Daniel Bao
CS171 Sec 02 - Machine Learning
015707445
*Problems 1-9*
HW1 Logistic Regression

**1. What is the train set Accuracy as printed by your code?**
Accuracy: 0.9818344920385736

**2. What are the train set Precision, Recall and F1-measure (also called F1-score) of the positive (spam) class as printed by your code?**
Precision: 0.8902627511591963
Recall: 0.9829351535836177
F1 Score: 0.9343065693430657

**3. What are the train set Precision, Recall and F1-measure (also called F1-score) of the negative (ham) class as printed by your code?**
Precision: 0.9816679576555641
Recall: 0.9816679576555641
F1 Score: 0.9816679576555641

**4. What is the confusion matrix for the train set as printed by your code?**
True Positive (Spam) : 576
False Positive (Spam) : 71
True Negative (Ham) : 3802
False Negative (Ham) : 10
[[3802, 71],
[10, 576]]

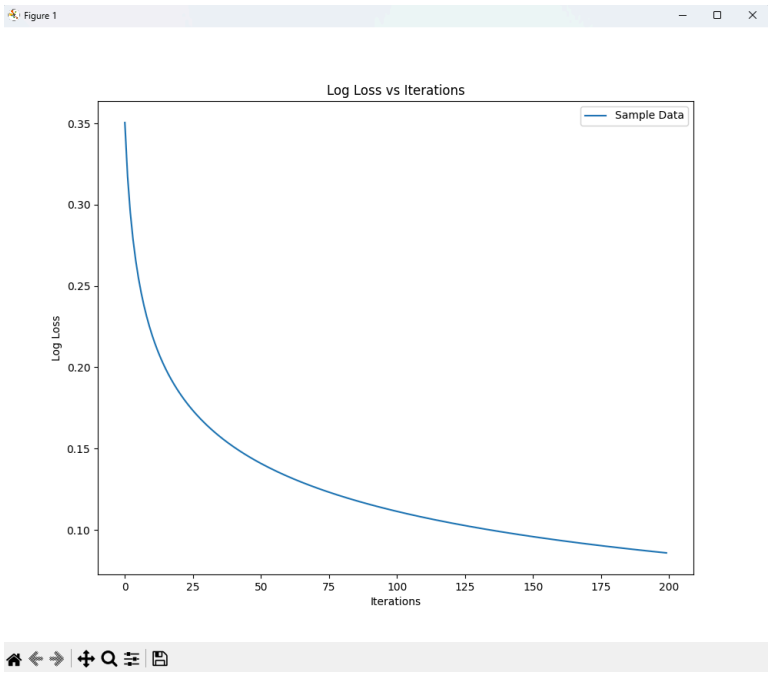
**5. What is the test set Accuracy as printed by your code?**
Accuracy: 0.9560538116591928

**6. What are the test set Precision, Recall and F1-measure (also called F1-score) of the positive and the negative class as printed by your code?**
[Positive]
Precision: 0.8111111111111111
Recall: 0.906832298136646
F1 Score: 0.8563049853372435

[Negative]
Precision: 0.9643605870020965
Recall: 0.9643605870020965
F1 Score: 0.9643605870020965

**7. What is the confusion matrix for the test set as printed by your code?**
Confusion Matrix:
True Positive (Spam) : 146
False Positive (Spam) : 34
True Negative (Ham) : 920
False Negative (Ham) : 15
[[920, 34],
[15, 146]]

**8. Draw a plot of log loss of the training data (y-axis) versus training iteration (x-axis), and include it in your report. You don't need to include the code for generating the plot. What is the total cost of your model? Use the below formulas (ignore b parameter (bias)):**
You can find this as Log_Loss_vs_Iterations.png
Total Log Cost: 0.12710826234773048 [25.421652469546096/200]




**9. Count the number of positive and negative instances in your train file (you don't need to report them). Keeping those counts in mind, what can you do to improve performance of your classifier on this corpus (apart from what you are already required to do in this homework)?**
Considering the class distribution in the training data, we can enhance the classifier's performance by adjusting the class weights,
we can experiment with different sampling techniques like oversampling, and exploring other ML algorithms such as ensemble methods or cost-sensitive learning.