CS178 Final Exam

Machine Learning & Data Mining: Winter 2014

Friday March 21st, 2014

Your name:

Your UCINetID (e.g., myname@uci.edu):

Your seat (row and number):

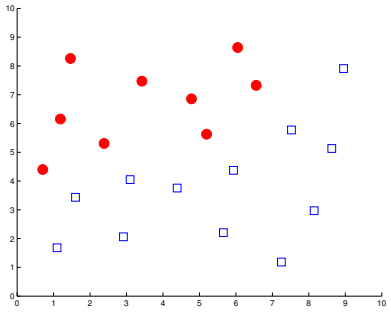
- Total time is 1 hour 50 minutes. READ THE EXAM FIRST and organize your time; don't spend too long on any one problem.
- Please **write clearly** and **show all your work**.
- If you need clarification on a problem, please raise your hand and wait for the instructor to come over.
- Turn in any scratch paper with your exam.

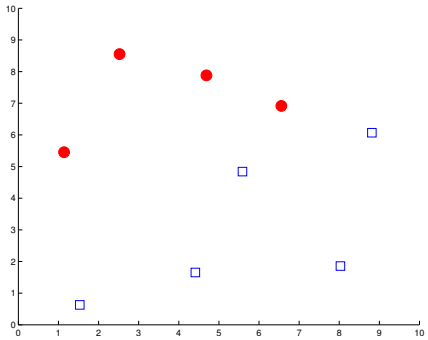
Good luck and enjoy your break!

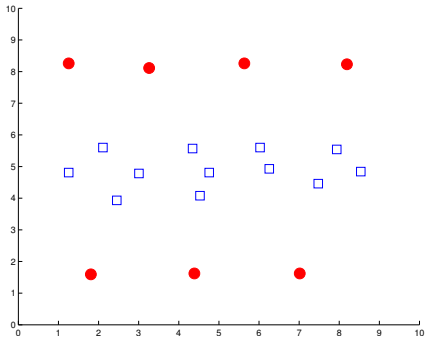
(This page intentionally left blank)

Problem 1: Separability

For each of the following examples of training data, sketch a classification boundary that separates the data. State whether or not the data are linearly separable, and if not, give a set of features that would allow the data to be separated.







Problem 2: Decision Trees

Training data is shown at left. Note that some data may be repeated (observed more than once).
You may find the following values useful (although you may also leave logs unexpanded):
$\log_2(1) = 0 \quad \log_2(2) = 1 \quad \log_2(3) = 1.59 \quad \log_2(4) = 2 \quad \log_2(5) = 2.32 \quad \log_2(6) = 2.59$
In case of ties, we prefer to use x_1 before x_2 before x_3, and prefer to predict class 1 over class -1.

(a) What is the entropy of the target variable, y?

Training data:

x_1	x_2	x_3	y
0	0	0	1
0	0	1	1
0	0	1	1
0	1	1	-1
1	1	0	1
1	1	0	-1

(b) Compute the information gain obtained if we split on feature x_1. (Again, you may leave logs unexpanded.)

(c) What is the best feature to split on first? Draw the full decision tree learned on these data.

(d) What is the **training** error for your model in part (c)?

(Problem 2 continued:)
(e) The table at left gives test data. What is the **test** error for your model in part (c)?

Test data:

x_1	x_2	x_3	y
0	0	0	1
0	1	0	-1
1	1	1	-1

Problem 3: Multiple Choice

For the following questions, assume that we have m data points $y^{(i)}$, $x^{(i)}$, $i = 1 \dots m$, each with n features, $x^{(i)} = [x_1^{(i)} \ \dots \ x_n^{(i)}]$. **Circle one answer for each:**

Suppose that we are training a decision tree classifier on our data. The parameter `minParent` prevented a node from splitting if there were fewer than `minParent` data at that node. Increasing `minParent` will make our classifier **more** **equally** **less** likely to overfit the data.

In class we learned that in the standard model, each node uses only one feature for the split. Instead, we decide to split on a linear combination of several features. This will most likely make our classifier **more** **equally** **less** likely to overfit the data.

We decide to use bagging to build an ensemble of classifiers. Compared to using just one classifier, this will likely make us **more** **equally** **less** likely to overfit the data.

We build an ensemble of regressors using gradient boosting. Compared to using just one regressor, this will likely make us **more** **equally** **less** likely to overfit the data.

Decreasing regularization on the weights in logistic regression will most likely
(**increase** **decrease** **not change**) the VC dimension.

