Name:

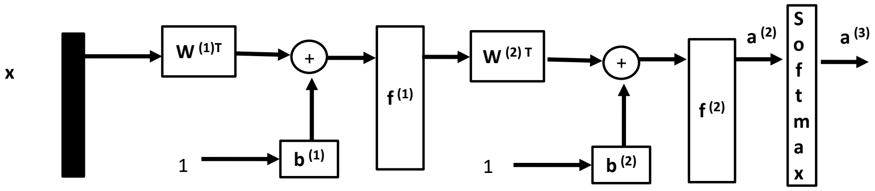
USC ID:

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.
- Please make sure to look at all pages in your exam.

Problem	Score	Earned
1	25	
2	20	
3	20	
4	20	
5	20	
Total	105	

1. Consider the following MLP



where

$$\mathbf{W}^{(1)} = \begin{bmatrix} -1 & -2 \\ -1 & 2 \end{bmatrix}$$
$$\mathbf{W}^{(2)} = \begin{bmatrix} 1 & 2 & 1 \\ -1 & -2 & 0 \end{bmatrix}$$
$$\mathbf{b}^{(1)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \mathbf{b}^{(2)} = \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix},$$

Also assume that the components of $\mathbf{f}^{(1)}$ are ReLus, i.e. the equation of each component is $f_1(x) = \max(0, x)$, the components of $\mathbf{f}^{(2)}$ are hyperpolic tangents and the equation of each component is $f_2(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. The second layer is followed by a softmax layer, and the output of each neuron in the softmax layer is calculated as:

$$a_j^{(3)} = \frac{e^{a_j^{(2)}}}{e^{a_1^{(2)}} + e^{a_2^{(2)}} + e^{a_3^{(2)}}}$$

where $a_j^{(3)}$ is the output of j^{th} neuron in the third layer and $a_j^{(2)}$ is the output of the j^{th} neuron in the second layer.

Answer the following questions:

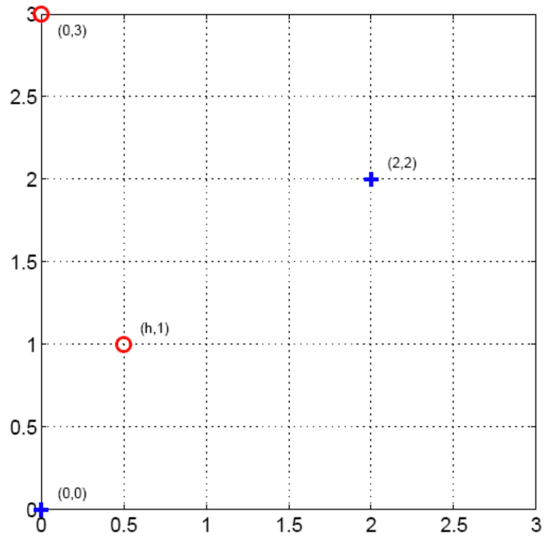
(a) How many neurons are there in the first, second, and third layers, respectively?

(b) How many elements are there in the output **a** (i.e. what is the dimension of the vector $\mathbf{a}^{(3)}$)?

(c) What is the output of the network when $\mathbf{x} = [2\ 1]^T$? In what class will it be classified by the network?

Solution:

2. Suppose we only have four training examples in two dimensions:



positive examples are $\mathbf{x}_1 = [0\ 0]^T, \mathbf{x}_2 = [2\ 2]^T$ and negative examples are $\mathbf{x}_3 = [h\ 1]^T, \mathbf{x}_4 = [0\ 3]^T$. h is a parameter.

(a) What is the largest value of h for which the training data are still linearly separable?

(b) Determine the support vectors when $h = 0.5$.

(c) When the training points are separable, does the slope of the maximum margin classifier change? Why?

(d) Assume that $h = .5$ and we have unlabeled data $\mathbf{x}_5 = [3\ 3]^T, \mathbf{x}_6 = [2\ 0.5]^T, \mathbf{x}_7 = [1\ 1.5]^T, \mathbf{x}_8 = [2.5\ 1.5]^T$. Which one will be labeled first, if we are performing self-training? Which one will be labeled first, if we are performing active learning?

3. Consider the unlabeled dataset with one feature: $\{0, 4, 5, 20, 25\}$. Assume that we want to obtain two top-level clusters in this dataset, using bottom-up hierarchical clustering. What will single linkage (minimum distance between members of clusters), complete linkage (maximum distance between members of clusters), and average linkage (average distance between members of clusters) output as the two clusters?

4. 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 0.25 and stays in San Diego office with probability 0.75. An employee in Los Angeles office is sent to the San Diego office with probability 0.3 and stays in Los Angeles office with probability 0.7. A new employee is assigned to Los Angeles office with probability 0.2 and to San Diego office with probability 0.8. An employee in San Diego office works between six and eight hours per day with probability 0.4, works more than eight hours with probability 0.4, and works less than six hours per day with probability 0.2. An employee in Los Angeles office works between six and eight hours per day with probability 0.1, works more than eight hours with probability 0.7, and works less than six hours per day with probability 0.2. 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 $4, 7, 10$, 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:

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

   (a) When the assumption of conditional independence of features holds, the Naïve Bayes' classifier provides the best accuracy among all possible classifiers. T F

   (b) The F1 score is not an appropriate measure for evaluating binarry classifiers when data are not imbalanced. T F

   (c) Leave-One-Out Cross Validation has less bias in estimating the error of a classifier for a large data set than 5 fold cross validation. T F

   (d) When classifying imbalanced data into two classes, we can decrease the threshold on class conditional probability $\Pr(Y = k|X_1 = x_1, \ldots, X_p = x_p]$ to increase the true positive rate at the expense of increasing the false negative rate. T F

   (e) Logistic regression assumes that the conditional odds of the outcome Y given the features, $\mathbb{O}[Y = k|X_1 = x_1, \ldots, X_p = x_p]$, is a logistic function of the features. T F

Scratch paper

Name:

USC ID:

Scratch paper

Name:

USC ID: