CS584: Machine Learning - Mid-term (15 points)

Name: **Illinois Tech A#:** **Score:**

This exam is open book. You may bring in your homework, class notes and textbooks to help you. Laptops are NOT allowed. “Check all that apply” in the questions means one or more answers are possibly correct.

Question 1 (0.5 Points)

A computer program is said to learn from experience E with respect to some task T and some performance measure P if its performance on T, as measured by P, improves with experience E. Suppose we feed a learning algorithm a lot of historical weather data, and have it learn to predict weather. In this setting, what is T, E and P?

Question 2 (1 Points)

Let f be some function so that f(θ_0,θ_1) outputs a number. For this problem, f is some arbitrary/unknown smooth function (not necessarily the cost function of linear regression, so f may have local optima). Suppose we use gradient descent to try to minimize f(θ_0,θ_1) as a function of θ_0 and θ_1. Which of the following statements are true? (Check all that apply.)

- (A) If θ_0 and θ_1 are initialized so that θ_0=θ_1, then by symmetry (because we do simultaneous updates to the two parameters), after one iteration of gradient descent, we will still have θ_0=θ_1.
- (B) Setting the learning rate α to be very small is not harmful, and can only speed up the convergence of gradient descent.
- (C) If the first few iterations of gradient descent cause f(θ_0,θ_1) to increase rather than decrease, then the most likely cause is that we have set the learning rate α to a large value.
- (D) If the learning rate is too small, then gradient descent may take a very long time to converge.

Question 3 (1 Points)

Suppose that for some linear regression problem (say, predicting housing prices), we have some training set, and for our training set we managed to find some θ_0, θ_1 such that the loss function J(θ_0,θ_1)=0. Which of the statements below must then be true? (Check all that apply.)

- (A) For this to be true, we must have $\theta_0 = 0$ and $\theta_1 = 0$ so that $h_\theta(x) = 0$
- (B) Our training set can be fit perfectly by a straight line, i.e., all of our training examples lie perfectly on some straight line.
- (C) For this to be true, we must have y(i)=0 for every value of i=1,2,. . . ,m.

- (D) 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.)

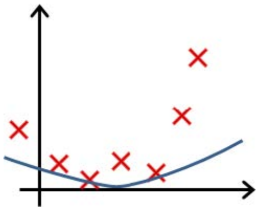
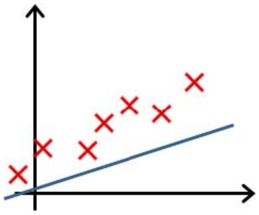
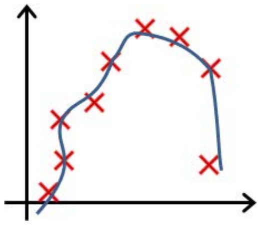
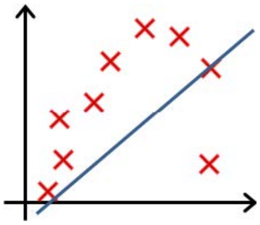
Question 4 (1 Points)

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

- (A) Adding many new features to the model helps prevent overfitting on the training set.
- (B) Introducing regularization to the model always results in equal or better performance on examples not in the training set.
- (C) Introducing regularization to the model always results in equal or better performance on the training set.
- (D) Adding many new features to the model makes it more likely to overfit the training set.

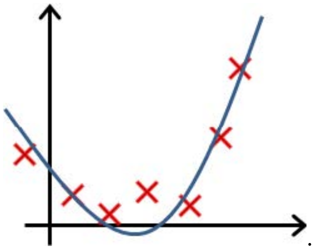
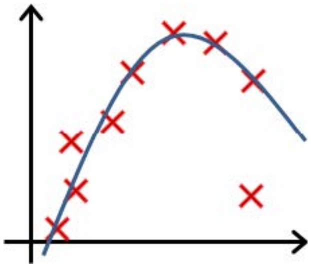
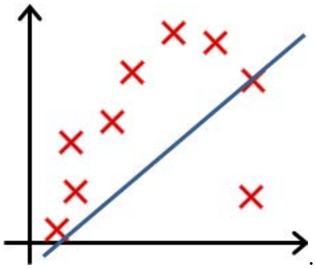
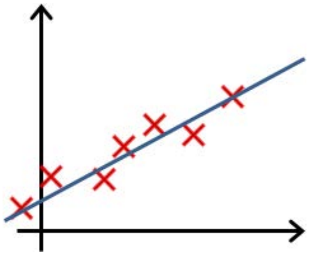
Question 5 (1 Points)

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

- (A) . (B) 
- (C) . (D) 

Question 6 (1 Points)

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

- (A)  . (B) 
- (C)  . (D) 

Question 7 (0.5 Points)

Please use a graph to describe one neural unit with 4 inputs, 1 bias term and 1 output. The activation function is Sigmod function.

