IT3071 – Machine Learning and Optimization Methods 2022

Assignment 02

Take Home Assignment

Spam email is unsolicited and unwanted junk email sent out in bulk to an indiscriminate recipient list. Typically, spam is sent for commercial purposes. It can be sent in massive volume by botnets, networks of infected computers. While some people view it as unethical, many businesses still use spam. The cost per email is incredibly low, and businesses can send out mass quantities consistently. Spam email can also be a malicious attempt to gain access to your computer. "Ham" is e-mail that is not Spam. In other words, "non-spam", or "good mail". It should be considered a shorter, snappier synonym for "non-spam".

- Here the spam.CSV dataset has been attached and it contains both spam and ham emails.
- You must develop a program for predicting whether an email is a spam or ham by checking its content.
- You must use Support Vector Machine algorithm.
- You can first split the dataset into independent & dependent.
- Then using **CountVectorizer** class inside **sklearn** library, convert the independent text variable into the table form for showing the word frequencies and using these data, train the algorithm.
- Use 80% of data for training with Random State 0.
- Try to reach above 95% accuracy for testing data.
- You're allowed to perform any accuracy enhancing technique.
- Using the trained model, predict whether the following five emails are spam or ham.

Hey, you have won a car !!!!. Conrgratzz
Dear applicant, Your CV has been recieved. Best regards
You have received $1000000 to your account
Join with our whatsapp group
Kindly check the previous email. Kind Regards