MATH-GA.2047-001 Data Science in Quantitative Finance

Homework 9

Profs. Ivailo Dimov and Petter Kolm
Due: November 10, 2020

Instruction

This homework is to be done individually. No collaboration and/or code sharing permitted.

Objective

In this assignment, you will:

- Use cross-validation to optimize model parameters of the various classifiers for the Reuters dataset demonstrated in class.
- Compare the classifier performance of the optimal models of each model class.

Methodology and Deliverables

1. Classifier Performance Comparison.
   In `09_2_lsa_demo_with_classifier_performance.ipynb` we demonstrated how to classify Reuters articles as acquisition-related by using a K-nn classifier on (A) the TF-IDF vectorized Reuters dataset (B) an LSA-enhanced TF-IDF vectorized Reuters dataset where the LSA vectors of the top 200 components were concatenated with the TF-IDF vectors. We compared the classifier performance of the A and B flavors of each K-nn model by calculating the ROC and precision recall curves.
   (a) Pipeline all of this functionality in a single Scikit-Learn `ReutersClassifier` model. The model parameters should be:
      - `num_lsa_components` which should be an integer between 0 and 500 corresponding to the number of LSA components concatenating the TF-IDF vectorization of the input data. The default should be 0 corresponding to pure TF-IDF vectorization.
      - `n_neighbors` which should be an integer between 0 and 20 corresponding to the number of nearest neighbors used in the `KNeighborsClassifier`

(b) For each of the following target labels (1) ‘earn’ (2) ‘usa’ (3) ‘corn’ create a train and test set of the articles which are tagged with a given label or not. Should you use stratified shuffle split for any of these variables, and if yes for which ones?

(c) Using `GridSearchCV` find the optimal `num_lsa_components` and `n_neighbors` for each of the training datasets (1)-(3) above. Use ‘f1’ a score evaluation metric. Plot the ROC and Precision-Recall (PR) curves for the optimal parameter. Repeat your analysis but instead of the ‘f1’ metric, use the ‘roc_auc’ score. Compare the ROC and PR curves for each of the two optimal models. Discuss your results.