CSCE 633: Machine Learning

Sample EXAM # 2

Total Time: 75 minutes

Name: Solution

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.

$$MSE(y, \hat{y})$$

Sample response: Autoencoders use the above loss function, with y denoting label and $\hat{y}$ denotes reconstructed prediction, and the loss calculates and minimizes reconstruction error

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?

Sample response: GMM provide a potential because it is probabilistic: you have p(z | x) for all clusters so you can view any assignment as a vector of probabilities. To enable soft clustering, you could create a vector of distances for K-means then normalize these to certain "probability/SoftMax"

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?

Sample response: they are similar to gradient descent in that - they are based on gradient descent and improving weights by minimizing error/loss.

They are different because they add concept of momentum - and modify a bit how they work specifically to help get out of saddle points that exist in very high dimensions.

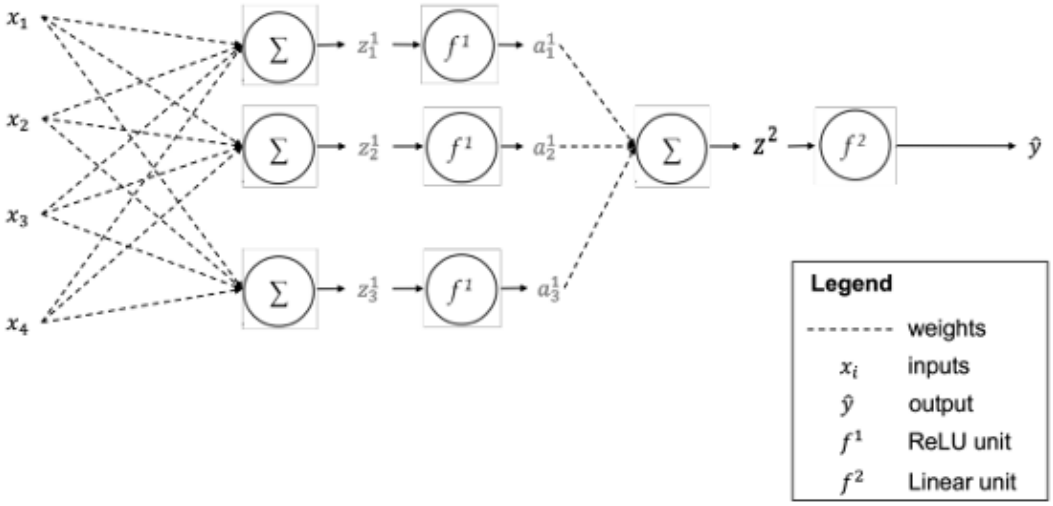
d) (10 points) What advantages do LSTMs have over RNNs?

Sample response: LSTM's memory component allows for finding repeated short sequent patterns repeated over long time sequences

This Page Intentionally Left Blank

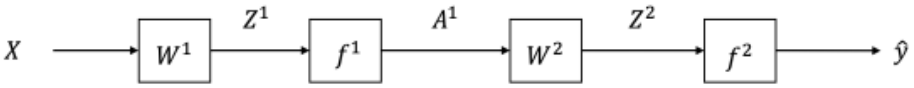
2) (30 points) Neural Networks

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 $Z^2 = (W^2)^T A^1$



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].

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

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

4 * 3

X is a 4 * 1 matrix and Z^1 is a 3-row matrix, $Z^1 = (W^1)^T X$, so W^1 must be 4 * 3.

Explain your answer or list steps

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

1 * 1
W[2] is a 3 * 1 vector, A[1] is a 3 * 1 vector, and $Z^2 = (W^2)^T A^1$, so Z[2] is 1 * 1.
Explain your answer or list steps

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

(12- 1) ^ 2 = 121

$$W^1 = \begin{matrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{matrix} \text{ and } W^2 = \begin{matrix} 1 \\ 1, \\ 1 \end{matrix}$$

$$\text{So } Z^1 = A^1 = \begin{matrix} 4 \\ 4 \\ 4 \end{matrix} \text{ and } Z^2 = (1 \quad 1 \quad 1) \times \begin{pmatrix} 4 \\ 4 \\ 4 \end{pmatrix} = 12.$$

Therefore, loss is $(y - \hat{y})^2 = 121$.

Explain your answer or list steps

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?

(0 - 1) ^ 2 = 1

$$W^1 = \begin{matrix} -1 & -1 & -1 \\ -1 & -1 & -1 \\ -1 & -1 & -1 \\ -1 & -1 & -1 \end{matrix} \text{ and } W^2 = \begin{matrix} 1 \\ 1, \\ 1 \end{matrix}$$

$$\text{So } Z^1 = \begin{matrix} -4 \\ -4. \\ -4 \end{matrix}$$

The negative ones in W[1] will ensure that the input to the ReLU unit (f[l]) is negative. This leads to A[1] = 0 thus Z[2] = ^y = 0.

This Page Intentionally Left Blank

We now use backpropagation 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.

Sample response: 0 components will be updated., The negative ones in W1 will lead to the outputs of the ReLU unit being zero which leads to $\partial A^1 \, \partial Z^1 = 0$, and that will zero out the gradient, $\partial L \, \partial W^1$.

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.

Sample response: 0 components will be updated.
X = 0 will zero out the gradient!

3) (30 Points) Support Vector Machines

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

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

(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

Sample response: this is support vector machine;
$-C\sum_{i=1}^{N}\epsilon_i$ is the slack variable; SVM-> $y_i(w^T\phi(x_i) + w_0) + 1 + \epsilon_i$

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

$$\min_{w}\{\frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{N}\epsilon_i\}$$

such that $y_i(w^T\phi(\mathbf{x_i}) + w_0) \geq 1 - \epsilon_i$ and $\epsilon_i > 0$ for i=1,2,⋯,N

This Page Intentionally Left Blank

For each - 3 points for T/F - 7 points for correct reasoning

a) (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:

Sample response: TRUE because you can just use the maximum marginal classifier

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

T / F

Reason:

Sample response: TRUE - more generalizable; slack variables contribute to the SVM being tolerant to errors. Thus, slack variables can contribute to the decision boundary not deviating severely due to the outliers.

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

T / F

Reason:

Sample response FALSE, increasing the number of support vectors does not necessarily increase the performance of the classifier. A very large number of support vectors can be a sign of overfitting.