Ahmad Elshenawy
Ling 572 - Wi2014
February 11, 2014

Homework 5 – writeup

Table 2:

p_0	$\mathbf{X}_0^2$ score	# of related features	Test Accuracy
Baseline	0.0	32846	0.72
0.001	13.816	2401	0.75
0.01	9.21	3895	0.75
0.025	7.378	5223	0.756666666667
0.05	5.991	7484	0.733333333333
0.1	4.605	8499	0.746666666667

Table 3:

p_0	Test Accuracy
Baseline	0.813333333
0.001	0.856666666667
0.01	0.846666666667
0.025	0.846666666667
0.05	0.82
0.1	0.853333333333

Q5) The differences between the results in Table 2 and Table 3 is quite noticeable, with the accuracies in Table 3 about 10% better than in Table 2. I'm not quite sure, just from these two runs, whether or not this is because of the different training sets, or the different k-values used in classification. At first blush, though, it would seem that using the new training data, which I believe is binarized, makes a difference regarding accuracy.

As a general trend, the highest accuracies were generally for more restrictive Chi score thresholds, but there is also the strange occurrence that $p_0 = 0.05$ performs worse than $p_0 = 0.025$ and $p_0 = 0.1$, in both result sets. I have no idea what this might be the result of, but it is consistent in both Tables.