IST736 Text Mining

HW4

Multinomial Naïve Bayes

**~~Use Weka’s~~**

**Use PYTHON’s Multinomial Naïve Bayes algorithm to build models to classify the customer reviews by (1) sentiment (positive or negative) (2) authenticity (true or fake, lie detection). ~~(6 points)~~**

**Use the dataset that is specific for this assignment.** You can create your own test dataset to test your code (which is always good), but then do the assignment with the dataset from class.

This is important as this dataset has two labels and must be (HINT HINT) separated in to two different datasets and then vectorize each. So, you will have one dataset for sentiment (the p or n) and one dataset for fake or true (f or t).

For each of the two classification tasks, use MNB to build the models, and evaluate them using 10-fold cross validation methods. (5-fold is fine too) ~~You can choose your own options for tokenization, and make sure to report if you change any default setting.~~

First – perform NB on a small sample of a dataset that should be easy to use for prediction. You will need labeled data because you are **training** the model and then testing it. You will need to vectorize the data first, and then separate it into a training set and a testing set. Then, remove and keep the labels. Labels CANNOT be part of the data.

Train the model.

Test the model.

Compare the model’s test data prediction results to the actual labels.

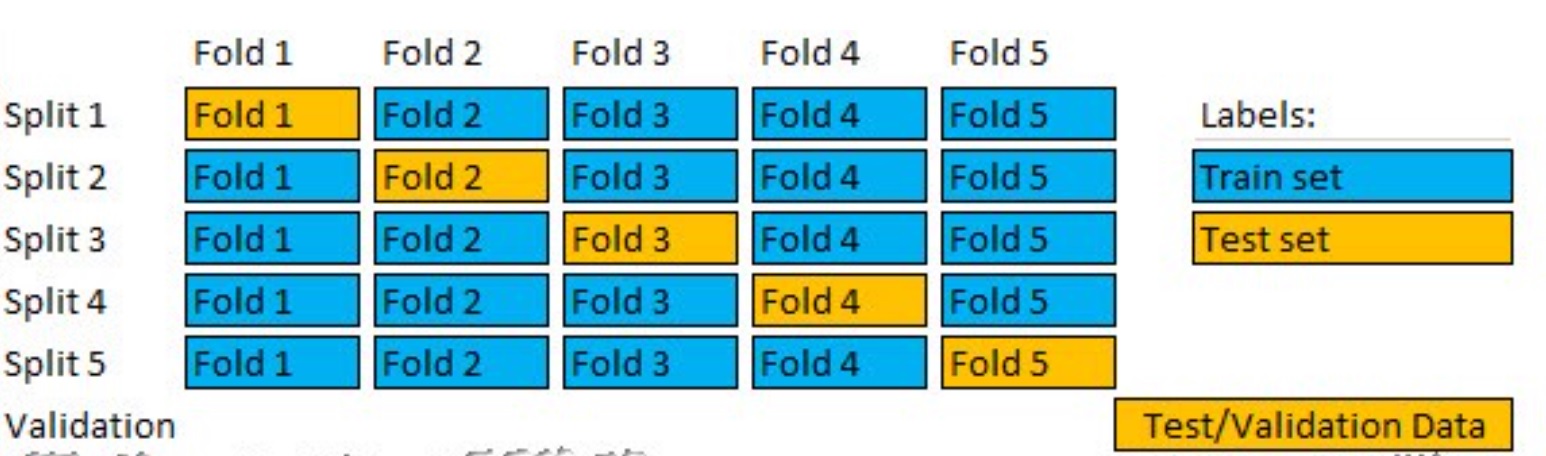
Use a confusion matrix to look at the results.

Next, do the above on the full dataset.

Once that is done you have succeeded in 1-fold cross validation ☺ You did this ONE TIME. Your test set is ONE sample of your data.

Now, to perform 5-fold cross validation, you do the above 5 times (so that each test/train set is different).

There are many ways to do this. One is to place the above in a loop and make sure your test set is created via random sampling or a sample method that you control and so can make sure is different each time.



The image above is ONE POSSIBLE WAY to do this. The “Test/Validation” data is optional and is a collection of the data that is never part of the training. It is “set aside fully”.

**We will talk about all of this in class as well.**

**Other cross validation methods….**

**What** is “cross validation”?

<https://machinelearningmastery.com/repeated-k-fold-cross-validation-with-python/>

<https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6>

**Use CountVectorizer and Python. Create labeled data. Train the NB model and test it.**

**As part of your Results section:** For each model (lie detection and sentiment classification), report the 20 most indicative words that the models have learned.

**As part of your Results (the techy part) and Conclusions (the non-techy part) include discussion of:** Based on these words, do you think the models have learned the concepts (lie or sentiment) that they are expected to learn?

**As Part of Results**: Also, compare the difficulty level of sentiment classification vs. lie detection. Discuss whether you believe computers can detect fake reviews by the words.

**Submission: Always use the required Homework Assignment FORMAT.**

~~Submit a word document in research paper format, similar to previous homework.~~

**Grading rubrics: Same as always.**

~~Same as HW1~~.