

CS178 Final Exam
Machine Learning & Data Mining: Winter 2016
Thursday March 17th, 2016

Your name: SOLUTIONS

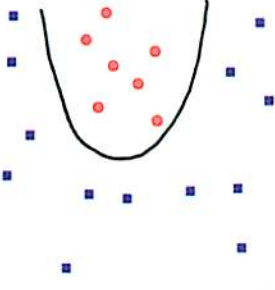
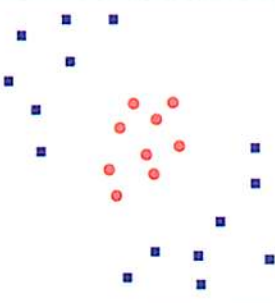
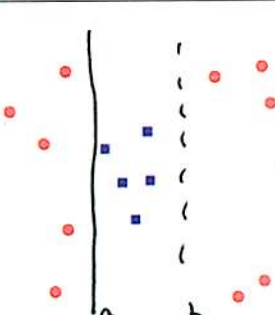
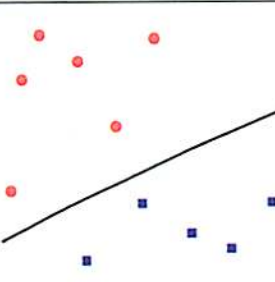
Your ID Number and UCINetID
(e.g., 12345678 / myname@uci.edu): 31415926 / parreale Cui.

Your seat (row and number): 6600

- 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

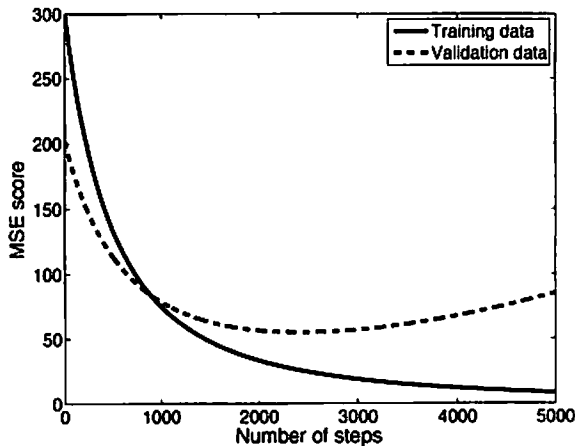
Problem 1: (8 points) Separability & Classifiers

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).

	<p>Linear classifier with quadratic features:</p> <p>Yes - sketched boundary is</p> $x_2 = ax_1^2 + bx_1 + c$
	<p>Depth-two decision tree:</p> <p>No - no 2 level axis-aligned split.</p>
	<p>Depth-two decision tree:</p> <p>Yes, eg:</p> <pre> graph TD A((x1 > a)) --> B(+1) A --> C((x1 > b)) C --> D(-1) C --> E(+1) </pre>
	<p>Linear perceptron classifier:</p> <p>Yes; linear decision boundary.</p>

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.

overfitting - the model is becoming too tuned to the training data, degrading validation performance.

Use early stopping, regularization, or some other form of complexity control.

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.

No / Similar - maybe slightly higher, esp. at the far right.

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

Yes - flatter, probably no or less upturn at the far right



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):

$$\log_2(1) = 0 \quad \log_2(2) = 1 \quad \log_2(3) = 1.59 \quad \log_2(4) = 2 \\ \log_2(5) = 2.32 \quad \log_2(6) = 2.59 \quad \log_2(7) = 2.81 \quad \log_2(8) = 3$$

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 ? $p(y) = 4/7$

$$H = 4/7 \log_2 7/4 + 3/7 \log_2 7/3 \approx .985 \text{ bits}$$

(b) Which variable would you split first? Justify your answer.

x_3 clearly has lowest entropy after split (highest info gain)

$$x_1: \begin{matrix} 0 \Rightarrow 1100 \\ 1 \Rightarrow 110 \end{matrix} \quad x_2: \begin{matrix} 0 \Rightarrow 100 \\ 1 \Rightarrow 1110 \end{matrix} \quad x_3: \begin{matrix} 0 \Rightarrow 1000 \\ 1 \Rightarrow 111 \end{matrix}$$

