

CS273 Final Exam  
Introduction to Machine Learning: Fall 2016  
**Friday December 9th, 2016**

Name: \_\_\_\_\_

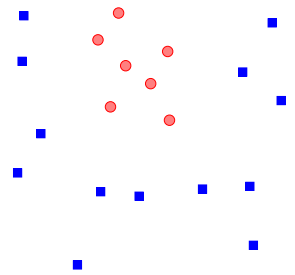
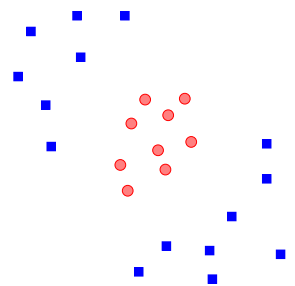
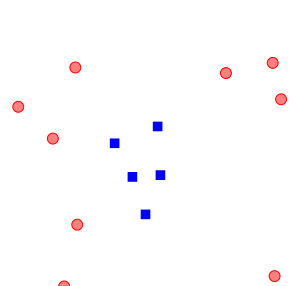
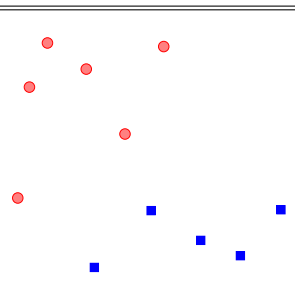
ID #: \_\_\_\_\_

UCINetID: \_\_\_\_\_@uci.edu      Seat (Row, #): \_\_\_\_\_

- The exam will begin on the next page. **Please DO NOT turn the page until told.**
- Pages are printed double-sided; when you are told to begin the exam, please check both sides of each page and verify that you have all **10** pages.
- Total time is 1 hour 50 minutes. **READ THE EXAM FIRST** and organize your time; don't spend too long on any one problem.
- Please **write clearly** and **show all your work**.
- If you need clarification on a problem, please raise your hand and wait for the instructor to come over.
- You may use one sheet of your own, handwritten notes for reference, and a calculator.
- Turn in any scratch paper with your exam

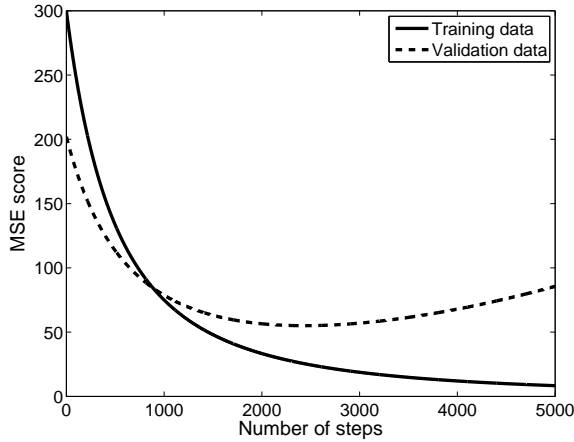
### Problem 1: (8 points) Separability & Classifiers

For each of the following examples of training data and classifiers, state whether there exists a set of parameters that can separate the data and justify your answer briefly (~1 sentence).

 A scatter plot showing two classes of data points: blue squares and red circles. The blue squares are distributed in a ring-like pattern around the perimeter of the plot, while the red circles are clustered in the center. This configuration is linearly separable in a 2D space with quadratic features.	Linear classifier with quadratic features:
 A scatter plot showing two classes of data points: blue squares and red circles. The blue squares are located on the left side of the plot, and the red circles are on the right side. A vertical line can separate the two classes.	Depth-two decision tree:
 A scatter plot showing two classes of data points: blue squares and red circles. The blue squares are clustered in the center, and the red circles are on the outer edges. This configuration is not linearly separable in a 2D space with quadratic features.	Depth-two decision tree:
 A scatter plot showing two classes of data points: blue squares and red circles. The blue squares are clustered in the bottom-left area, and the red circles are in the top-right area. A diagonal line can separate the two classes.	Linear perceptron classifier:

## Problem 2: (9 points) Training & Test Error

Consider the following plot, which shows the training set error and the validation test set error for a neural network model as it is trained, i.e., the horizontal axis indicates the number of iterations of training (gradient steps). Note that the training error decreases monotonically, while the test error does not.



