CMPE 462 - Spring 2021
Assignment 2

Introduction

This assignment consists of 2 parts.
The first one is about logistic regression and the second part is about naive bayes.

Part1

In Part 1 of this assignment, you will implement **Logistic Regression** (LR) from scratch. We will give you the dataset as a csv file. Each line is a sample and each column is a feature. The last column is the class value.

The dataset file is vehicle.csv. It is the Vehicle Silhouettes dataset from UCI Machine Learning Repository which includes attributes for 4 classes. You can find more information on the website https://archive.ics.uci.edu/ml/index.php. For this assignment we will be using classes "saab" and "van" and all features. You will implement LR with 5-fold cross-validation.

The steps of Part1 are below:

- Step1: Implement LR with batch gradient descent (update weights after a full pass over data) and apply on the dataset.
- Step2: Implement LR with stochastic gradient descent (update weights after mini batches) and apply on the dataset.

For each step, save the loss value at each iteration and after convergence plot the loss over iterations graph. Try 3 different (small, medium and big) step sizes in range 0 and 1. Provide plots for each step size.

You can use mathplot library. But special functions or libraries like scikit-learn is forbidden.

In your assignment report, include results of your runs for each step. Place the plots and discuss over them. Discuss the effect of gradient descent model on convergence. Compare your runs with different step sizes. Also discuss the number of iterations and time you need for each step size.

Part2

In Part 2 of this assignment, you will focus on **Naive Bayes**.

Consider the table below. Use Naive Bayes to classify the test sample in the last row. Include your full solution in the assignment report. (Name column is informative.)

Name	**GiveBirth**	**CanFly**	**LiveInWater**	**HaveLegs**	**Class**
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals
test	yes	no	yes	no	**???**

Base Environment

You will be implementing your code with Python 3.6.
You need to create a python virtual environment with Anaconda for your project. After installing Anaconda, a base environment can be created with below commands:

```
conda create -n 462assignment python=3.6
conda activate 462assignment
```

While you keep working on your models, you will need to import additional libraries. List these libraries in a requirements.txt file. State any special versions if needed. A sample requirements file can be as below:

```
scikit-learn >= 0.22.2
scipy
pandas
sentencepiece==0.1.91
```

For grading, we will load your requirements with the command below:

```
python3 -m pip install -r requirements.txt
```

Before submission, test your code on a clear new conda environment by installing additional libraries from your requirements file. Because, there will be penalty if your code doesn't run like this.

Grading Details

The assignment will be graded over 100 points. You will be graded for your code and report.

- 60 points for report
  - 20 points for Part1
  - 40 points for Part2
- 40 points for code (Part1)
  - 20 points for step 1
  - 20 points for step 2

We will run your code on a clear new conda environment. First we will load your requirements.txt file. Then we will test your code with below commands:

- Part1

```
python3 assignment2.py part1 step1
python3 assignment2.py part1 step2
```

Consider second command, you will run LR with stochastic gradient descent.

Submission Details

This is an individual assignment. Your code should be original. Any similarity between submitted assignments or to a source from the web will be accepted as cheating.

If you have any further questions, send an e-mail to the course page on Piazza.

- The deadline for submitting Assignment 2 is **May 18, 2021 - 23:59**.
- There will be 2 submissions open for this assignment.
- Submission 1:
  - You should submit 3 items:
    - your Python script, assignment2.py
    - your requirements.txt file, blank file if no additional library is needed
    - your assigment report in pdf, 462_assignment2_<studentid>_report.pdf, example 462_assignment2_20181123456_report.pdf
  - You should compress all submission items in a zip file with name as 462_assignment2_<studentid>.zip, example 462_assignment2_20181123456.zip
  - The zip will be submitted on Moodle.
- Submission 2:
  - You should also submit your reports in Turnitin submission on Moodle.