Supervised Methods for Classifying and Scaling Texts: Lab Exercise

Kenneth Benoit

This exercise involves using the automatic document classification features of WordStat, using texts from movie reviews (files here) [Pang and Lee, 2004, Pang et al., 2002] and then from Evans et al. [2007] *amicus curiae* briefs (files here).

Instructions

1. Load the movie review texts into QDA Miner. After creating a new project, begin by loading the positive reviews, and use the spreadsheet editor to code all of these with under a new variable type - Sentiment - with the value POS. Then load the reviews from the negative folder and give them the variable value NEG. Make sure these are Categorical variable types.

2. Open WordStat with the parameters as follows: 'Analyse all text in relation with Variable SENTIMENT'.

3. Choose the automated text classification button (3rd from the left, bottom row, in the Crosstab panel)

4. Try the different options in the 'Learn and Test' panel and observe the results. Note the different options for performing cross validation.

5. Construct a systematic exploration of the parameter space with the experiment button on the history panel.

6. Repeat the experiment, but choose a much smaller set of examples. What is the relationship between the accuracy and the size of the training set?

7. Create a new project for the Evans et al amicus briefs. Import all of the texts in the "training" and "testing" folders. Create a variables for "SET" (training or test) and "Class" (petitioner or respondent).

8. Predict the category of petitioner versus respondent for the *amicus* briefs using only the training briefs. You can choose which documents to predict from the 'Apply' tab by selecting 'list of documents' and 'edit list'.

9. Experiment with feature selection to see if predictive accuracy can be improved.

References

Michael Evans, Wayne McIntosh, Jimmy Lin, and Cynthia Cates. Recounting the courts? applying automated content analysis to enhance empirical legal research. *Journal of Empirical Legal Studies*, 4(4):1007–1039, December 2007.

B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, 2004.