Name:

USC ID:

Notes:

- Write your name and ID number in the spaces above.
- No books, cell phones or other notes are permitted. Only one letter size cheat sheet (back and front) and a calculator are allowed.
- Problems are not sorted in terms of difficulty. Please avoid guess work and long and irrelevant answers.
- Show all your work and your final answer. Simplify your answer as much as you can.
- Open your exam only when you are instructed to do so.

Problem	Score	Earned
1	25	
2	20	
3	20	
4	20	
5	25	
Total	110	

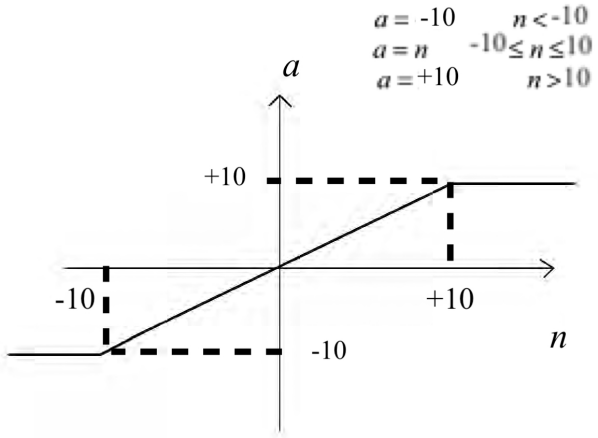
1. The purpose of this question is to design a Convolutional Neural Network to classify the word "REZA" encoded as class $C_1 = [1\ 0]^T$ using one-hot encoding from the word "JACK" encoded as $C_2 = [0\ 1]^T$. Here, the convolution operator acts on "letters" instead of pixels. Letter are encoded using the following table:

Conversion Table

A = 1	K = 11	U = 21
B = 2	L = 12	V = 22
C = 3	M = 13	W = 23
D = 4	N = 14	X = 24
E = 5	O = 15	Y = 25
F = 6	P = 16	Z = 26
G = 7	Q = 17	
H = 8	R = 18	
I = 9	S = 19	
J = 10	T = 20	

Each word is represented as a *row vector*.

Only one feature map is created using the Kernel $[-1\ 2\ -1]$ with stride 1. The resulting feature map is then passed through the saturating linear activation function described in the following figure:



Note that the mathematical formula for the saturated linear function $\mathbf{f}^{(1)}$ is given in the figure. A maxpooling operator is then applied to the *whole* feature map that is output of the saturating linear function. For example, if the output of the saturating linear function is $\mathbf{a}^{(1)} = [1\ 10\ -10\ 5]^T$, then $a^{(2)} = \text{maxpooling}(\mathbf{a}^{(1)}) = 10$. The output of the maxpooling is treated as the input to a Feedforward Neural Network with one layer whose weight matrix is $\mathbf{W}^{(3)}$. For simplicity, we assume that the network does not have bias. This one layer neural network has its own activation function $\mathbf{f}^{(3)}$.

(a) Draw a block diagram of this network.

(b) Determine $\mathbf{W}^{(3)}$ and $\mathbf{f}^{(3)}$ such that REZA is classified as $[1\ 0]^T$ and JACK is classified as $[0\ 1]^T$ and show all the calculations that are needed to determine the output of the network for each of these two words.

2. A company with headquarters in the Bay Area has two offices in Los Angeles and San Diego. An employee in San Diego office is sent to the Los Angeles office the next day with probability 0.35 and stays in San Diego office with probability 0.65. An employee in Los Angeles office is sent to the San Diego office with probability 0.8 and stays in Los Angeles office with probability 0.2. A new employee is assigned to Los Angeles office with probability 0.4 and to San Diego office with probability 0.6. An employee in San Diego office works between six and eight hours per day with probability 0.7, works more than eight hours with probability 0.2, and works less than six hours per day with probability 0.1. An employee in Los Angeles office works between six and eight hours per day with probability 0.15, works more than eight hours with probability 0.25, and works less than six hours per day with probability 0.6. A manager in the headquarters can only observe the number of hours each employee worked each day.

