Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 7

Due: 11th April, 2024(100 points)

Dharun Suryaa Nagarajan

HOMEWORK GIT REPOSITORY

Homework Questions

1. Download the data from the homework 5 data folder. Train a logistic regression predicting who would survive with titanic.train.csv. Test to see your accuracy on titanic.test.csv

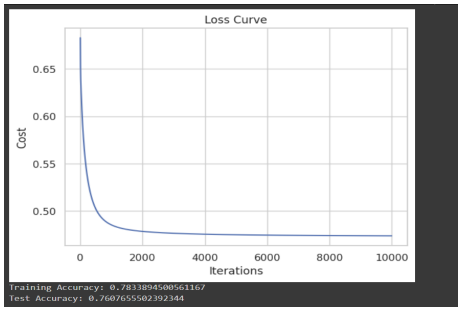

Figure 0.1: File 1

1. Overall accuracy on training=81.56
2. Overall accuracy on testing= 77.03
3. Include what features you used:
   $pclass, survived, sexmale, embarkedQ, embarkedS, farebin, agebinb.(13, 28], agebinc.(28, 44], travelsized$
   $ppl, titlegroupother$
   a. Used one hot encoding to get categorical to numerical features
   b. Used Min Max Scalar to normalize the dataset.
   c. created buckets of other features that will give more information to the model.
4. Include what features you used : 0.05