- (a) Explain what is happening and why; suggest a possible solution.

Now suppose that we were to re-train the model with 10 times as much data, while keeping all other aspects (initialization, etc.) the same.

- (b) Would you expect the training curve to be different? If so, sketch how it might change.
- (c) Would you expect the validation (test) curve to be different? If so sketch how it might change.

### Problem 3: (12 points) Decision Trees

Consider the table of measured data given at right. (Note that some data points are repeated.) We will use a decision tree to predict the outcome  $y$  using the three features,  $x_1, \dots, x_3$ . In the case of ties, we prefer to use the feature with the smaller index ( $x_1$  over  $x_2$ , etc.) and prefer to predict class 1 over class 0. You may find the following values useful (although you may also leave logs unexpanded):

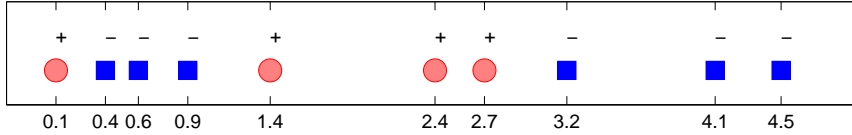
$$\begin{aligned} \log_2(1) &= 0 & \log_2(2) &= 1 & \log_2(3) &= 1.59 & \log_2(4) &= 2 \\ \log_2(5) &= 2.32 & \log_2(6) &= 2.59 & \log_2(7) &= 2.81 & \log_2(8) &= 3 \end{aligned}$$

$x_1$	$x_2$	$x_3$	$y$
0	0	1	1
0	1	0	1
1	1	1	1
1	1	1	1
0	0	0	0
0	0	0	0
1	1	0	0

- (a) What is the entropy of  $y$ ?
  
  
  
  
  
  
  
  
  
  
- (b) Which variable would you split first? Justify your answer.
  
  
  
  
  
  
  
  
  
  
- (c) What is the information gain of the variable you selected in part (b)?
  
  
  
  
  
  
  
  
  
  
- (d) Draw the rest of the decision tree learned on these data.

#### Problem 4: (12 points) Classification in 1D

We observe a collection of training data with one feature, “ $x$ ” and a class label  $y \in \{-, +\}$ , shown here; class  $+$  is indicated by circles and  $-$  by squares, and also labeled with text for redundancy. Answer each of the following questions. Express error rates as the fraction of data points incorrectly classified.



