University of Washington Bothell
Computing & Software Systems

Course title:	Introduction to Machine Learning
Course number:	CSS 581
Term:	Winter 2014
Instructor:	Jeff Howbert

Exercises 4

Date assigned:	Jan. 17, 2014
Date due:	Jan. 22, 2014

IDM refers to our textbook, "Introduction to Data Mining", by Tan, Steinbach, and Kumar.

1) [3 points] Exercise 2(a) in Section 4.8 of IDM (p. 198). Show your work. For this exercise, you are essentially computing the Gini index for a single node. Review pp. 158-159 in IDM and slides 31-32 from Lecture 4.

2) [5 points each] Exercises 2(c) and 2(e) in Section 4.8 of IDM (p. 198). Show your work. For these exercises, you are computing the combined, weighted Gini index for two or more child nodes. Review pp. 160-162 in IDM and slides 33-35 from Lecture 4.

3) [2 points] In Question 2), you computed the Gini index for two attributes, gender and shirt size. Which attribute is better for splitting? Briefly explain your answer. Review slide 30 from Lecture 4 if necessary.

4) [15 points] This Question explores the effect on classification accuracy of pruning in a small decision tree. All the code you need is exemplified in Part B of matlab_demo_04.m.

Do the following:

i. Divide the Fisher iris dataset into training and test sets exactly as done in Part B.
ii. Train a decision tree on the training set exactly as done in Part B.
iii. For each possible pruning level in the trained tree, evaluate the classification accuracy of the tree on both the training set and the test set.
    a. You can set the pruning level as an argument in `eval()`.
    b. The code in Part B shows one way to find the maximum pruning level.
    c. Remember that the minimum pruning level is 0.
    d. For each pruning level and evaluation dataset, output something like this to the console:
    `pruning level = ppp  correct predictions on ttt set : xxx / yyy`
    e. Iterate through the pruning levels using a `for` loop, where the index of the loop is the pruning level.
iv. In your answer document include:
    a. The multiple lines of output from iii.d. above.
    b. All the code you used for the exercise.

c. ***DO NOT*** generate or turn in confusion matrices for these evaluations.

5) [15 points] This Question explores the effect on classification accuracy of changing the training parameters on a small decision tree. All the code you need is exemplified in Part B of matlab_demo_04.m.

Do the following:

i. Divide the Fisher iris dataset into training and test sets exactly as done in Part B.

ii. Train a decision tree several times on the training set exactly as done in Part B, but with certain variations on the settings of the training parameters:
   a. `minparent = 10, minleaf = 1` (default values)
   b. `minparent = 2, minleaf = 1` (default values)
   c. `minparent = 10, minleaf = 5` (default values)
   d. `minparent = 50, minleaf = 1` (default values)
   e. `minparent = 100, minleaf = 1` (default values)

iii. Train a decision tree several more times on the training set exactly as done in Part B, but using each of the three possible values of `'splitcriterion'` (keep `minparent = 10, minleaf = 1`).

iv. Prior to running each variation above, make sure you reset `RandStream` with a `seed = 1`.

v. For each of the above eight variations on the settings, output the classification accuracy and confusion matrix for both training and test sets to the console. You can use the evaluation and output code in Part B exactly as is.

vi. In your answer document include:
   a. The output from v. above.
   b. Brief explanations for the differences in accuracy (or lack thereof) seen between:
      i. Variation ii.a and variation ii.b
      ii. Variation ii.e relative to variations ii.a, ii.b, ii.c, and ii.d.
      iii. The three `'splitcriterion'` variations under iii.
   c. ***DO NOT*** include any of your code for this Question.