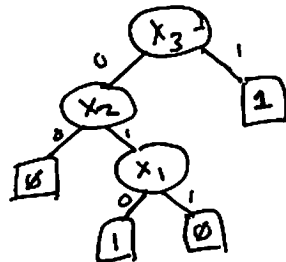
(c) What is the information gain of the variable you selected in part

(b)? $H(4/7) - 4/7 H(4/4) - 3/7 \cdot 0.$

$$\approx .985 - .463 = .522$$

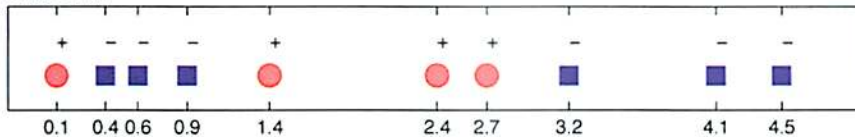
$$\approx .522 \text{ bits.}$$

(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.

3/10

+ \rightarrow (-)
 0 | 0 0 0 0 0 0 0 0 0 0
 or
 0 0 0 0 0 0 | 0 0 0 0

set the decision boundary, $ax + b = 0$,
 at $x = .3$ or $x = 3$, for example.

- (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.

1/10

0 0 0 0 (-) (+) 0 0 0 0
 $ax^2 + bx + c$

$ax^2 + bx + c = -(x-1)(x-3)$.
 (for example)

- (c) What is the best training error we can achieve from a decision tree classifier? Explain briefly how it's achieved.

0/10

keep splitting until all the training data
 are correct.

- (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., # of hidden nodes & what they look like).

0/10

Use say 3 hidden nodes, activating at $x = .3$, $x = 1$, and $x = 3$.

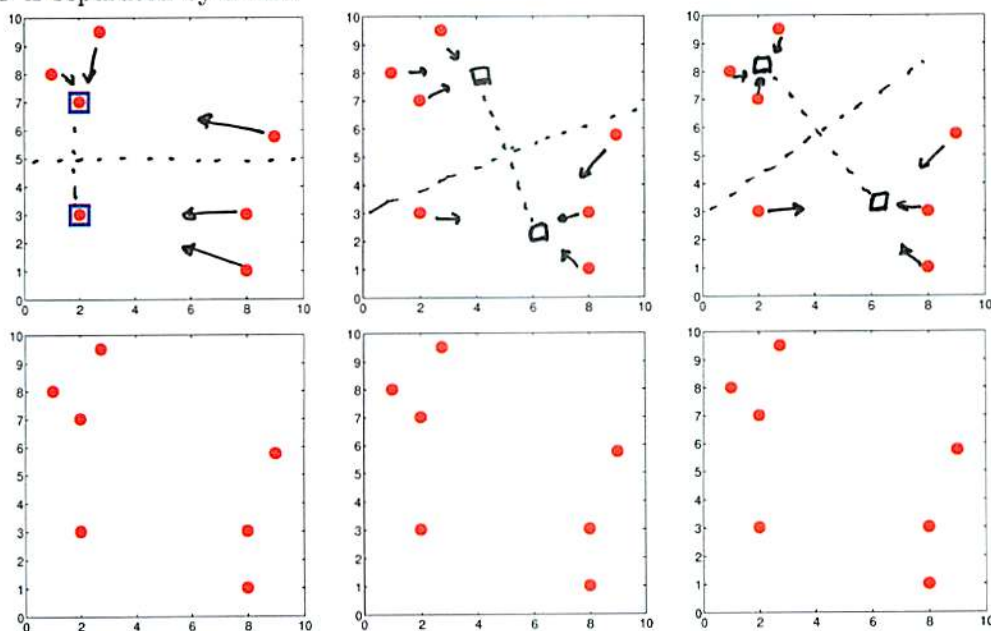
then, a linear combination of these values can separate the data.

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.



50p.

