Assignment 2

CSCI 6515 - Machine Learning for Big Data – Fall 2023

Posting date: Oct 16, 2023, 9 AM (Halifax Time)

Due Date: Oct 30, 2023, 11.59 PM (Halifax Time)

Background Information

Heart disease is dangerous and responsible for many deaths every year. It describes any condition that affects the heart, leading to complications, such as heart failure, heart attack, and stroke. Many risk factors contribute to the development of heart diseases, such as age, smoking, and obesity. Tests and exams are performed to diagnose the disease, which includes exercise or stress tests and electrocardiograms. Motivated by this fact, we try to integrate machine learning research with Public Health Dataset to assist in the identification of the most influential factors or exams that help in heart disease diagnosis. We explore the anonymized dataset of Cleveland, Hungary, Switzerland, and Long Beach V from 1988. Simultaneously, we investigate if ML models may assist doctors in the diagnosis by providing a list of patients that may need more investigation. Finally, we decided to do a few experiments. We will create two different models, one using logistic regression and another using Naïve Bayes to classify the patients that may have heart disease assisting the doctors in the analysis.

To access and collect the heart disease data follow the link: Heart Disease Dataset | Kaggle (https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset)

Note that you may use open-source libraries such as SK-learn, NumPy, and Pandas. If you are interested in using any other available library, you can consult with TAs.

Your Tasks:

(Points on the right side of each question indicate the weight of the individual question)

[1] *Your first task is doing some research on the dataset.*

- **a)** Process or pre-process the heart disease data. Show the outputs of the pre-processing phase. **3**
- **b)** Visually analyze the data after processing or pre-processing. **2**
- **c)** Explain the method(s) you chose for processing/preprocessing and provide a descriptive analysis to justify your choice. You can resort [1(b)] to justify. **3**

[2]

- **a)** Apply Naive Bayes classifier on your dataset. **3**
- **b)** Evaluate the performance for disease classification. **2**

[3]

a) Repeat task [2] by applying logistic regression with gradient descent. **3**

b) Compare the performance of task 3[a] with batch gradient descent, mini-batch gradient descent, and stochastic gradient descent. **3**

c) Display training, validation, and test loss over 30 epochs. **2**

[4] *Compare the two models in task [2] and [3], and provide a detailed analysis by answering the following questions*

a) Which evaluation metric did you use? Explain your reasoning. **3**

b) Which model provided the highest score? Explain the reasons. **3**

[5] *Suppose that you have the following training data.*

RID	age	income	student	credit	C_i: buy
1	youth	high	no	fair	C_2: no
2	youth	high	no	excellent	C_2: no
3	middle-aged	high	no	fair	C_1: yes
4	senior	medium	no	fair	C_1: yes
5	senior	low	yes	fair	C_1: yes
6	senior	low	yes	excellent	C_2: no
7	middle-aged	low	yes	excellent	C_1: yes
8	youth	medium	no	fair	C_2: no
9	youth	low	yes	fair	C_1: yes
10	senior	medium	yes	fair	C_1: yes
11	youth	medium	yes	excellent	C_1: yes
12	middle-aged	medium	no	excellent	C_1: yes
13	middle-aged	high	yes	fair	C_1: yes
14	senior	medium	no	excellent	C_2: no

We would like to determine if Jack can buy a computer or not. The following information is available about him: age = middle-aged, income = medium, student = yes, credit = fair. Estimate whether or not he can buy a computer according to Naive Bayes Classification.

a) Show the probability of all the components of Bayesian approach. **2**

b) Given the above dataset, what is the prediction of the Naïve Bayes Classifier about Jack? Why? **1**

Supporting Document:

In order to learn more about how to run your experiments [1-4] and write the report, check out this paper (https://scholarworks.lib.csusb.edu/cgi/viewcontent.cgi?article=2924&context=etd).

Instructions:

1. Completing Assignment:

- Use **Jupyter Notebook** for Assignment 2.
- Follow the provided *Jupyter notebook* **template with the assignment in Brightspace's**. You can create as many code and markdown cells necessary for program executions and answering the questions following the template format.
- **Retain all the cell outputs** of Jupyter notebook.
- If there is any manual calculation instructed in question, do it in notebook or attach images of the calculation in the notebook. For safety purposes, you can attach images in the zip file as well.
- Include a list of all references used in this assignment in the reference section. Follow *APA referencing format* for the references.

******Note:* Failing to follow the template will result in a deduction of up to 2 points**
******Note:* Failing to retain cell outputs raises suspicions of cheating and will result in point deductions.**

2. Submission Guideline:

i. **Preparing File**: Your submission should be a **ZIP (.zip)** file containing two files**:**

1. A Jupyter notebook file **(.ipynb)** with cell execution results**.**
2. A **PDF** of the Jupyter notebook with cell execution results.

ii. **Naming the Submissions:** As part of the final printout, name your files as follows:

a. *Naming **Individual** Files*

i. **Jupyter Notebook:** A2_<banner_id>.ipynb
ii. **PDF file:** A2_<banner_id>.pdf

b. *Naming the **ZIP** file*

- A2_<your_first_name>_<your_banner_id>.zip

**Here A2 means, Assignment-2.*

******Note:* Failure to follow the naming convention may result in a deduction of up to 2 points.**
****** Note:* Failing to provide both required files (.ipynb and .pdf) will result in a score of zero (0) for the assignment. Any file other than notebook and/or pdf will result in 0 points. No word file or its pdf will be accepted.**

iii. **Submission Process: S**ubmissions should be made through Brightspace, adhering to the due date and time specified time.

3. **Late Submission Policy:** Late submission will follow the following policy-

- Day 1 (within first 24 hours) – 15% reduction of the total marks
- Day 2 (24-48 hours) – 20% reduction of the total marks
- Day 3 (48-72 hours) – 25% reduction of the total marks

****** Note:*** ***No submissions will be accepted after the specified time; resulting in zero (0) points for the assignment.***

****** Note: Any suspicion of plagiarism will be subjected to penalty and may lead to a score of 0 points***

****** NOTE: Submissions that do not follow the instructions and/or format of answering will not be graded.***