**CSC529: Advanced Data Mining**

**Assignment #1 (review homework)**

**Total: 44 points**

**Assigned: Monday, January 15th, 2018**

**Due: Monday, January 29th, 2018, by midnight**

The goal of this assignment is to review three main classification approaches used in data mining: Naïve Bayes, decision trees, and k-nearest neighbors.

The dataset used for this assignment is the Wine Recognition Dataset that can be downloaded it from: <https://archive.ics.uci.edu/ml/machine-learning-databases/wine/> The data are the results of chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. There are 178 instances in total with 59, 71, and 48 instances in class 1, class 2, and class 3, respectively.

**Problem 1 (10 points): Data Visualization**

1. (2 points) Select randomly two variables and plot the data using different colors/signs to plot the points belonging to different three classes. What can you say about the class separability in this space?
2. (8 points) Repeat part a. but this time using two variables that you found to be the most relevant for the classification process. Explain the approach you applied to select these two variables and include the analysis you performed in your answer.

**Problem 2: (30 points) Classification**

First divide the data using the holdout partition (67% training and 34% testing).

1. (10 points) Create a decision tree model and report:
   * 1. What are the parameters that you used for building the decision tree and which parameter configuration gave you the best performance?
     2. For the best parameter configuration, report
        1. Accuracy on training data
        2. Accuracy on testing data
     3. Are there any assumptions on the distributions of the data that have to be made in order to apply a decision tree model? Explain your answer.
   1. (10 points) Create a Naïve Bayes classifier and report
      1. Accuracy results on both training and testing data
      2. Naïve Bayes classifier makes certain assumptions about data. State these assumptions and check if these assumptions are satisfied.
   2. (10 points) Create a K-nearest neighbor classifier and report
      1. Accuracies results on both training and testing data for values of k equal to 1, 3, and 5.
      2. What was the similarity measure you used? Was there any need to normalize the data? Explain your answer.

**Problem 3: (4 points) Basic data mining concepts explained via pictures**

Sometimes it is easier to use pictures to explain machine learning and data mining concepts. You can review examples of such pictures on the KdNuggets website at:[**http://www.kdnuggets.com/2014/03/machine-learning-7-pictures.html**](http://www.kdnuggets.com/2014/03/machine-learning-7-pictures.html)

Based on these pictures describe:

1. Why lower training error is not always a good thing?
2. Why collectively relevant features may look individually irrelevant?
3. Why irrelevant features hurt kNN, clustering, and other similarity based methods?
4. What does it mean Occam razor?

**Matlab hints:**

* Classification examples (decision trees and naïve Bayesian) at:

<http://www.mathworks.com/examples/statistics/mw/stats_featured-ex39502418-classification>

* Classification examples (k-nearest neighbors) at: <http://www.mathworks.com/examples/statistics/2163-find-nearest-neighbors-using-a-custom-distance-metric>

**Submission Instructions**

1. Answer the problems and write your answers in a Word/Pdf document file. Include your all code in the Word/PDF file as an Appendix.
2. Submit your file online at <http://d2l.depaul.edu>
3. Check that you submitted correctly your file and keep a copy of all your submissions!
4. If you have questions about the homework, email me BEFORE the deadline.
5. Late submissions are allowed with a 5%, 10%, and 15% penalty for a one day, two days, and three days, respectively.
6. No late work will be accepted after three days since the assignment was due.