ECE 473 : Introduction to Artificial Intelligence

Assignment 2

Instructions on programming assignments

You are asked to modify the code in `hw2_submission.py` between

```
###################################################################
###################################################################
# BEGIN_YOUR_CODE
# HINTS OR INSTRUCTIONS

pass;

# END_YOUR_CODE
###################################################################
###################################################################
```

- Please use Python3 to do all assignments in the course.
- You are allowed to make your own helper functions outside the skeleton functions.
- Do not change other than `hw2_submission.py`. You are only allowed to use `numpy` package to compute and construct functions. Do not use `scikit-learn, pytorch`, or `tensorflow` etc to implement your functions.
- You can test your code with test functions in `test.py`.
- For your reference, you can compare your own classifiers with the ones in sklearn package that is implemented in `test.py`.
- Do not change `seed` in `test.py`.
- You need to get no more than 5% difference of accuracy from the reference to get full score.
- Only submit `hw2_submission.py`

Problem 1: Logistic Regression

In this problem, you will implement `class Logistic_Regression` and test it on 'breast cancer wisconsin' dataset. The dataset has 30 features including the information about cell nucleus, such as radius, texture, concave points, and etc. The goal is to classify rather the cell (data point) is malignant ($y = 1$) or benign ($y = 0$).

1. Implement function `Logistic_Regression.fit`
2. Implement function `Logistic_Regression.sigmoid`
3. Implement function `Logistic_Regression.loss`
4. Implement function `Logistic_Regression.predict`

Problem 2: Spam Naive Bayes Classifier

Similar to the previous problem, you will implement Naive Bayes Classifier to classify a bunch of emails that they are rather spam or not. You can download the dataset from the link "Spam Dataset" in the BrightSpace. Locate the `ham, spam` folders in `spam_dataset` directory of your homework folder. This is real email data which contains spam emails and ham (non-spam) ones from the Enron Corporation after the company collapsed.
(Hint: You will count the number of occurrence of each word to compute likelihood of each word given a spam/ham label.)
(Note: Since this dataset is a real dataset of emails, it contains real spam messages. Your anti-virus may prune some these emails because they are spam. Let your anti-virus prune as many as it wants. This will not affect our code as long as there are some spam and ham messages still there.)
(Note: Test accuracy above 70% will sufficient for this problem)

1. Implement function `Spam_Naive_Bayes.get_word_counts`
2. Implement function `Spam_Naive_Bayes.fit`
3. Implement function `Spam_Naive_Bayes.predict`