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

$$J = \sum_i \|x^{(i)} - \mu_{z_i}\|_2^2$$

$$\|v\|_2^2 = \sum_j v_j^2 \text{ is Euclidean length squared}$$

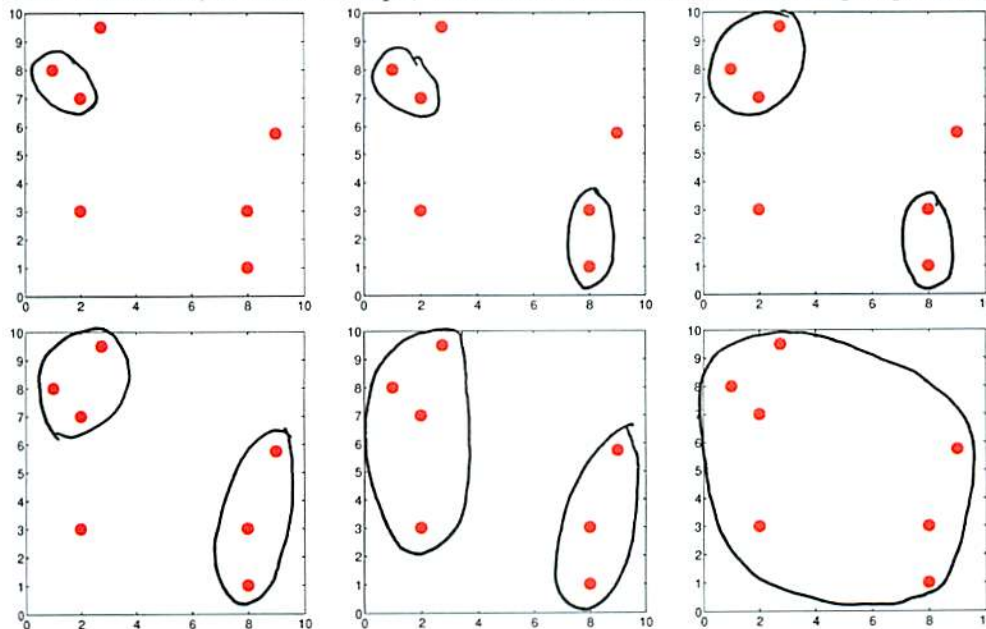
$x^{(i)}$ = data point i .

z_i = cluster assignment of $x^{(i)}$

μ_c = cluster center of cluster c .

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.

$$O(m^2 \log m)$$

Compute $O(m^2)$ cluster distances for the original m points.

Heaps law sorted $\Rightarrow O(m^2 \log(m^2)) = O(m^2 \log m)$.

Each iteration i (m times):

merge closest cluster pair (constant time)

update $(m-i)$ cluster distances in sorted structure $\Rightarrow O((m-i) \log m)$

\Rightarrow total $O(m^2 \log m)$.

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:

$$\begin{aligned}
 p(y=1) &= 4/7 & p(y=0) &= 1 - p(y=1) = 3/7 \\
 p(x_1=1 | y=1) &= 3/4 & p(x_1=1 | y=0) &= 1/3 \\
 p(x_2=1 | y=1) &= 1/2 & p(x_2=1 | y=0) &= 2/3 \\
 p(x_3=1 | y=1) &= 1/2 & p(x_3=1 | y=0) &= 1/3
 \end{aligned}$$

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

$$\begin{aligned}
 & 3/7 \cdot 2/3 \cdot 1/3 \cdot 2/3 & \text{vs} & & 4/7 \cdot 1/4 \cdot 1/2 \cdot 1/2 \\
 & 4/9 & & \text{vs} & 1/4 & \Rightarrow \text{predict } y=0.
 \end{aligned}$$

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

$$p(y=1 | 011) = \frac{4/7 \cdot 1/4 \cdot 1/2 \cdot 1/2}{4/7 \cdot 1/4 \cdot 1/2 \cdot 1/2 + 3/7 \cdot 2/3 \cdot 2/3 \cdot 1/3} = \frac{4 \cdot 827}{4 \cdot 27 + 3 \cdot 4 \cdot 4 \cdot 4} = \frac{9}{9+16} = 9/25.$$

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.

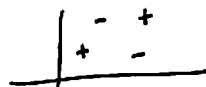
(a) What is the VC dimension of the linear classifier,

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

with parameters (a, b, c) ?

3 - perceptron in 2 features has vc dim d+1

Recall: Can shatter: but cannot separate

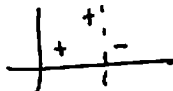


(b) What is the VC dimension of a decision stump,

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

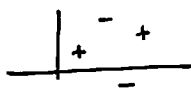
with parameters (a, b, t, i) ?

3 -



etc.

Not required, but again, cannot shatter



No single axis split.

(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)?

For (a) - No. Just place the 3 points at 2 still shatters.



For (b) - Yes. For point placement like ↗,

we cannot separate with a single, axis aligned split.