DS 200 Introduction to Data Sciences

A Mini Data Science Project

Fall 2018

Instructor: John Yen

Learning Assistant: Luwei Lei

Motivating Questions

The goal of the mini project is to provide you with a hands-on personal experience regarding the construction, interpretation, refinement, deployment, and evaluation of a twitter-based opinion/stance classification pipeline to explore answers to the following questions:

- How can we build and improve decision tree-based models by adjusting model parameters (e.g., maximum tree depth and minimum number of samples in the leaf node)?
- How can we extract knowledge and important features from decision trees generated from Startified k-fold cross validation?
- How can we interpret the roles (or meaning) of these features for the predictive model?
- How can we improve a predictive model after applying it to a new dataset (that does not have tags)?

You will be randomly assigned to one of two topics for the mini project. Based on your topic, you will be first provided with a set of tagged tweets (both for relevant/irrelevant tag, and for stance tag). Using this data set, your first project deliverable (PD #1) will include the following:

A. Model Construction: Using the set of provided tweets (which have been tagged both for Relevant and for Stance) to construct two Decision Tree-based predictive models for two purposes: (1) filtering tweets that are irrelevant, and (2) predicting the stance of relevant tweets.
B. Model Assessment: Use Stratified k-fold cross-validation to rigorously assess the quality of each predictive model.
C. Model Interpretation: Using the trees constructed from part A/B to extract knowledge and insights (i.e., rules) for predicting relevant tweets and for predicting stance of tweets. Comment on the confidence level about the rule. Compare multiple decision trees generated during k-fold cross validation to identify important features.

Your second project deliverable (PD #2) is the best relevant tweet classification model and stance classification model (Model 2) you were able to construct by systematically adjusting two parameters for constructing decision trees: (1) maximum tree depth and (2) minimum samples in leaf nodes. The model will be applied to a new set of tagged tweets to assess the performance of their prediction. All models created for a common topic will be ranked by their performance.

During the final phase of your project, you will be provided with additional twitter data (Dataset 2) for improving the two classification models you have developed. The resulted model will be applied to the last set of twitter data (Dataset 3), which simulates the performance of your model after deployment.

The final project deliverable (PD#3 Final Project Report) includes the following elements:

- The construction of your relevant tweet classification model and stance classification models for PD1, PD2, and PD3.
- The Stratified k-fold cross validation results of PD1, PD2, and PD3.
- The interpretation of the models constructed for PD1, PD2, and PD3, including similarities and differences of the features/rules between these models.
- The impacts of using new data (Dataset 2) on the prediction performance of the models.