University of Washington Bothell
Computing & Software Systems

Course title:	Introduction to Machine Learning
Course number:	CSS 490 / 590
Term:	Winter 2012
Instructor:	Jeff Howbert

Exercises 5

Date assigned:	Feb. 2, 2012
Date due:	Feb. 7, 2012, 10:00 PM

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

1) [7 points] The accuracy of a 1-nearest neighbor classifier will always have the same value, regardless of the dataset.

a) What is that value?
b) Explain why that value is always obtained.

Question 2) explores the effects of varying the number of nearest neighbors on classification accuracy in a k-nearest neighbor classifier. You will apply the accompanying script, `knn.m` (see link in Schedule) to the MNIST handwritten digits dataset (`mnistabridged.mat`). The only thing you need to do to answer this Question is change the value of k in the script. I recommend you save the modified script and run it from the command line, rather than executing by highlighting + F9.

2) Run `knn.m` with values of k = 1, 3, 5, 7, 11, 17, and 27. Each run will take about 2 minutes.

a) [5 points] Create a table that lists the number of correct predictions on the test set for these values of k.
b) [3 points] What value of k (among those run) gives the best overall accuracy? Which gives the worst?
c) [5 points] Recall that the diagonal of a confusion matrix tells how many of each class were correctly predicted. Let's compare the predictions of the overall best and worst k in more detail. Look at the diagonals of the confusion matrices for these two values of k. Which class (i.e. digit) suffered the worst loss in accuracy going from best to worst k?
d) [5 points] Now let's see what we can learn from the off-diagonal entries. In a confusion matrix generated by MATLAB's confusionmat() function, the rows correspond to true class labels, and the columns to predicted labels. Look at the confusion matrix for the worst value of k. For the digit that suffered the worst loss of accuracy (i.e. answer from part c)), what other digit was it most frequently misclassified as?

3) [10 points each part, except 5 points for part (e)] Exercises 7(a), 7(b), 7(c), 7(d), and 7(e) in Section 5.10 of IDM (p. 318). Show your work. Some clarifications:

- For 7(a) there are a total of 12 conditional probabilities, not 6, because there are two possible values, 0 and 1, for each of the attributes A, B, and C. Show your results for all 12:
  - p(A = 0 | +)
  - p(A = 0 | -)
  - p(A = 1 | +)
  - p(A = 1 | -)
  - p(B = 0 | +)
  - p(B = 0 | -)
  - p(B = 1 | +)
  - p(B = 1 | -)
  - p(C = 0 | +)
  - p(C = 0 | -)
  - p(C = 1 | +)
  - p(C = 1 | -)
- For 7(b) you need to also calculate two class prior probabilities before you can use the naïve Bayes approach. In addition to the answer requested in the book, show your results for the two class priors:
  - p(+)
  - p(-)
- For 7(c): same as 7(a).