ELEG 6913: Machine Learning for Big Data

Fall 2016

Department of Electrical and Computer Engineering

Prairie View A&M University

Project 1

(a) Text classification is to assign the text to one or more predefined classes or categories. The text could be a document, news article, search query, email, tweet, support tickets, customer feedback, user product review etc. We can apply this technique to information retrieval, recommendation system, filtering spam email, sentiment analysis, and analyzing customer feedback. The goal of this project is to construct a text classifier to classify texts into two predefined categories. Text data for the project is available at https://github.com/BruceDong/Resources-for-Projects-of-Machine-Learning/tree/master/Datasets/classification.

(b) Use the feature extraction method: “bag of words” to represent the text data as feature data.

(c) Divide the feature data into 5 parts: 4 parts as training data and 1 part as testing data.

(d) Choose a classification algorithm from logical regression, neural network, and support vector machine, and train it on the training data to construct a text classifier.

(e) Evaluate the text classifier on the testing data with evaluation metrics, namely precision, recall, and F-score.

(f) Repeat everything in (b) with another feature extraction method: term frequency–inverse document frequency (TF-IDF).

(g) Repeat everything in (d-e) with the same classification algorithm. Compare your results with (e).

(h) Divide the feature datainto 10 parts: 9 parts as training data and 1 part as testing data.

(i) Repeat everything in (d-g) with the same classification algorithm.

(j) Summarize your observations and conclusions.

Submit your programs and supporting documents as one zip file in ecourses by December 6, 2016.