CSC570 Machine Learning
Final Project

Machine learning in politics

(due by the end of the day on Sunday, December 10th)

Download the dataset house-votes.txt. The dataset includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by Congressional Quarterly Almanac, 98th Congress, 2nd session 1984, Volume XL: Congressional Quarterly Inc. Washington, D.C., 1985. The CQA lists nine different types of votes: voted for, paired for, and announced for (these three simplified to yea), voted against, paired against, and announced against (these three simplified to nay), voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an unknown disposition).

The task is to predict whether the voter is a republican or a democrat based on their votes.

Attribute Information:

1. Class Name: 2 (democrat, republican)
2. handicapped-infants: 2 (y,n)
3. water-project-cost-sharing: 2 (y,n)
4. adoption-of-the-budget-resolution: 2 (y,n)
5. physician-fee-freeze: 2 (y,n)
6. el-salvador-aid: 2 (y,n)
7. religious-groups-in-schools: 2 (y,n)
8. anti-satellite-test-ban: 2 (y,n)
9. aid-to-nicaraguan-contras: 2 (y,n)
10. mx-missile: 2 (y,n)
11. immigration: 2 (y,n)
12. synfuels-corporation-cutback: 2 (y,n)
13. education-spending: 2 (y,n)
14. superfund-right-to-sue: 2 (y,n)
15. crime: 2 (y,n)
16. duty-free-exports: 2 (y,n)
17. export-administration-act-south-africa: 2 (y,n)

1. Import the txt file into a data frame.

2. Prepare the data. Impute missing values in a way that agrees with the goal of the project.

3. Apply two machine learning algorithms to the dataset. One of the algorithms should have not been used in previous homeworks. In other words, one of the algorithms must be different from kNN, Naïve Bayes and linear regression. Use confusion matrices and different performance measure to compare the algorithms.

4. Use 10-fold CV to estimate how well the learning algorithms will perform on new datasets. Compare the algorithms.

5. Perform automated parameter tuning for both models (if they allow it) using the caret package.

6. Try to improve the performance of each algorithm by using ensemble learning (one method of ensemble learning of your choice) and the caret package. Compare the algorithms.

Write a report (between 3 and 5 pages) describing what you have done, the algorithm comparisons and summarizing the results.

Your submission must consist of two files:
- a report as a txt, docx, or a pdf file
- a text file, script.txt, (without any drafts, errors, or debugging) of your program. The script must represent your final program and should be directly executable. That is, one can directly copy, paste it, and execute it into the R console.