improve the performance of that system. It the character recognition system has high variance, additional data will improve its performance.
☑ Performing the ceiling analysis shown here requires that we have ground-truth labels for the text detection, character segmentation and the character recognition systems.	✔	0.25	At each step, we provide the system with the ground-truth output of the previous step in the pipeline. This requires ground truth for every step of the pipeline.
Total		1.00 / 1.00	

Question 5

What are the benefits of performing a ceiling analysis? Check all that apply.

Your Answer		Score	Explanation
☑ It helps us decide on allocation of resources in terms of which component in a machine learning pipeline to spend more effort on.	✔	0.25	The ceiling analysis reveals which parts of the pipeline have the most room to improve the performance of the overall system.
☐ A ceiling analysis helps us to decide what is the most promising learning algorithm (e.g., logistic regression vs. a neural network vs. an SVM) to apply to a specific component of a machine learning pipeline.	✔	0.25	A ceiling analysis works with different components of a pipeline under a fixed algorithm setup.
☐ It is a way of providing additional training data to the algorithm.	✔	0.25	Ceiling analysis works with the data already present.
☑ It can help indicate that certain components of a system might not be worth a significant amount of work improving, because even if it had perfect performance its impact on the overall system may be small.	✔	0.25	An unpromising component will have little effect on overall performance when it is replaced with ground truth.
Total		1.00 / 1.00	