CSCE 633: Machine Learning

Sample EXAM # 2

Total Time: 75 minutes

Name: _____

UID: _____

Question	*Point*	*Grade*
1	*40*	
2	*30*	
3	*30*	
Total	*100*	

People Sitting to Your Left:

Person Sitting to Your Right:

1) (40 points) Concepts.

For each of the following questions, please provide your answer and an explanation/justification. Please answer the following questions

a) (10 points) Please describe the loss function used for autoencoders.

b) (10 points) Why does GMM provide a potential for soft clustering where K-means is a strict clustering? What small modification could you make to K-means to make it a soft clustering?

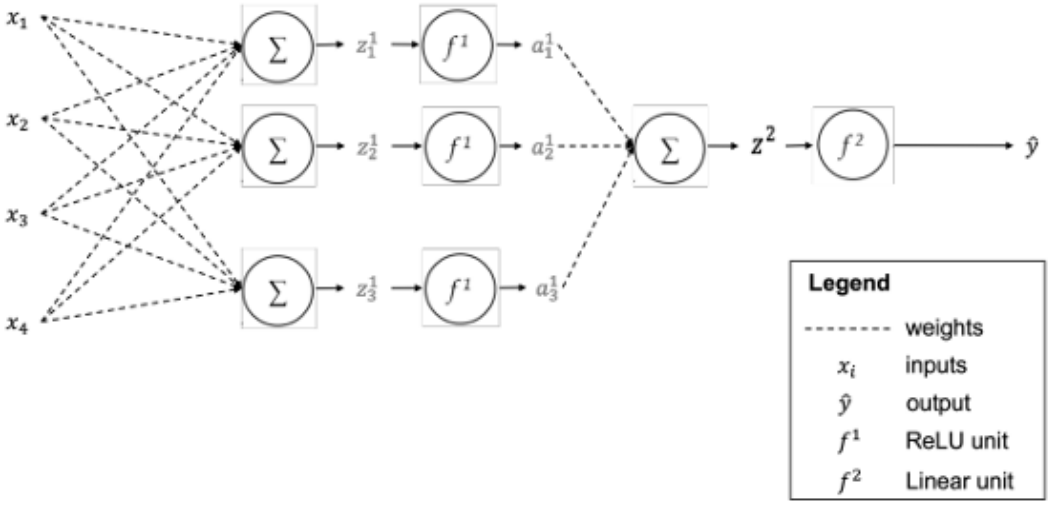
c) (10 points) There are a series of optimizers used for training Neural Networks (e.g. RMSProp, Adam). What (if anything) are the similarities or differences between these optimization techniques and Gradient Descent?

d) (10 points) What advantages do LSTMs have over RNNs?

This Page Intentionally Left Blank

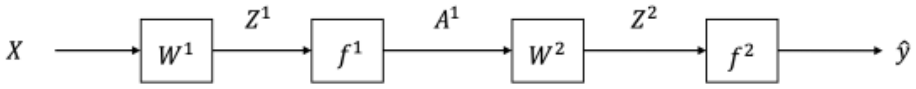
2) (30 points) Neural Networks

A feed-forward neural network is shown below.



- The dashed lines represent the weights that the neural network learns
- There are no bias/constant terms among the weights
- We are using squared-loss, i.e., $L(y, \hat{y}) = (y - \hat{y})^2$
- Inputs are represented by $X = [x_1, x_2, x_3, x_4]^T$
- The true output value is denoted by y, and the estimation output by the network is $\hat{y}$
- f^1 is ReLU unit
- f^2 is a linear unit

The neural network shown above can also be represented by the network shown below, which uses matrix notation. Specifically, the input X is a 4 × 1 column vector, $\hat{y}$ is a 1×1 scalar. W^2 is a 3×1 vector. We also know that, $Z^1 = (W^1)^T X$ and $Z^2 = (W^2)^T A^1$



Important: for each question, provide you justification/reasoning.

a. (4 points) What are the dimensions of the matrix W^1?

b. (4 points) What are the dimensions of the matrix Z^2?

For both parts below, you are told that there is only one data point which is: X =$[1,1,1,1]^T$ and y = [1].

c. (5 points) If W^1 and W^2 are both matrices/vectors of all ones, what is the resulting Loss?

d. (5 points) If W^1 is a matrix of all −1's (all negative ones) and W^2 is a vector of all 1's (positive ones), what is the resulting Loss?

This Page Intentionally Left Blank

We now use back-propagation to update the weights during each iteration. For all questions below, assume that we only have one data point (X, y) available to use, and the step-size parameter is 0.01. You are asked to determine how many components of W^1 will get updated (i.e., have their value changed) in each scenario below.

e. (6 points) Assume X =$[1,1,1,1]^T$, y = [1]. Further assume that we start with W^1as a matrix of −1's (negative ones) while W^2 is a vector of 1's (positive ones). How many components of W^1 will get updated (i.e., have their value changed) after one iteration of backprop? Explain your answer.

f. (6 points) Assume X =$[0,0,0,0]^T$, y = [0]. Further assume that we start off with $W^1$and W^2 as matrices/vectors of all ones. How many components of W^1 will get updated (i.e., have their value changed) after one iteration of back-propagation? Explain your answer.

3) (30 Points) Support Vector Machines

Assume we have the following Lagrangian-Optimized Loss function for support vector machines:

$$L = \frac{1}{2}\|\boldsymbol{w}\|_2^2 - C\sum_{i=1}^{N}\epsilon_i - \sum_{i=1}^{N}\alpha_i(y_i(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i) + w_0) - 1 + \epsilon_i) - \sum_{i=1}^{N}\mu_i\epsilon_i$$

a) (5 points) Is this the Maximal Marginal Classifier, the Support Vector Classifier, or the Support Vector Machine? Please list the formula elements that allow you to make this determination

b) (10 points) Please re-write this in the original constraint-optimization form.

This Page Intentionally Left Blank

c) (15 points) Please explain if the following statements are True or False AND provide you justification/reasoning

c.i) Data that is linearly separable does not need the use of slack variables in SVM.

T / F
Reason:

c.ii) SVMs are likely better at handling outlier data than logistic regression due to the use of slack variables.

T / F
Reason:

c.iii) Increasing the number of support vectors always leads to an increase in performance of an SVM classifier.

T / F
Reason: