CSCI4360 Term Project Abstract

Group Members(first name alphabetical order): Mingyu Sun, Shubhangi Rai, Ye Tian
Instructor: John A. Miller

1. **Purpose of Study:**

   Classification based on given data has been an important topic in many fields such as marketing protocols and internet content due to its worldwide accessibility. Spam emails are one of the major problems of today's internet causing email based threats such as financial damage to the organizations. Spam emails are occupying the user's mail boxes consuming time and capacity without the user's consent. Eradicating spam emails without eliminating legitimate ones need some important measures. Spam filtering is one of the most important techniques which is centered more on the classifier-related issues. The effectiveness of this project identifies the use of different learning algorithms for classifying spam emails from the legitimate emails. A comparative analysis among the algorithms would also be presented.

   In this report, we investigate the effectiveness of different modeling techniques on classification, including Regression models and Neural Networks. We study the Regression models first, starting with the NullModel, then Multiple Linear Regression, and more complicated Regression models like quadratic regression and Transform Regression. As the datasets are pretty large and complex, we will mainly focus on Neural Networks with different numbers of layers. We will compare all the models, by the coefficient of determination R^2 , adjusted coefficient of determination $\bar{R}^2$, and that with cross validation R^2_{cv} . We will also compare their running time. Apart from these two, we will also consider how the model generalizes, that is, if the models could work reasonably well for datasets from different fields. We hope to choose the best modeling technique considering quality of fit, universality, as well as the cost. Besides, we also explore the relative importance of different features via forward selection, backward elimination with the coefficient of determination R^2 as our criteria. Considering the running time, we will begin with simple datasets, like auto-mpg, annealing, Algerian Forest Fires datasets first, but our focus will be the two large-scale datasets for this report. This also gives us a chance to investigate if our models generalize for datasets with different sizes.

   a. Apply feature selection process and several classification models, compare and choose the better combination of attributes and models to predict the results on the testing set. Specifically:
      i. For the bank marketing dataset, we hope to figure out which attributes have larger influence on the decision of the clients, and from the feature selection process, study how we can help the bank to improve client submission rates.
      ii. For the spambase dataset, we want to find out the "keywords" that could most accurately identify spam emails with least mis-identifications

possible. We would like to remove the unwanted expression from the dataset and tokenize it using count vectorizer to form features. This could be done by different spam filtering techniques and Classification Algorithms such as Random Forest Classification, Naive Bayes, Support Vector machine , etc

b. Since there are datasets chosen that are recommended only for classification models, we want to experiment with some regression models and see how they perform on the datasets, compare them with the prediction performance given by our best classification models, so as to improve our understanding of the regression models chosen.

2. **Datasets selected:**
   a. Bank Marketing Datasets: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing
   "The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe to a term deposit (variable y)."
   b. Spambase Dataset: https://archive.ics.uci.edu/ml/datasets/spambase
   "Classifying Email as Spam or Non-Spam"
3. **Data Preprocessing Techniques:**
   a. Remove identifiers (if applicable), convert string columns to numeric columns, throw away missing values, detect outliers.
   b. Tokenize it using count vectorizer to form features
   c. Create a matrix/array/dataframe for the dataset.
   d. Scikit-learn provides some pre-packaged tools for splitting datasets into training and testing parts, such as sklearn.model_selection.train_test_split.
   e. Different Algorithms of Classification such as Random Forest Classification, Logistic Regression, Naive Bayes Classification, K mean clustering and Support Vector Machine.
4. Models planned for experiment: NullModel, Multiple Linear Regression, transformed regression, quadratic regression, quadratic regression with crossed terms, Neural Networks, etc.