CS464 Introduction to Machine Learning
Spring 2024
Homework 1

Due: March 12, 2024, 11:55 PM

Instructions

- Submit a soft copy of your homework of all questions to Moodle. Add your code at the end of your homework file and upload it to the related assignment section on Moodle. Submitting a hard copy or scanned files is NOT allowed. You have to prepare your homework digitally (using Word, Excel, Latex etc.).
- This is an individual assignment for each student. That is, you are NOT allowed to share your work with your classmates.
- For this homework, you are allowed to use any programming language you would prefer.
- To submit the homework file, please package your file as a gzipped TAR file or a ZIP file with the name `CS464_HW1#Firstname_Lastname`.

  As an example, if your name is Sheldon Cooper and you are from Section 1 for instance, then you should submit a file with name `CS464_HW1#Sheldon_Cooper`. Do NOT use Turkish letters in your package name.

  Your compressed file should include the following:

  - Report: Your answer to each question. (Need to be in pdf format)
  - p2main.*: The source code for question 2. (* mean you are allowed to use any programming language format)
  - q3main.*: The source code for question 3. (* mean you are allowed to use any programming language format)
  - README.txt : You must also provide us with a README file that tells us how we can execute/call your program. README file should include which parameters are the default values, what is the terminal command to execute your file and how to read the outputs.
- You are NOT allowed to use any machine learning packages, libraries or toolboxes for this assignment (such as scikit-learn, tensorflow, keras, theano, MATLAB Statistics and Machine Learning Toolbox functions, e1071, nnet, kernlab etc.).
- In the attached file, assets related to each question are given inside the directory with the question name.
- Your codes will be evaluated in terms of efficiency as well. Make sure you do not have unnecessary loops and obvious inefficient calculations in your code.
- If you do not follow the submission routes, deadlines and specifications (codes, report, etc), it will lead to significant grade deduction.
- For any questions regarding this homework, contact to m.poorghaffar@bilkent.edu.tr (Masoud poorghaffar Aghdam).

1 Programmers and Energy drink [10 pts]

Question 1.1 [**2 pts**] In a study of programmers' typing speed, 60% of participants consumed the energy drink before the test. Among those who drank the energy drink, 70% showed an improvement in typing speed. Among those who didn't drink the energy drink, 40% showed an improvement. What is the probability that a randomly selected participant showed an improvement in typing speed?

Question 1.2 [**2 pts**] If it is known that 50% of the participants who drank the energy drink were experienced programmers and 80% of those experienced programmers showed an improvement in typing speed, what is the probability that a randomly selected participant who showed improvement after drinking the energy drink was an experienced programmer?

Question 1.3 [**3 pts**] Among the participants who didn't drink the energy drink, 60% were beginners, and 41% of them had no improvement. If 56% of the beginners showed an improvement in typing speed, what is the probability that a randomly selected participant who showed improvement was a beginner?

Question 1.4 [**3 pts**] It is observed that 70% of experienced programmers and 30% of beginner programmers who demonstrated improvement consume caffeine daily. Furthermore, approximately 20% of participants who did not exhibit any improvement consume caffeine daily. Given this scenario, what is the probability that a participant who showed improvement is consuming caffeine daily?

2 Height of giants!? [20 pts]

In an imaginary world, there exists a species of giants. The height of these giants is varying. The goddess who created these giants claims that the population height follows a normal distribution with mean 50 and variance 5. However, recently some giant scientists studied the heights of their own species. They estimated mean and variance using maximum likelihood estimation (MLE), but their results are different from what goddess claims; they are thinking their goddess is not straightforward with them. As CS464 students, you are asked to resolve this issue.

Question 2.1 [**15 pts**] We are given a list of heights that giant scientists used. You will repeat the procedure of estimation just like giant scientists and validate their claims. Code your solution and explain your reasoning.

Instructions

- Utilize the dataset provided in the file `giant_heights.txt`, containing the heights used by the giant scientists in their study.
- Employ Maximum Likelihood Estimation (MLE) to calculate both the mean and variance of the giant population heights. If the difference is lower than 0.01 you can say it is a match and consider it as a computer error.
- Compare the estimated parameters with the goddess's claims to validate her claims. If they are not equal, the differences should be included.

Submission Requirements

- Implementation: Your script (source code) which does the computations.
- Methodology Explanation: A detailed explanation of the formulation process for Maximum Likelihood Estimation (MLE). Justify the reason behind your approach and explain how you computed the mean and variance from the dataset, and in addition, with your code output add it to the report.

Question 2.2 [2 pts] Plot both of the distributions. That is the distribution goddess claims and the distribution giant/you inferred. (For plotting you are allowed to use third party libraries)

Question 2.3 [3 pts] What happens if the giant scientists increase the size of the sample they are studying? Give a detailed explanation for probable scenarios.

3 UTI Diagnosis [70 pts]

UTI diagnosis refers to the process of identifying and confirming the presence of a urinary tract infection (UTI) in an individual. A UTI is an infection that occurs anywhere in the urinary tract, including the kidneys, ureters, bladder, and urethra, and it can be caused by bacteria, viruses, or fungi.

A clinic in Northern Mindanao, Philippines, collected UTI data from April of 2020 to January of 2023 in a dataset [1]. You are asked to create a naive Bayes classifier for diagnosing UTI.

Instructions

- You can find the dataset attached to this homework which contains four files:
  - `x_data_*`: The files starting with this pattern contain the features of each record in rows.
  - `y_data_*`: The files starting with this pattern contain the labels corresponding to each row in the features file.
  - `*train`: These files have records for training.
  - `*test`: These files have records for testing.
- The provided dataset has been simplified from the original version. Therefore, the evaluation will be conducted using the attached dataset.
- To learn about dataset, and its features you can read `README.txt` in assets related to this question.

Question 3.1 [50 pts] Train and test a Naive Bayes classifier using the preprocessed data obtained from the previous question. Then test the trained classifier on the test data, and report the testing accuracy and confusion matrix in addition with the number or correct/incorrect predictions. To estimate model parameters use the MLE estimator that you used in question 2. You need to submit both code and output. Read the global instructions, no libraries!

Question 3.2 [5 pts] How many parameters do we need to estimate for this model? Write both names and counts.

Question 3.3 [5 pts] If a patient's test result is positive, what is the probability that the patient actually has the disease? Out of all the patients who truly have the disease, what proportion did your model correctly identify as having the disease? Support your answer with empirical results.

Question 3.4 [5 pts] Find the most and least effective feature in positive diagnosis.

Question 3.5 [5 pts] The ratio of the classes in a dataset is close to each other, it is a called "balanced" class distribution if not it is skewed. What is the percentage of POSITIVE diagnosis in the train dataset. Is the training set balanced or skewed towards one of the classes? Do you think having an imbalanced training set affects your model? If yes, please explain how it affects and propose a possible solution if needed.

References

1. Urinalysis Test Results.
https://www.kaggle.com/datasets/avarice02/urinalysis-test-results/data