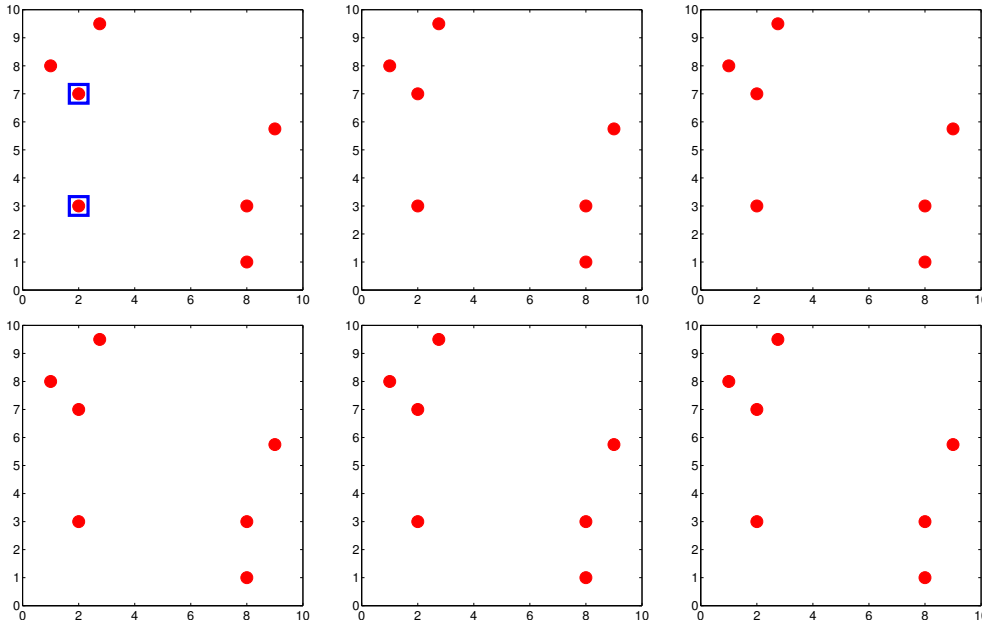
- (a) What is the best training error rate we can achieve on these data from a linear classifier on the original input features ( $f(x) = \text{sign}(ax + b)$ )? Explain briefly (sketch + 1-2 sentences): how it is achieved.
- (b) What is the best training error we can achieve from a linear classifier with quadratic features, e.g.,  $f(x) = \text{sign}(ax^2 + bx + c)$ ? Explain briefly how it's achieved.
- (c) What is the best training error we can achieve from a decision tree classifier? Explain briefly how it's achieved.
- (d) What is the best training error we can achieve from a two-layer neural network (multi-layer perceptron) with input features “ $x$ ” and “1”? Explain briefly how it's achieved (e.g., approximate # of hidden nodes & what they might look like).

### Problem 5: (12 points) Clustering

Consider the two-dimensional data points plotted in each panel. In this problem, we will cluster these data using two different algorithms, where each panel is used to show an iteration or step of the algorithm.

#### k-means

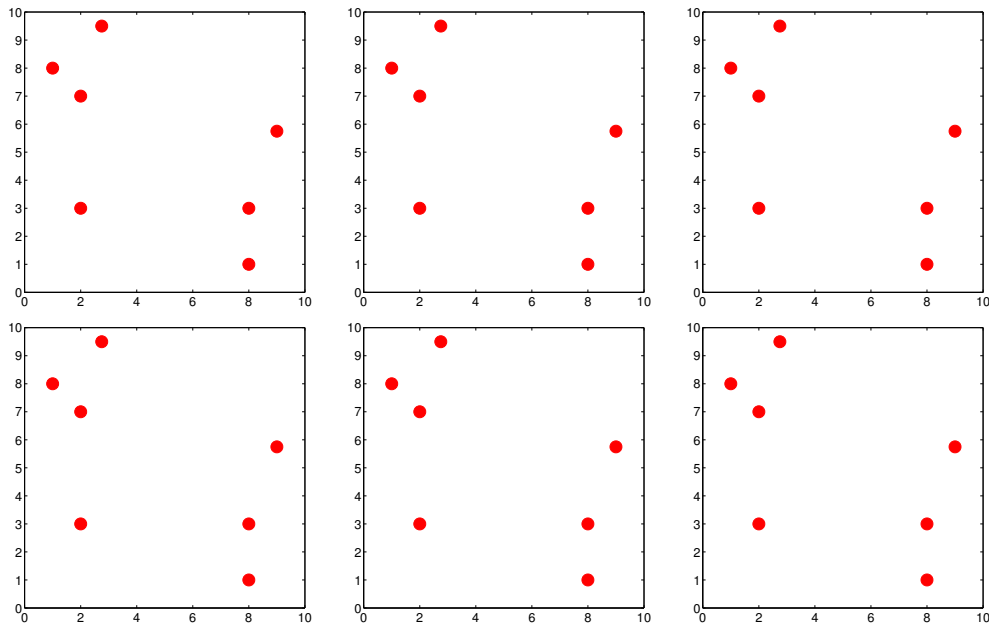
(a) Starting from the two cluster centers indicated by squares, perform k-means clustering on the data points. In each panel, indicate (somehow) the data assignment, and in the next panel show the new cluster centers. Stop when converged, or after 6 steps (3 iterations), whichever is first. It may be helpful to recall from our nearest-neighbor classifier that the set of points nearer to  $A$  than  $B$  is separated by a line.



