CS 5593 – Section 001 – Fall 2023 - Dr. Le Gruenwald
HOMEWORK 5
Assigned: 11/15/2023; Due: 11/27/2023 by 3:00 PM to the class website on Canvas
Maximum Points: 100 points

Notes

- **Late homework will be accepted on the class website on Canvas 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.**
- **ACADEMIC INTEGRITY: Homework is individual work; it must be done by you only; no collaboration with anyone is allowed. The use of generative AI tools (including ChatGPT, Bard, Bing Chat, and other AI writing and coding assistants) is not allowed in all homework questions, except for those questions where the use of these tools is explicitly required. Violations of any of these rules will be considered academic misconduct and will result in action as specified in the Academic Integrity Code at The University of Oklahoma: http://www.ou.edu/integrity. Consult also the following web page for a Student's Guide to Academic Integrity at The University of Oklahoma: http://www.ou.edu/integrity/students.**

Problem 1 (15 points) (your answers to this question must be typed): Using ChatGPT, generate prompts asking it to explain how outlier detection is used in the Knowledge and Data Discovery (KDD) process. Also, request an explanation of the negative impacts that outliers would have on the KDD process if outlier detection algorithms were not implemented. Then ask it to provide three application examples of how outlier detection is applied within the KDD process in three different industries – pharmaceutical, aerospace, and computers. You can try at most three prompts. From the answers given by ChatGPT, elaborate on whether the answers are sufficient or insufficient. If you only state that ChatGPT's answers are sufficient or insufficient without providing detailed explanations, you will get a zero score for this question. If any answer you think is insufficient, modify it to ensure a comprehensive understanding of how outlier detection is used in the KDD process, what negative impacts that outliers would have on the KDD process, and how outlier detection is used in three applications from the three given industries. Show your prompts, ChatGPT's answers, your detailed explanations, and your modifications (if any).

Problem 2 (65 points) (your answers to this question 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 the *w* standard deviation from the mean.

(c) 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.

(d) Run the function implemented in Task (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 k = 2 and 3. For each value of k, output a list of data values with their corresponding outlier scores.

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

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 data set with PlayTennis as the binary class label (Yes or No). Compute the necessary probabilities and apply Naïve Bayes Classifier to predict the class label of the following test example:

X= *(Outlook= Sunny, Temperature = Cool, Humidity = High, Wind=Strong)*.

Show your work.

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Notes on submission: Submit two files to the class website. The first file is a complete PDF file named *HW5_Your Last Name_Your First Name.pdf* containing your answers to ALL the three problems. For Problem 2, 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 *HW5_Your Last Name_Your First Name.extension* containing the program for Problem 2 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 and py for a program written in Python). Failure to submit this program file will result in a zero grade for your Problem 2.