Project D: Natural Language Processing and Machine Learning Classification

Objective: Develop and apply algorithmic classifiers for news stories.

In this project you will prepare text from the RCV1 corpus for text analytics research, and use it for training, evaluation, and selection of supervised machine learning algorithms.

Universe: The set of objects that your classifier needs to handle consists of over 800,000 news articles in English that have been conventionally pre-processed: stemming, stop word removal, capitalization, punctuation, etc. Every article has been manually coded and mapped to a hierarchical category of subject topics. Details are provide in Ref. 1.

Data and Computing Environment: The RCV1 data has been extracted and loaded into the SQL Server database **rcv1** on hedge.mit.edu. The articles have been split as in Ref. 1 into two groups, stored in tables **news** and **news_test**. However you may choose to select training and test sets differently if you wish. The table **topics** contains the subject headings in parent-child format. For convenience, a numerical id key column has been added. The same data has also been broken out by hierarchy level in tables `h1,h2,h3,h4,h5`. The assignments of articles to topics is contained in the table **news_topics**. You should browse the data and become familiar with the structure, style, and content of the assignments.

The core R packages for this assignment are **tm** (text mining), **RTextTools**, **tidytext**, **e1071**. You may use others as well.

1. **Model training, evaluation, and comparison:**

   In this first part of this assignment you will perform cross-model comparison. It is important to prepare the datasets, train the models, and compute performance analytics in a consistent manner.

   There are two tasks to perform with each model:

   (a) Assign articles to the four top-level topics (i.e., Corporate/Industrial, Economic, Government/Social, Market).

   (b) Assign articles to the most specific topic levels you can.

Select at least four different learning algorithms among those supported in the **RTextTools** package. Train the models and evaluate their performance on classifying out-of-sample test data.

For each model, give a description of how it is supposed to work. Evaluate how the models did on each learning task and compare. Did one algorithm dominate? Do the algorithms show different strengths and weaknesses?

Include diagnostics such as the following: accuracy, recall, precision, and the confusion matrix.

Describe the procedure you used to arrive at these results. For instance, how did you decide on the size of the training vs. test sets, the size of the active vocabulary, etc.

2. **Model selection:**

   Select one model to use for further tuning and classification. Pick the best model, or your favorite, from those above. If you prefer, you can construct an ensemble model that combines the outputs of more than one base model.

   Explain the reasons for your selection based on the cross-model comparison above. Do you expect your choice to outperform in all situations?

3. **Model tuning:**

   Explore parameters of the model, such as cost, gamma, tree pruning, etc. to optimize its performance. Also consider parameters related to the preparation of training data, such as term frequency metrics (e.g., Tf vs. Tf-Idf) and term sparseness (i.e., threshold for removing sparse words in constructing the document-term matrix).

   Develop your best classifier(s) for RCV1 news articles for the two tasks. How much better (if any) is it compared to the default settings you used in the initial cross-model evaluation? How do you expect it to perform on fresh out-of-sample data?

4. **Model competition:**

   After your report is submitted and your final code/model are set, everyone will be given fresh out-of-sample data to classify. Let's see whose predictions hold up best!

References (available on Canvas): These references provide relevant background on the data, the software, and machine learning analytics, respectively.

1. Lewis, Yang, Rose, and Li, "RCV1: A New Benchmark Collection for Text Categorization Research," *Journal of Machine Learning Research* 5 (2004) 361-397

2. Jurka, Collingwood, Boydstun, Grossman, and van Atteveldt, "RTextTools: A Supervised Learning Package for Text Classification," *R Journal* 5 (2013) 6–12

3. James, Witten, Hastie & Tibshirani, "An Introduction to Statistical Learning, 2nd Ed.," Springer, 2009