ML Assignment 1 Report

Aryan GD Singh
2019459

Q1.

We first read the dataset and then start preprocessing

Preprocessing:

1. We examine the data and notice that some Height values are 0, which is impossible, so those samples are removed from the data.
2. Rings column is renamed to Age, and all its values are increase by 1.5, as that is the way to calculate Abalone age form number of rings
3. Next we check which features are highly correlated and remove them as needed. This removes the extra weight parameters.
4. We standardize the data using StandardScaler since the input is real life samples, which can be modeled by the gaussian distribution.
5. Gender is encoded as integer, **I = 0, M = 1, F = 2**
6. Data is split into training and test sets, and also into input and labels

Linear regression:

Linear regression is performed with a learning rate of 0.1 and we get an RMSE of 2.53 on the training set, and 2.81 on the testing set.

Part 2:

We test our model with different values of penalty(alpha) on the coefficients. The coefficients decrease in value as the value of alpha increases. RMSE remains stable for low values of alpha but increases after a certain point. What this means is that our feature selection was correct, so the penalty does not affect model performance a lot.

GridSearch also performs similar to our models, only performing slightly better.

Q2.

Preprocessing:

1. We plot the correlation matrix and see that the features are not highly correlated, so there is no need to drop any features.
2. We standardize the data using StandardScaler since the input is real life samples, which can be modeled by the gaussian distribution.
3. Data is split into training and test sets, and also into input and labels

Logistic regression:

Learning rate = 0.1
BGD takes around 200 epochs to converge while SGd takes only around 10/15 epochs to converge, this is because in SGD the weights are updated for each sample input.
Final loss on the validation set is a bit higher than on the training set.

Testing learning rates:

- 0.01 learning rate is slower to converge than the original 0.1 taken by us, but it does reach convergence after a few thousand iterations

- 0.0001 is too low for the learning rate, it will take an unfeasibly large number of iterations to converge.

- 10 is too high, after some iterations the loss starts bouncing between 2 high loss values and does not reduce any further, it cannot reduce the loss to an acceptable level

Part 2:

Learning rate = 0.01

- sklearn SGDClassifer converges in around 10 epochs while our algorithm took more than 25 epochs, this may be because the sklearn model may use some extra optimisations that we have not

- Confusion matrix for both the SGDClassifier and our model is the same, so both perform similarly in terms of accuracy, precision, recall and f1score

Q3.

Preprocessing:

1. Training and testing data is loaded in.
2. Trouser and Pullover images are separated out as we only need these for our task.
3. The images are binarised to 0 and 1.
   if pixel value > 127 → 1
   else → 0

Our model performs with an accuracy of 93.25 %.