Hong Kong University of Science and Technology

COMP327/527: Pattern Recognition

Fall 1996

Final Examination

19 December 1996, 8:30–11:30am

Student Name: __________________________

Student Number: __________________________

Instructions

1. This is an open-book, open-notes examination.
2. Check that you have all 14 pages (excluding this cover page).
3. Write your name and student number on this page.
4. Answer all questions in the space provided. Rough work should be done on the back pages.
5. Make your answers as concise as possible.

Question 1 (16%):	________________
Question 2 (13%):	________________
Question 3 (10%):	________________
Question 4 (15%):	________________
Question 5 (10%):	________________
Question 6 (8%):	________________
Question 7 (8%):	________________
Question 8 (8%):	________________
Question 9 (12%):	________________
TOTAL (100%):	________________

1. (16%)

   Suppose the probability density functions of a univariate feature x for two classes are as follows:

$$p(x \mid \omega_1) = \begin{cases} 5a & 3 \le x < 4 \\ 3a & 4 \le x < 6 \\ 4a & 6 \le x \le 7 \\ 0 & \text{otherwise} \end{cases}$$

$$p(x \mid \omega_2) = \begin{cases} b & 1 \le x \le 10 \\ 0 & \text{otherwise} \end{cases}$$

   Assume that $Pr(\omega_1) = Pr(\omega_2) = 1/2$.

   (a) [2%] What should be the values of a and b?

   (b) [4%] What are the decision regions of the two classes for the domain $1 \le x \le 10$ when optimal (Bayesian) decision is made.

(c) [3%] What is the Bayes error?

(d) [7%] Repeat parts (b) and (c) if the prior probabilities are changed to 1/3 and 2/3 for $Pr(\omega_1)$ and $Pr(\omega_2)$, respectively.

2. (13%)

Consider the following four training examples:

x_1	x_2	x_3	x_4	t
1	1	-2	-4	1
2	-1	2	3	-1
5	2	-5	0	-1
0	2	1	2	1

Since these four examples are linearly separable, a simple perceptron with four input units is trained to perform the classification task.

(a) [8%] Suppose all the weights of the perceptron are initialized to 0. Show the execution trace of the perceptron learning algorithm by listing the sequence of weight changes until a solution is found.

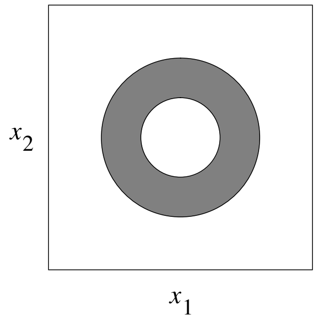
(b) [5%] Suppose features x_1 and x_3 are ignored so that only x_2 and x_4 are used to describe the training examples:

x_2	x_4	t
1	-4	1
-1	3	-1
2	0	-1
2	2	1

Show that the four examples become linearly non-separable.

3. (10%)

Consider the following two-dimensional feature space, where the shaded area corresponds to one class and the rest corresponds to another class:



The shaded area is formed by two concentric circles with center (a, b) and radii r_1 and r_2 $(r_1 < r_2)$.

Suppose we define two additional features x_3 and x_4 as:

$$x_3 = x_1^2 \qquad x_4 = x_2^2$$

Show that there exists a network composing of two layers of simple perceptrons with four inputs (x_1, x_2, x_3, x_4) and one output that can represent the decision regions above corresponding to two classes. Assuming that all the simple perceptrons have output values in $\{0, 1\}$, give the values of all the weights and biases in the network. (Hint: One solution has two hidden units and one output unit.)

(cont'd)

4. (15%)

Consider an extended version of adaline, called q-adaline here for convenience. A q-adaline with two inputs (x_1 and x_2) and one output (y) has weights w_{12}, w_1, w_2, and θ that relate the inputs with the output by the following equation:

$$y = w_{12}x_1x_2 + w_1x_1 + w_2x_2 + \theta$$

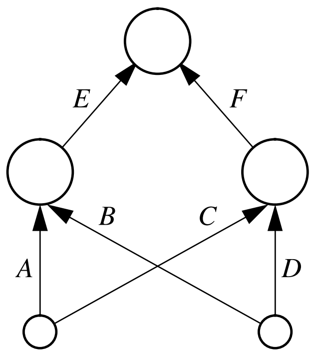
(a) [8%] Using the gradient-descent approach, derive the weight updating rules for w_{12}, w_1, w_2, and θ. The total quadratic error over the entire training set of n examples should be used as the error function.

(b) [4%] What is the advantage of q-adaline over adaline? What is its disadvantage?

(c) [3%] Does the error function have local minima? Explain your answer.

5. (10%)

Consider the following neural network classifier with two input units and one output unit:



(a) [6%] Suppose we begin with all weights having exactly the same value and run the back-propagation learning algorithm on some training data. Which statements below will always be true? Explain your reasoning for each.

i. Weights A and B never differ.

ii. Weights A and C never differ.

iii. Weights E and F never differ.

(b) [4%] Give a short explanation of one good reason not to initialize all weights to the same value when running the back-propagation learning algorithm.

6. (8%)

Compare a typical feedforward neural network classifier with a typical nearest-neighbor (1-NN) classifier with respect to computational (i.e. time) and storage (i.e. space) requirements during the training and classification stages.

7. (8%)

"Since reducing the dimensionality of a feature space from d to m ($<d$) typically results in a loss of data variance, the classification rate obtained using the m transformed features can never exceed that obtained using the d original features." Is this statement correct? Explain why or why not.

8. (8%)

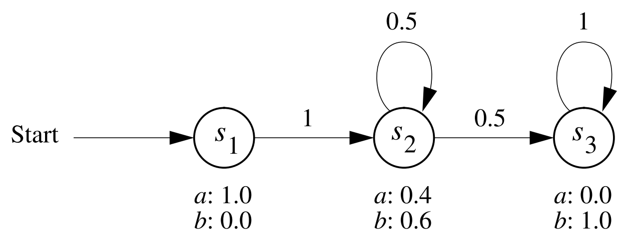
Suppose n (>1) unique sample points in a feature space form a single cluster. It has been shown in Problem Set 4 that the sample mean $\boldsymbol{\mu}$ is the location of the cluster center $\mathbf{m}$ that minimizes the sum of squared errors

$$J = \sum_{k=1}^{n} \|\mathbf{x}^{(k)} - \mathbf{m}\|^2$$

Suppose a two-cluster partition is applied to the n points using the K-means clustering algorithm. The definition of J above will then be generalized to the optimality criterion used in K-means clustering. Prove that the minimum attainable value of J for the two-cluster case is strictly smaller than that for the single-cluster case.

9. (12%)

Consider the following hidden Markov model (HMM):



Associated with each state transition is a number which represents the probability of that transition. For each state, the probability of seeing each observation symbol is shown. For example, the probability of making transition from state s_2 to state s_3 is 0.5, and the probabilities of seeing symbols a and b at state s_2 are 0.4 and 0.6, respectively.

(a) [8%] What is the probability that this HMM will generate the observation sequence $aabb$ as its first four symbols?

(b) [4%] What is the most likely sequence of states for the observation sequence in part (a)? Justify your answer.

***** THE END *****