Decreasing the maximum depth of a decision tree will most likely
(**increase** **decrease** **not change**) the VC dimension.

True or **false**: Linear regression can be solved using either matrix algebra or gradient descent.

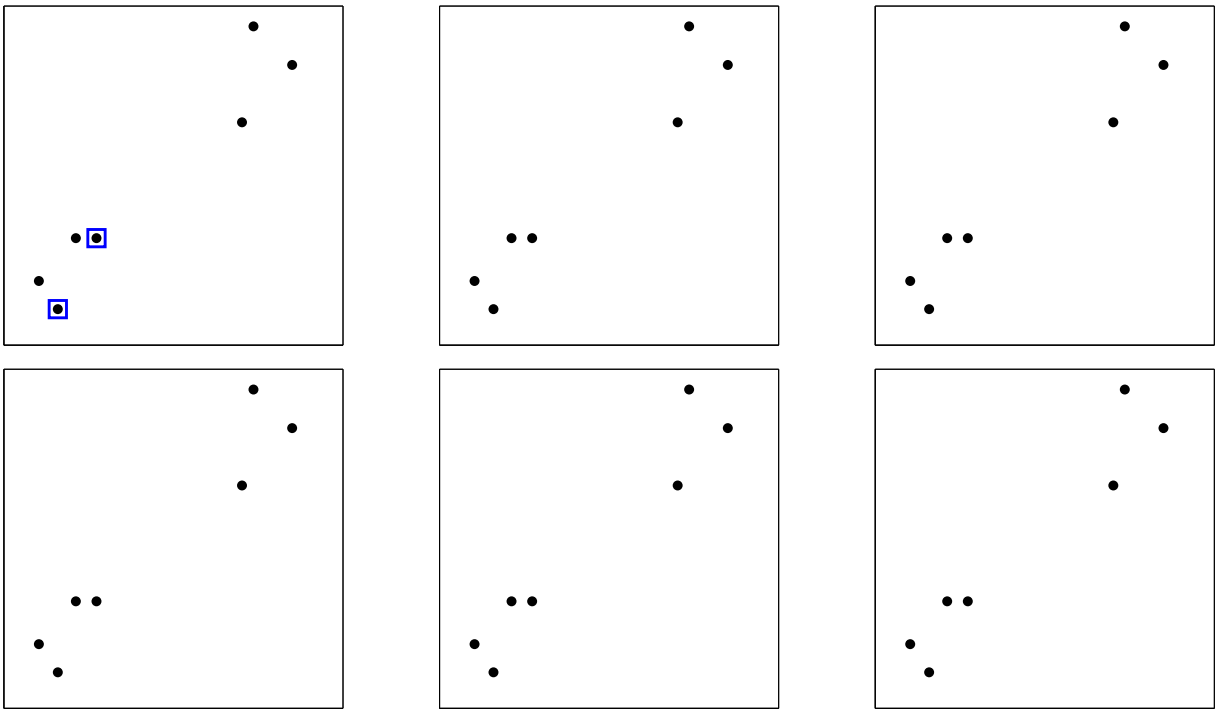
True or **false**: The predictions of a classification (decision) tree will be affected by pre-processing to normalize the magnitude of each feature.

True or **false**: With enough hidden nodes, a Neural Network can separate any data set.

Problem 4: K-Means Clustering

(a) Execute the k-means algorithm (as accurately as possible) on the following data. Dots indicate data points; the two squares show the initial cluster centers. In your plots, draw the cluster centers as squares and also show the decision boundary that defines each cluster. Use as many pictures as you need for convergence, or until all are full. Also, explain your process and placement during the algorithm.

Explanation:

