CS 5100: Foundations of Artificial Intelligence (Fall 2019)

Final Project Proposal

Team: Nikhita Singh and Amey Arya

Topic: Building a Spam Filter Using Machine Learning

Problem Statement:

With the entire world moving towards a digital age, emails have become an integral part of our lives. Since, most, if not all the information important to us is communicated through emails, it has become an arduous task to keep track of what is more important. This is where spam filters come in. These filters classify the emails as spam and non-spam to make our lives easier.

We are trying to tap into this problem and understand how spam filters work using Machine Learning techniques. We plan to build our classifier and train it so that it predicts if an email is spam or not with some substantial degree of accuracy.

Algorithms:

We are implementing an example of supervised learning using the following machine learning algorithms:

1. Naïve Bayes
2. Decision Tree
3. Random Forest
4. Support Vector Machine
5. K Nearest Neighbour

- Input: The dataset that we are going to use is a pre-processed subset of the Ling-Spam Dataset, provided by Ion Androutsopoulos.
- Output: Trying to achieve high accuracy using this classifier o filter emails.

Outcome:

To learn how these different Machine Learning techniques help us achieve the goal and thereby develop an intuition for real-world problems. We also plan to do an analysis of how each technique differs from the other and which one gives a better result.

What is to be learned:

Naïve Bayes: Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

Decision Tree: Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modeling approaches used in statistics, data mining and machine learning.

Random Forest: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

Additionally, we will try to implement SVM and kNN to the same problem.

All the mentioned libraries are available on scikit-learn (machine learning libraries built on python).

Dataset:

The Ling-Spam Dataset: A dataset that contains spam messages and messages from the Linguist list.

http://www.aueb.gr/users/ion/docs/ir_memory_based_antispam_filtering.pdf

Milestones:

- 11/11:
  - Learning Naïve Bayes, Random Forest, and Decision Tree classifiers
  - Setting up the environment on the local machine as well as GitHub
  - Cleaning and scraping of the above-mentioned dataset
  - Deciding train: test ratio

- 11/25:
  - Implementing Naïve Bayes, Random Forest, and Decision Tree classifiers
  - Generating graphs to analyze the comparison between the three classifiers
  - Experimenting with SVM and kNN classifiers

Week-by-Week plan:

Week of 10/28: Start learning about Machine Learning (algorithms, classifiers).

Week of 11/4: Work towards milestone 1.

Week of 11/11: Implementing Naïve Bayes, Random Forest, and Decision Tree classifiers.

Week of 11/18: Work towards milestone 2.

Week of 11/25: Work on graphs and presentations.

Week of 12/2: Prepare for the interview and final tweaks on the project.

Week of 11/9: Final submission.