Question 8 (1 Points)

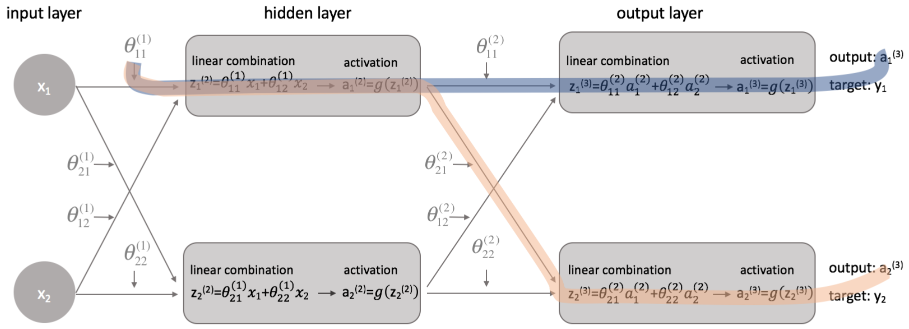
If the inputs are 1,2,-1,3 respectively, the bias is 1, all weights are 1, what is the output value for the neural unit in Question 7?

Question 9 (1 Points)

Please draw a 3-layer fully-connected neural network and describe how this neural network used for multi-class classification tasks.

Question 10 (3 Points)

What is the chain rule? Assuming the lost function $J(\theta)$ is the least squared loss. Please calculate the total expression for $\partial J(\theta)/\partial\theta_{11}^{(1)}$.


input layer
hidden layer
output layer
x_1
x_2
$\theta_{11}^{(1)}$
$\theta_{21}^{(1)}$
$\theta_{12}^{(1)}$
$\theta_{22}^{(1)}$
linear combination
activation
$z_1^{(2)}=\theta_{11}^{(1)}x_1+\theta_{12}^{(1)}x_2 \longrightarrow a_1^{(2)}=g(z_1^{(2)})$
linear combination
activation
$z_2^{(2)}=\theta_{21}^{(1)}x_1+\theta_{22}^{(1)}x_2 \longrightarrow a_2^{(2)}=g(z_2^{(2)})$
$\theta_{11}^{(2)}$
$\theta_{21}^{(2)}$
$\theta_{12}^{(2)}$
$\theta_{22}^{(2)}$
linear combination
activation
$z_1^{(3)}=\theta_{11}^{(2)}a_1^{(2)}+\theta_{12}^{(2)}a_2^{(2)} \longrightarrow a_1^{(3)}=g(z_1^{(3)})$
linear combination
activation
$z_2^{(3)}=\theta_{21}^{(2)}a_1^{(2)}+\theta_{22}^{(2)}a_2^{(2)} \longrightarrow a_2^{(3)}=g(z_2^{(3)})$
output: $a_1^{(3)}$
target: y_1
output: $a_2^{(3)}$
target: y_2

Question 11 (1 Points)

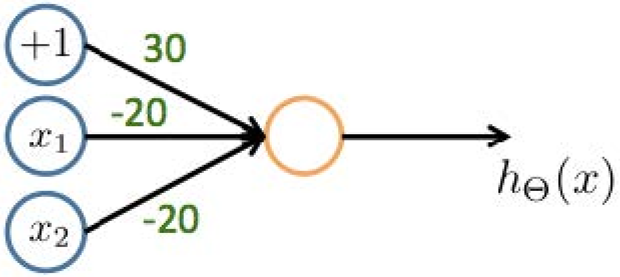
Please describe why convolutional neural network is suitable for image. Moreover, please describe how is convolution operator used for the image.

Question 12 (1 Points)

Please draw and describe a typical pipeline of convolutional neural network for image classification.

Question 13 (2 Points)

Consider the following neural network which takes two binary-valued inputs x_1,$x_2 \in \{0, 1\}$ and outputs $h_\Theta(x)$. The activation function is the sigmod function. What's the logical functions (AND, OR, XOR, NAND, *etc.*) does this neural network (approximately) compute and why?


+1
x_1
x_2
30
-20
-20
$h_\Theta(x)$