Statistical Significance

9-20-2018

Working on doing the corpus combinations with the two different sets of data (Wanga and another Luyia language). However, I got an error that the corpora are actually dictionaries. I believe the error is because I'm pushing the objects that are stored in the shelve files when really I should be pushing the oobjects that create those objects. 9-20

9-24-2018

I have fixed the issues with the corpus objects in the shelves by storing the non-vectorized versions. There are improvements to be made to my code to make it a bit slimmer. However, it works decently well as is. I have a script that has been running for two days now trying to build all the corpora. All the different corpora except the case with the merged swahili corpus finished about a day ago. However, the merged swahili corpus takes quite a while. Changes have also been made to the `training_template.py` file so that when I finish building the extracted feature shelves, it should be ready to run. However, a few errors that I had not anticipated are sure to emerge.

9-26-2018

Writing the dictionaries into a pandas dataframe is currently not working. I'm getting an error that the arrays in the dictionary are not the same length.

11-12-2018

I now have ran through nearly all the settings 1000 times. The only exceptions are the Swahili tests at 0.5. I stopped these because they were taking too long to finish. Once I get a second server up and running consistently, I will go ahead and try the Swahili stuff again for this setting.

I have also set-up an r-shiny application online here. The data looks roughly normally distributed

which is good. I was concerned that this would not be the case.

03-15-2018

My best bet for a statistical test is Wilcoxon's signed rank test as described in Japkowicz & Shah 2011.

This test is ideal because "the t test can be more powerful than the Wilcoxon's Signed Rank test when the parametric assumptions made by the t test are met. However, Wilcoxon's test is the method of choice when this is not the case" (p. 236).

I believe this can be extended to the multiclass domain by using a monotonic multiclass performance measure like macro average f-score.

The data need to be paired though and there is no sense in which the pairs between different language datasets could ever be dependent. E.g. no matter how sampling is done, there is no way to create a pair of trials such that both are matched with one using Swahili data and the other Tiriki. By definition of using different training datasets in different languages, they are not dependent trials.

The solution is to use the related Mann-Whitney U test. My performance metrics are all ordinal. However, this does assume independence. I could do a cross validation setup where the training data is split into say 10 folds and then 5 folds (corresponding to the .1 and .2 ratios used previously.

Revision #1

Created Thu, Feb 6, 2020 4:42 PM by kenneth

Updated Thu, Feb 6, 2020 4:43 PM by kenneth