Computational Methods for American Politics

Problem Set # 2: Classification & Unfolding

In this assignment, you will be asked a series of questions covering classification and unfolding.

Note: Each student must submit a **single rendered PDF** with code and output produced in line, in addition to responses to all questions. Submit to Github **by Monday, November 11 at 5 pm.**

1 Part 1: Classification

For the first part on classification, you will use data on Senators' votes on Supreme Court nominees. The key explanatory variables are the squared distance between the senator's ideal point and the nominee's inferred ideal point (`EuclDist2`), the perceived qualifications (`qual`) or lack of qualification of the nominee (`lackqual`), a dummy variable indicating that the president is strong (`strngprs`), and a dummy variable indicating that the senator shares the president's party affiliation (`sameprty`).

After loading the data, run the following line to drop observations for Justice Sam Alito, because the JOP paper used only nominees through Chief Justice John Roberts.

```
conf06 <- subset(data, data$nominee!="ALITO")
```

Further restrict the data to only include variables of interest (for ease of working with the data), and also create your dependent variable, `numvote`, based on the `vote` variable.[1]

```
vars <- c("vote", "nominee", "sameprty", "qual", "lackqual",
"EuclDist2", "strngprs") # vector of vars
conf <- conf06[vars] # retain only key vars from above object
conf$numvote <- as.numeric(conf$vote)-1 # from 1/2 to 0/1
conf$numstrngprs <- as.numeric(conf$strngprs)-1 # same as above
```

Your goal in the first part of this assignment is to build and compare two major classification approaches (logistic regression and LDA). The outcome is a Senator voting "yes" on a Supreme Court nominee, as a function of covariates of interest. With the data loaded, answer the following questions.

1. Define an 80/20 train/test split.

2. Build logit classifier. Present the results in a useful way (e.g., a confusion matrix, etc.). Discuss the output in substantive terms.

3. Build an LDA classifier. Present the results in a useful way (e.g., a confusion matrix, etc.). Discuss the output in substantive terms.

[1] Be sure to inspect your data everytime you manipulate it; this is a vital habit to get into if you aren't already.

4. Calculate and plot the predicted probabilities from **only the logit model**, over the range of the nominees' perceived qualifications, holding all other variables at their mean values (*hint*: we did this in class). Describe these results in substantive terms (e.g., how does the probability of a yes vote change as the nominees' qualifications lessen?).

5. Offer a couple paragraphs explaining all of your findings for the reader. For example, talk about different variables' impacts on the ability to correctly classify a Supreme Court nominee being supported. Or consider talking about how politicized the nominations process has become, and how we do (or do not) see this in these data. Just present a nice summary; short, to the point.

6. **BONUS QUESTION**: Generate the predicted probabilities plot in the previous question 4, **but** include separate curves for each level of the `sameprty` variable (i.e., a conditional plot).

2 Part 2: Unfolding

The study of legislative behavior through analysis of roll call votes has proliferated with the development of complex scaling techniques, such as NOMINATE (Poole & Rosenthal 1997). In this next section, you will focus on unfolding via the W-NOMINATE algorithm. Here, you will use the 113th U.S. House roll call data (`.ord`). You can load it with the following chunk of `R` code:

```
house113 <- readKH(
        file.choose(), # locate the .ord file saved locally
        dtl=NULL,
        yea=c(1,2,3),
        nay=c(4,5,6),
        missing=c(7,8,9),
        notInLegis=0,
        desc="113th_House_Roll_Call_Data",
        debug=FALSE
)
```

With the data loaded, answer the following questions.

1. Fit a W-NOMINATE algorithm. Present a plot of members and discuss the results in substantive terms. What do you see?

2. Discuss the dimensionality of the space. You can present and inspect fit via the aggregate proportion reduction in errors (APRE), the geometric mean prediction (GMP) rate, scree plots, or any other diagnostic tool (visual or numeric) to inspect the overall fit of the algorithm.

3. Based on the previous inspection, discuss the advantages and disadvantages of major unfolding approaches (discussed in class) to studying roll-call voting and binary choice data more generally? When might one approach perform better than other approaches? Etc.