(b) Write down the cost function optimized by the k-means algorithm, explaining your notation.

## Linkage

(a) Now execute the hierarchical agglomerative clustering (linkage) algorithm on these data points, using “complete linkage” (maximum distance) for the cluster scores. Stop when the algorithm would terminate, or after 6 steps, whichever is first. Show each step separately in a panel.



(b) What is the algorithmic (computational) complexity of the hierarchical clustering algorithm? Briefly justify your answer.

**Problem 6: (9 points) Bayes Classifiers and Naïve Bayes**

Consider the table of measured data given at right. We will use the three observed features  $x_1, x_2, x_3$  to predict the class  $y$ . In the case of a tie, we will prefer to predict class  $y = 0$ .

$x_1$	$x_2$	$x_3$	$y$
0	0	0	0
0	0	0	1
0	1	1	0
1	1	0	0
1	1	0	1
1	0	1	1
1	1	1	1

(a) Write down the probabilities necessary for a naïve Bayes classifier:

(b) Using your naïve Bayes model, what value of  $y$  is predicted given observation  $(x_1, x_2, x_3) = (000)$ ?

(c) What is the class probability  $p(y = 1 | x_1 = 0, x_2 = 1, x_3 = 1)$ ?



**Problem 7: (9 points) VC Dimension**

Consider the following classifiers  $f(x)$ , defined on data with two real-valued features  $x = (x_1, x_2)$  and predicting a binary class  $y \in \{-1, +1\}$ . Answer the following questions about their VC dimension, by showing that it is at least as large as the value you give.

- (a) What is the VC dimension of the linear classifier,

$$f(x) = \text{sign}(a + bx_1 + cx_2),$$

with parameters  $(a, b, c)$ ?

- (b) What is the VC dimension of a decision stump,

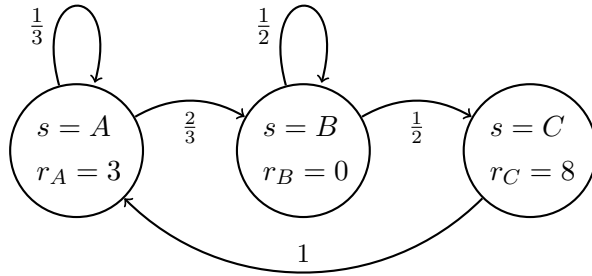
$$f(x) = a \text{ if } x_i < t \text{ else } b,$$

with parameters  $(a, b, t, i)$ ?

- (c) Suppose that the data points were forced to have *binary valued* features, e.g.,  $x_i \in \{0, 1\}$ , rather than real values. Would this change your answer for (a) or (b)?

### Problem 8: (9 points) Markov models

Consider the Markov model shown here:



$$\Pr[A \rightarrow A] = 0.33$$

$$\Pr[A \rightarrow B] = 0.66$$

$$\Pr[B \rightarrow B] = 0.5$$

$$\Pr[B \rightarrow C] = 0.5$$

$$\Pr[C \rightarrow A] = 1.0$$

where the transition probabilities are shown next to each arc and at right, and the rewards  $r_s$  associated with each state  $s$  are shown inside the circles. Assume a future discounting factor of  $\gamma = \frac{1}{2}$ .

- Compute  $J^1(s)$ , the expected discounted sum of rewards for state sequences of length 1 (e.g.,  $[A]$ ) starting in each state  $s$ .
- Compute  $J^2(s)$ , the expected discounted sum of rewards for state sequences of length 2 (e.g.,  $[C \rightarrow A]$ ) starting in each state  $s$ .
- Compute  $J^3(s)$ , the expected discounted sum of rewards for state sequences of length 3 (e.g.,  $[B \rightarrow B \rightarrow C]$ ) starting in each state  $s$ .