Applied Scaling & Classification Techniques in Political Science

Lab 6

Dictionaries and Supervised classification methods

Difference between procedures

When you want to do a text classification (with a training-set and a test-set), the procedures are different when you want to use a random forest/support vector machine algorithm vs. a naive bayes algorithm

- *When you want to use the former, those are your steps:*

1. *You create a unique corpus of the texts including both the training-set and the test-set corpus*
2. *You create a unique DFM*
3. *You transform such unique DFM into a unique data frame and you add to such data frame the column of the «sentiment» originally included in the unique corpus*
4. *You split the unique data frame in 2 (one for the training-set and one for the test-set) according to the presence or absence of «sentiment» information in the unique data frame*
5. *You run the analysis*

Difference between procedures

When you want to do a text classification (with a training-set and a test-set), the procedures are different when you want to use a random forest/support vector machine algorithm vs. a naive bayes algorithm

- *When you want to use the latter, those are your steps:*

1. *You start with two different corpus of texts (one for the training-set and one for the test-set)*
2. *You create two separated DFM*
3. *Naive Bayes can only take features into consideration that occur both in the training set and the test set, so we have to make the features identical!*
4. *You run the analysis*

Difference between procedures

Why this difference? Because the Naive Bayes algorithm based on a multinomial distribution (i.e., the suggested algorithm when dealing with texts) is implemented at the moment only within Quanteda!

Difference between procedures

When you want to run a k-fold cross-validation on your training-set, the procedures are different when you want to use a random forest/support vector machine algorithm vs. a naive bayes algorithm

- *When you want to use the former, those are your steps:*

1. *You start with the data-frame that you got at point 4 above related only to the training-set*
2. *You randomly split it according to the value of K you want (K=2, 5, 10) by creating different data-frames (i.e., data-frames both including Ki (if K=3, i runs from 1 to 3), as well as data-frames NOT including Ki)*
3. *You run cross-validation analysis*

Difference between procedures

When you want to run a k-fold cross-validation on your training-set, the procedures are different when you want to use a random forest/support vector machine algorithm vs. a naive bayes algorithm

- *When you want to use the latter, those are your steps:*

1. *You start with the corpus inlcuing only the training-set*
2. *You randomly split it according to the value of K you want (K=2, 5, 10)*
3. *You create a DFM for both the corpus including Ki (if K=3, i runs from 1 to 3), and the corpus NOT including Ki*
4. *You run cross-validation analysis*

Difference between procedures

There are also some packages in R that allow you to run the cross-validation procedure in just one line of command (such as *Caret*), but better first learning what you have to do step-by-step!