   (a) Construct a Hidden Markov Model that models the observations of the manager in their headquarters. Clearly show the parameters with matrices and vectors and draw a state transition graph for the model.

   (b) If the manager observes the number of hours a new employee worked in the first three consecutive days of work to be $6.5, 10, 7$, what is the most likely sequence of places at which the employee worked in those three days?

   (c) What sequence of three places has the maximum expected number of correct places?

Solution:

3. Consider the following data set: In class 1, we have $[0\ 0]^T, [0\ 1]^T, [1\ 1]^T$. In class 2, we have $[0.5\ 0.5]^T$.

   (a) Sketch the data set and determine whether or not it is linearly separable.

   (b) Regardless of the answer to 3a, find a quadratic feature $X_3 = f(X_1, X_2) = aX_1^2 + bX_2^2 + cX_1X_2 + d$, that makes the data linearly separable; that is, $X_3 \geq 0$ for members of class 1, and $X_3 < 0$ for members of class 2. Find the maximum margin classifier only based on X_3. Hint: The equation of the maximum margin classifier based on only one feature is $X_3 = \beta_0$ and you should determine β_0.

   (c) By solving $X_3 = f(X_1, X_2) = \beta_0$ for X_2, find the equation of the decision boundary in the original feature space and sketch it. Show the regions in the feature space that are classified as class 1 and class 2. You do not need to be very precise.

4. Suppose that for a particular data set, we perform hierarchical clustering using single linkage (minimal intercluster dissimilarity) and using complete linkage (maximal intercluster dissimilarity). We obtain two dendrograms.

   (a) At a certain point on the single linkage dendrogram, the clusters $\{1, 2, 3\}$ and $\{4, 5\}$ fuse. On the complete linkage dendrogram, the clusters $\{1, 2, 3\}$ and $\{4, 5\}$ also fuse at a certain point. Which fusion will occur higher on the tree, or will they fuse at the same height, or is there not enough information to tell?

   (b) At a certain point on the single linkage dendrogram, the clusters $\{5\}$ and $\{6\}$ fuse. On the complete linkage dendrogram, the clusters $\{5\}$ and $\{6\}$ also fuse at a certain point. Which fusion will occur higher on the tree, or will they fuse at the same height, or is there not enough information to tell?

5. Choose either T (True) or F (False):

   (a) One can design a classifier for the XOR problem using a MLP with linear activation functions in the hidden layer and sigmoids in the output layer. T F

   (b) $\mathscr{L}_2$ regularization in the back-propagation+Stochastic Gradient Descent training of MLPs is equivalent to adding a forgetting factor to the weight update equation. T F

   (c) When any linear binary classifier results in classification close to random guessing, an RBF Kernel is the best kernel of choice to expand the feature space for a Support Vector Machine. T F

   (d) In Co-training, we use a labler or "Oracle" along with a classifier in a collaborative manner to train the classifier. T F

   (e) The function of hidden layers of MLPs is equivalent to the function of Kernels in SVMs. T F

   (f) Convolutional Neural Networks act as feature extractors from images. T F

   (g) The responses of support vector classifiers and unreqularized logistic regression trained on the same data set are very similar because of having very similar loss functions. T F

   (h) We cannot encode each binary label of a multi-label problem into an output of a MLP, because that architecture is reserved for multi-class classification. T F

   (i) The Naïve Bayes' classifier cannot yield decision boundaries that are the same as those given by a support vector classifier. T F

   (j) To find K in the K-means algorithm, we can penalize the objective function WCV (Within Cluster Variation) using AIC or BIC, in the same way we use AIC or BIC in model selection for linear regression. T F

Scratch paper

Name:

USC ID:

Scratch paper

Name:

USC ID: