Rensselaer Polytechnic Institute

Department of Electrical, Computer, and Systems Engineering

ECSE-6610: PATTERN RECOGNITION

Spring 2018

Final Exam

Name: _____ RIN: _____ Score: _____

- This is a OPEN BOOK & OPEN NOTE exam. But you cannot access the Internet or use your laptop computer. Do the exam independently.
- There are a total of 100 points in the exam. Plan your work accordingly.
- Write out the steps for all problems to receive the full credit. Use additional pages if necessary.
- Date: May 7th, 2018.
- Time: 3:00 pm - 6:00 pm.

Problem	Points	Scores
Problem 1: True or False	20	
Problem 2: AdaBoost	20	
Problem 3: Hidden Markov Model	20	
Problem 4: Multilayer Neural Networks	15	
Problem 5: Unsupervised Learning	25	

Problem 1: True or False (20 points)

(a) We can get multiple local optimum solutions if we solve a linear regression problem by minimizing the sum of squared errors using gradient descent.

○ True ○ False

(b) When the hypothesis space is richer, over fitting is more likely.

○ True ○ False

(c) In terms of feature selection, L2 regularization is preferred since it comes up with sparse solutions.

○ True ○ False

(d) A weak learner with less than 50% accuracy does not present any problem to the Adaboost algorithm.

○ True ○ False

(e) EM algorithm is always guaranteed to converge to a local minima.

○ True ○ False

(f) We can use gradient descent to learn a Gaussian Mixture Model.

○ True ○ False

(g) Assuming Boolean attributes, the depth of a decision tree classifier can never be larger than the number of training examples

○ True ○ False

(h) Logistic loss is better than L2 loss in classification tasks.

○ True ○ False

(i) If the data is not linearly separable, then the gradient descent algorithm for training a logistic regression classifier will never converge.

○ True ○ False

(j) k-means clustering is a special case of hard EM.

○ True ○ False

Problem 2: AdaBoost (20 points)

Suppose you have two weak learners, h_1 and h_2, and a set of 17 points

(a) **[3 points]** You find that h_1 makes one mistake and h_2 makes four mistakes on the dataset. Which learner will AdaBoost choose in the first iteration (namely $t = 1$)? Justify your answer.

(b) **[3 points]** What is α_1?

(c) **[4 points]** Calculate the data weighting co-efficients D_2 for the following two cases: (1) the points on which the chosen learner made a mistake and (2) the points on which the chosen learner did not make a mistake.

(d) **[10 points]** Consider a simple modification to the AdaBoost algorithm with the update rule

$$D_{t+1}(i) = \begin{cases} D_t(i) & y_t(x_i) = t_i \\ D_t(i)e^{2\alpha_t} & y_t(x_i) \neq t_i \end{cases}$$

where $\alpha_t = \frac{1}{2}\ln(\frac{1-\epsilon_t}{\epsilon_t})$. To normalize the weights, we continue to replace $D_{t+1}(i)$ by $D_{t+1}(i)/Z_{t+1}$, *i.e.*, $D_{t+1}(i) \leftarrow D_{t+1}(i)/Z_{t+1}$ where $Z_{t+1} = \sum_{i=1}^{N} D_{t+1}(i)$.

Prove that $Z_{t+1} = 2(1-\epsilon_t)$.

Hint: Notice that if the weights are normalized, then $\epsilon_t = \sum_{i=1}^{N} D_t(i) I(y_t(x_i) \neq t_i)$.

Problem 3: Hidden Markov Model (20 points)

Consider an HMM with an explicit absorber state w_0 and unique null visible symbol v_0 with the following transition probabilities a_{ij} and symbol probabilities b_{jk} (where the matrix indexes begin at 0):

$$a_{ij} = \begin{pmatrix} 1 & 0 & 0 \\ 0.2 & 0.3 & 0.5 \\ 0.4 & 0.5 & 0.1 \end{pmatrix} \qquad \text{and} \qquad b_{jk} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.7 & 0.3 \\ 0 & 0.4 & 0.6 \end{pmatrix}$$

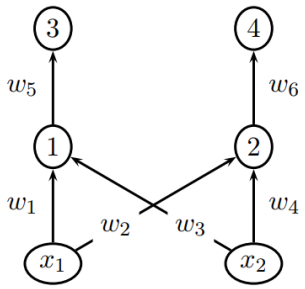
(a) **[7 points]** Give a graph representation of this Hidden Markov Model.

(b) **[10 points]** Suppose the initial hidden state at $t = 0$ is w_1. Starting from $t = 1$, what is the probability it generates the particular sequence $V^3 = \{v_2, v_1, v_0\}$?

(c) **[3 points]** Given the above sequence V^3, what is the most probable sequence of hidden states?

Problem 4: Multilayer Neural Networks (15 points)

Consider the Neural network given below.



Assume that all internal nodes and output nodes compute the tanh function. In this question, we will derive an explicit expression that shows how back propagation (applied to minimize the least squares error function) changes the values of w_1, w_2, w_3, w_4, w_5 and w_6 when the algorithm is given the example (x_1, x_2, y_3, y_4) with y_3 and y_4 being ground-truth at unit 3 and 4 respectively (there are no bias terms). Assume that the learning rate is η. Let o_1 and o_2 be the output of the hidden units 1 and 2 respectively. Let o_3 and o_4 be the output of the output unit 3 and 4, respectively.

Hint: Derivative of $\tanh(x) = 1 - \tanh^2(x)$.

(a) **[5 points]** Forward propagation. Write equations for o_1, o_2, o_3 and o_4.

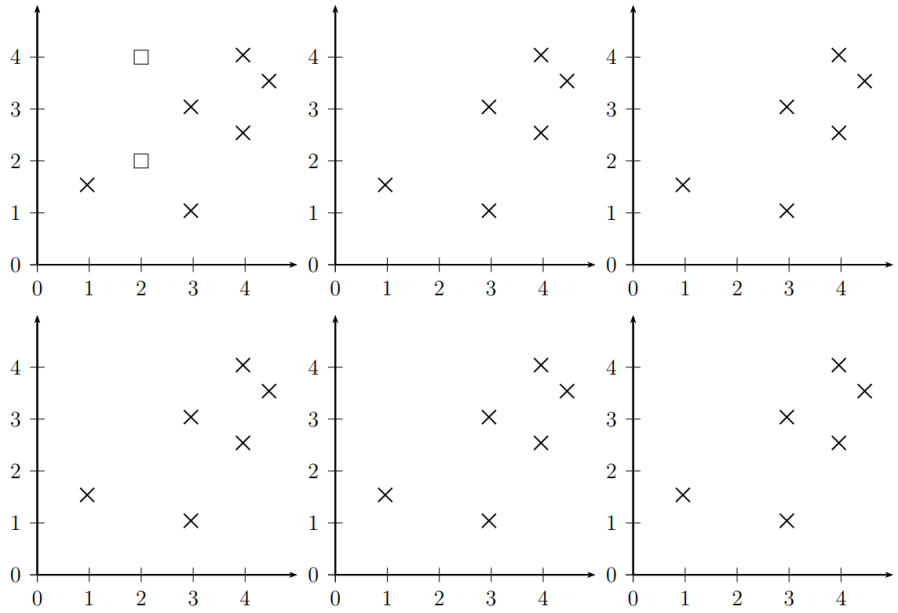
(b) **[5 points]** Backward propagation. Write equations for σ_1, σ_2, σ_3 and σ_4 where σ_1, σ_2, σ_3 and σ_4 are the values propagated backwards by the units denoted by 1, 2, 3 and 4 respectively in the neural network.

(c) **[5 points]** Give an explicit expression for the new (updated) weights w_1, w_2, w_3, w_4, w_5 and w_6 after backward propagation.

Problem 5: Unsupervised Learning (25 points)

(a) **[5 points]** Suppose we want to cluster n samples into c clusters, (i) what is the definition of the "Sum-of-Square-Error" clustering criterion? and (ii) what is the interpretation of this criterion?

(b) **[10 points]** Starting with two cluster centers indicated by squares, perform k-means clustering on the six data points (denoted by $\times$). In each panel, indicate the data assignment and in the next panel show the new cluster means. Stop when converged or after 6 steps whichever comes first. (using Euclidean distance measure).



(c) **[10 points]** Draw two agglomerative clustering trees for the following data set. You must use single link clustering with distance measure $d_{min}(D_i, D_j)$ and $d_{max}(D_i, D_j)$, respectively. Disscuss your observations.