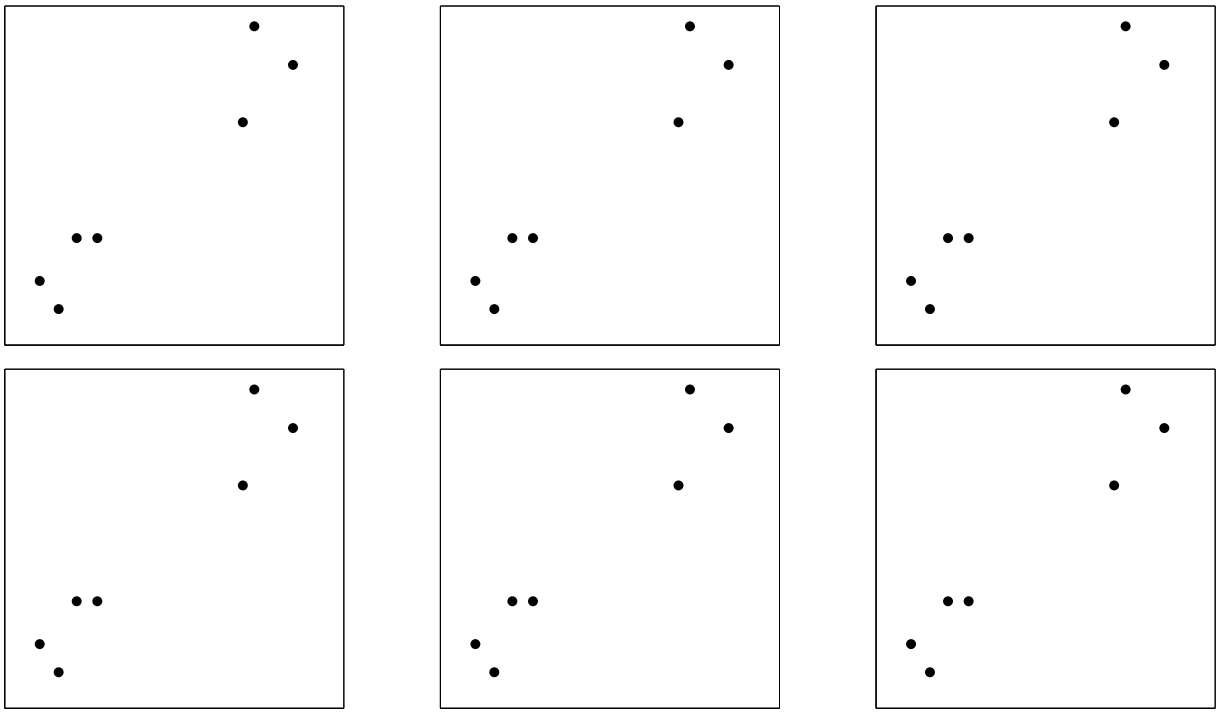


(b) Would increasing the number of clusters increase or decrease the value of the cost function?

(c) Describe how the K-Means algorithm might be used as part of a classification task.

Problem 5: Agglomerative Clustering

(a) Now execute the hierarchical agglomerative clustering (linkage) algorithm on these data points using "single linkage" (minimum distance) for the cluster scores. Stop when finished or after 6 steps, whichever is first. Show each step separately in a panel.

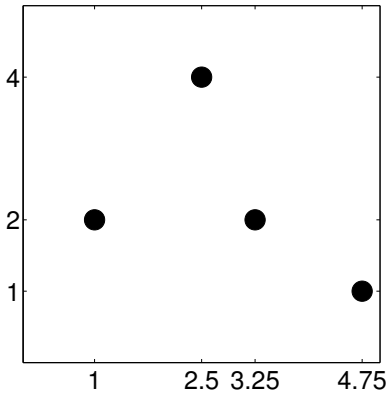


(b) How does "complete" linkage differ from "single linkage"? Would it behave differently on the above data, and if so, how?

(c) Describe (briefly) how we might decide on the correct number of clusters from this procedure.

Problem 6: Cross-validation

Consider a regression problem for predicting the following data points, using the k-nearest neighbor regression algorithm from class and the homework to minimize mean squared error (MSE). (Note: if you like, you may leave an arithmetic expression, e.g., leave values as "$(.6)^2$".)



(a) For $k = 1$, compute the training error on the provided data.

(b) For $k = 1$, compute the leave-one-out cross-validation error on the data.

(c) For $k = 3$, compute the training error on the provided data.

(d) For $k = 3$, compute the leave-one-out cross-validation error on the data.

Problem 7: Perceptrons and VC Dimension

In this problem, consider the following perceptron model on two features:

$$\hat{y}(x) = \text{sign}(\, w_0 + w_1 x_1 + w_2 x_2 \,)$$

and answer the following questions about the decision boundary and the VC dimension.

(a) Describe (in words, with diagrams if desired) the possible decision boundaries that can be realized by this classifier

(b) What is its VC dimension?

Now suppose that I also enforce an additional condition on the parameters of the model: that **at most two** of the weights w_i are **non-zero** (so, equivalently, at least one weight is zero). Note that the training algorithm can choose which parameter is zero, depending on the data.

(c) Describe (in words, with diagrams if desired) the decision boundaries that can be realized by this classifier

(d) What is its VC dimension?