Introduction:

The task is to use machine learning methods to determine whether a given email is spam or not, in order to filter out the spam emails.

Problem Formulation:

The dataset is from UCI Machine Learning Repository-Spambase Data Set (https://archive.ics.uci.edu/ml/datasets/spambase), which contains 4601 samples: 1803 spam emails and 2790 non-spam emails. There are 58 attributes to consider whether an email is spam or not, the first 48 attributes are percentage of words in the email which matches the spam word, and the next 6 attributes are percentage of characters in email matches the spam "char", the 55-57 attributes are length of uninterrupted sequences of capital letters in the email, and the last attribute is the class target (spam:1, not spam:0). (from spambase.DOCUMENTATION). The inputs are the 57 attributes (the last attribute is class label, so I exclude that), and I split the train, validate, and test samples to 2760, 921, 920 respectively, then I normalize these data samples. The outputs are the predictions whether the email is spam(1) or not spam(0) in test samples.

Approaches:

1.baseline
Majority guess:
Since most of the email in our train set is non-spam(0), we predict all test samples are non-spam.

2. Binary logistic regression:
Hyperparameters:

a. Learning rate: choose from 0.1, 0.01
b. Max number of epochs: choose from 20, 50

3. Decision tree

4. K nearest neighbor:
Hyperparameters:

a. K values: choose from 3, 4, 5 neighbors

5. neural network:
Hyperparameters:

a. Optimizer: choose from "sgd", "Adam"
b. Learn rate: choose from 0.01, 0.001

Evaluation metric:
1. Majority guess: acc = 0.584
This method is an approximation, it is reasonable as the majority of the email in the dataset is indeed non-spam.

Below methods are all real goals of the task.

2. Binary logistic regression:
   - Max epoch:20, learn rate: 0.1, acc:0.90554
   - Max epoch:50, learn rate: 0.1, acc:0.91531
   - Max epoch:20, learn rate: 0.01, acc:0.90554
   - Max epoch:50, learn rate: 0.01, acc:0.90337

3. Decision tree:
   - acc = 0.91857

4. K nearest neighbors:
   - Acc: 0.91531 with k = 3
   - Acc: 0.90988 with k = 4
   - Acc: 0.91314 with k = 5

5. Neural network:
   - Max epoch:100,optimizer:Adam, acc:0.91965
   - Max epoch:200,optimizer:Adam, acc:0.91314
   - Max epoch:100,optimizer:Adam, acc:0.90445
   - Max epoch:100,optimizer:Adam, acc:0.89142

Result:

As the statistics show above, all these machine learning methods perform much better than the baseline (majority guess). Neural network with 100 iterations and Adam optimizer reaches the best accuracy.