Name:

USC ID:

Notes:

- Write your name and ID number in the spaces above.
- No cell phone, no books or other notes are permitted. Only two letter size cheat sheets (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.

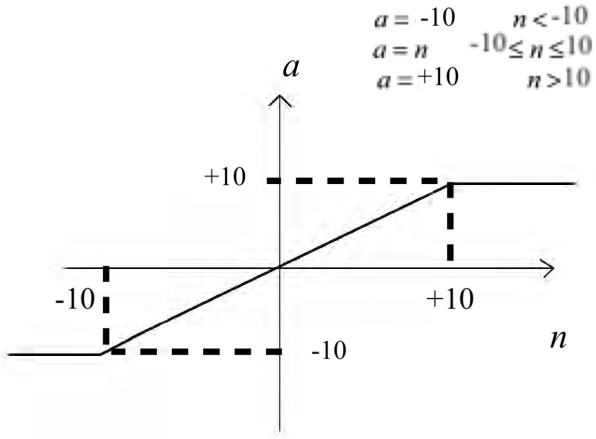
Problem	Score	Earned
1	20	
2	20	
3	20	
4	20	
5	20	
5	20	
Total	120	

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: Each word is represented as a *row vector*.

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		

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.

Solution:

2. 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 2a, find a quadratic feature $X_3 = f(X_1, X_2)$, that makes the data linearly separable. 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.

3. 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

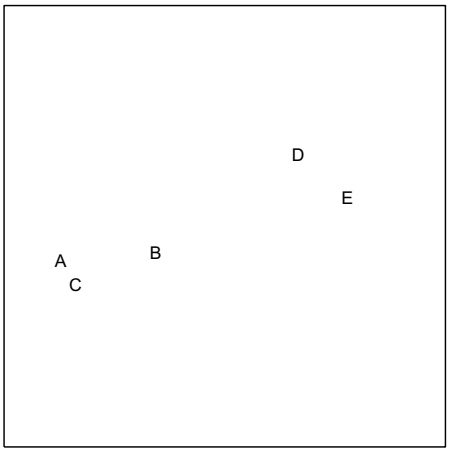
   (g) The responses of support vector classifiers and unreqularized logistic regression trained on the same data set are very similar because of having very similar loss functions. T F

   (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

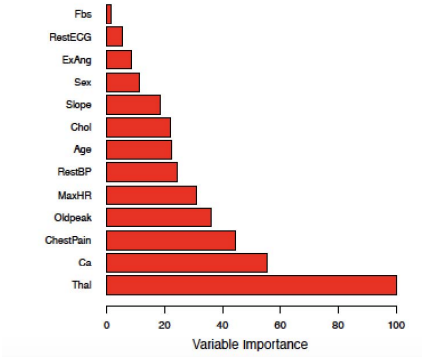
   (j) To find K in the K-means algorithm, we can penalize the objective function WCV (Within Cluster Variation) using AIC or BIC, in the same way we use AIC or BIC in model selection for linear regression. T F

4. Consider the unlabeled two-dimensional data represented in the following figure.
   (a) Draw hierarchical clusters for the data.
   (b) Draw a dendrogram for hierarchical clusters.

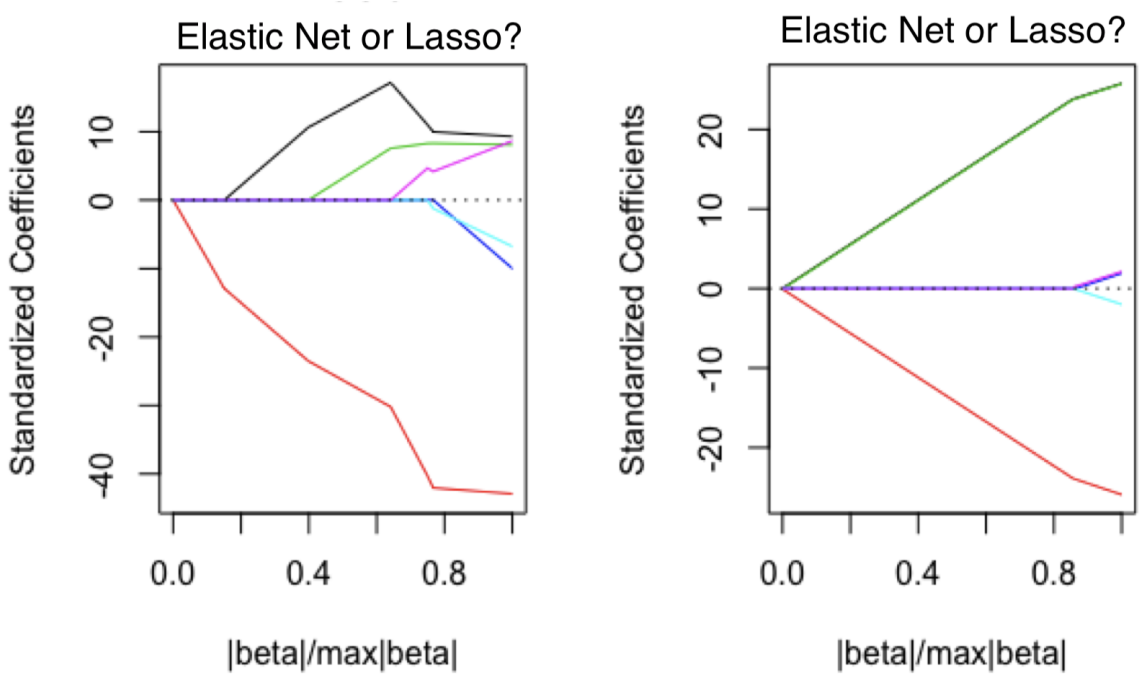


5. Assume that in a binary classification task, our data set has three highly correlated but relevant features X_1, X_2, X_3 and three irrelevant features X_4, X_5, X_6. X_1 and X_2 have very large positive correlation. X_1 and X_3 have very large negative correlation. X_4 and X_5 have very high positive correlation, andx X_5 and X_6 have very high negative correlation.

   (a) Assume that we use random forests for classification and it has low training error. Sketch a reasonable variable importance plot for this data set. As a reminder, a sample variable importance plot for random forests is shown below:



   (b) Assume that we perform classification using logistic regression with Elastic Net and with Lasso penalties. In the figure below, determine which penalty was used by circling either Elastic Net or Lasso on top of each subfigure. Also, on each figure, determine which coefficient path is associated with each of the variables $X_1, X_2, \ldots, X_6$.



6. Assume that we have a Ridge regression problem with only one predictor, and the true model is linear *without an intercept*, i.e. $Y = \beta_1 X + \epsilon$. Assume that we have n samples, $(x_i, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and we want to find the $\mathscr{L}_2$ regularized least squares estimate $\hat{\beta}_1$ from the data.

   (a) Formulate the objective function in terms of a candidate $\hat{\beta}_1$ and x_i's and y_i's, which are known. Assume that the regularization parameter is λ

   (b) Find $\hat{\beta}_1$ in terms of λ and the data.

Scratch paper

Name:

USC ID:

Scratch paper

Name:

USC ID: