CS 5593 – Section 001 - Fall 2020 - Dr. Le Gruenwald
HOMEWORK 5
Assigned: 11/18/2020; Due: 11/30/2020 by 2:00 PM to the class website
Maximum Points: 100 points

Notes:

- **Homework is individual work; it must be done by you only, no collaboration with anyone else is allowed.**
- **Late homework will be accepted on the class website until 11:59 PM on the date following the due date with 5% (of the maximum points) penalty. Late homework submitted after this time will not be graded.**

Problem 1 (60 points) (your answers to this problem must be typed): Using a programming language of your choice (e.g., C, C++, Java, Python, or R), write a program to implement the following tasks for univariate outlier detection. Except for Task (a), you must implement all the other tasks from scratch (i.e. in your code, you must not use any existing function/package that already implements these tasks or a part of them).

(a) Use a graphical tool to demonstrate and explain whether or not the following dataset is approximately normally distributed:
{152.36, 130.38, 101.54, 96.26, 88.03, 85.66, 83.62, 76.53, 74.36, 73.87, 73.36, 73.35, 68.26, 65.25, 63.68, 63.05, 57.53}

(b) Write a function to implement the "Parametric Method I: Outlier Detection for Univariate Outliers based on Normal Distribution" discussed in the Lecture Topic 8 "Anomaly Detection," where outliers are those values that do not lie within *w* standard deviations from the mean.

(c) Write a function to implement the "Parametric Method II: Grubbs Test" to detect outliers.

(d) Write a function to implement the k^{th}-nearest neighbor outlier detection where Euclidian distance is used to compute the distance between two data points.

(e) Run the function implemented in part (b) to detect outliers in the dataset given in Task (a) for each of the three cases (*w* = 1, 2, and 3):
Then run the function implemented in Task (c) to detect outliers in the same dataset with $\alpha = 0.05$.
Then run the function implemented in Task (d) to detect outliers in the same dataset with *k* = 2 and 3. For each value of *k*, output a list of data values with their corresponding outlier scores.

(f) Provide an in-depth comparison analysis of the results you obtained in Task (e).

Problem 2 (20 points) (your answers to this problem can be hand-written, but must be readable in the submitted PDF version): Consider the following training set for predicting whether there is traffic congestion in the morning on a highway for a particular day. There are 100 examples in the training set, with 40% positive (+) (congestion) and 60% negative (−) (no congestion) examples.

Accident	Weather	Construction	Number of positive training examples	Number of negative training examples
no	good	no	5	30
no	good	yes	10	20
yes	good	no	10	5
yes	bad	no	10	5
yes	bad	yes	5	0

(a) Compute the class conditional probabilities P(accident=yes|+), P(accident=yes|−), P(weather=good|+), P(weather=good|−), P(construction=yes|+), and P(construction=yes|−). Show your work.

(b) Use the class conditional probabilities computed in Task (a) to predict the class label of a test example with the following feature set: **(Accident = no, weather = bad, construction = yes)** by applying the naıve Bayes classifier. Show your work.

Problem 3 (20 points) (your answers to this problem must be typed): Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how EACH of the following techniques, clustering, classification, association rule mining, and anomaly detection, can be applied.

Notes on submission: Submit two files to the class website on Canvas. The first file is a complete PDF file named *Your Last Name_Your First Name_HW5.pdf* containing your answers to ALL the three problems. For Problem 1, this complete PDF file needs to also contain the source code of the program that you wrote to accomplish all the tasks, the screenshots of your program executions and the required output. Appropriate in-line comments must be included in the program, and appropriate labels must be provided for the required output. The second file is a program file named *Your Last Name_Your First Name_HW5.extension* containing the program for Problem 1 as we will test your program for correctness, where *extension* is the programming language you used to implement your program (e.g. r for a program written in R, java for a program written in Java, and py for a program written in Python). Failure to submit this program file will result in a zero grade for your Problem 1.