BAYESIAN MULTINOMIAL LOGISTIC REGRESSION FOR AUTHOR IDENTIFICATION

Dmitriy Fradkin[1], Alexander Genkin[1], David D. Lewis[2], David Madigan[1,3]
(1) DIMACS, Rutgers University
(2) David D. Lewis Consulting
(3) Department of Statistics, Rutgers University

Abstract

Author attribution has a long history that includes some famous disputed authorship cases and also has forensic applications. The focus of our work is on large collections of documents, large groups of potential authors (hundreds and thousands), and thousands of stylometric features. For "one of K" classification learning we chose polytomous logistic regression:

$$p(y_k = 1|\mathbf{x}, \mathbf{W}) = \frac{\exp(\mathbf{x}^T \mathbf{w}_k)}{\sum_{k'} \exp(\mathbf{x}^T \mathbf{w}_{k'})}$$

and Bayesian estimation. We implemented a program that handles problems of that size, supports Gaussian and Laplace priors, and is now publicly available.

We experimentally studied different sets of stylometric features and their combinations and found their relative value for classification to be fairly stable on diverse data collections. Trying to distill author identification from topic categorization leads to "cross-topic" experiments that show significant change in relative value of feature sets and help select feature sets that are less topic dependent.

Another interesting problem is "odd man out": check if a document belongs to any of the given set of people, each of them being represented by a set of documents they authored. This problem can be treated as binary classification by introducing "decoy" authors as opposing classes in learning. We show however that results can be improved with multi-class treatment: we first train a polytomous regression model on the target set of authors, then use regression model scores for all classes as features in binary training with a "decoy".

Key Words: Multinomial logistic regression, polytomous logistic regression, Bayesian regularization, sparsity, author identification, stylometry