Elijah Chou

Dr. Li Xiong

CS 334

11/8/2021

HW 3

1. (a) – (d): See q1.py for Q1. A total of 8 CSV files are created after running q1.py, including “binary\_train\_x.csv”, “binary\_train\_y.csv”, “binary\_test\_x.csv”, “binary\_test\_y.csv”, “count\_train\_x.csv”, “count\_train\_y.csv”, “count\_test\_x.csv”, “count\_test\_y.csv”.
2. (a): See perceptron.py for Q2(a). The main function was altered to default to using my binary datasets and maximum epoch number of 1000 when called on the command line.  
     
   (b): See q2bc.py for Q2(b). For the binary dataset, the optimal epoch number was found to be 9 from KFold cross validation with k=5, and it had a total of 1 mistake on the training set and a total of 32 mistakes on the test set using the optimal epoch number.  
     
   For the count dataset, the optimal epoch number was found to be 40 from KFold cross validation with k=5, and it had a total of 70 mistakes on the training set and a total of 54 mistakes on the test set.  
     
   (c): See q2bc.py for Q2(c). For the binary dataset, the 15 words with the most positive weights were “major”, “will”, “access”, “ever”, “paid”, “am”, “financi”, “increas”, “hour”, “websit”, “sight”, “click”, “market”, “guarante”, and “remov”. The 15 words with the most negative weights were “wrote”, “i”, “still”, “juli”, “re”, “the”, “network”, “seem”, “httpaddr”, “ll”, “review”, “post”, “comment”, “set”, and “prefer”.

For the count dataset, the 15 words with the most positive weights were “size”, “profession”, “insur”, “year”, “face”, “numberdnumb”, “pleas”, “market”, “call”, “email”, “compani”, “island”, “remov”, “numberb”, and “numberc”. The 15 words with the most negative weights were “numberp”, “numberd”, “but”, “re”, “file”, “cnet”, “wrote”, “date”, “if”, “url”, “razor”, “version”, “must”, “s”, and “messag”.  
  
Note: optimal epoch number, mistakes, and words with most positive/negative weights were printed after running q2bc.py.

1. (a) – (b): See q3.py for Q3.   
     
   The Multinomial Naïve Bayes model made 48 mistakes on the binary dataset and 50 mistakes on the count dataset. On the other hand, the Logistic Regression model made 31 mistakes on the binary dataset and 23 mistakes on the count dataset. All of this is printed in the terminal after running q3.py.