Olga Whelan
Mike Roylance
Ling 572

Homework 5

Overall notes: we created two additional scripts, produce_q4.sh and produce_q5.sh. If you would like to run these, they will produce the necessary files to answer q4 and q5. We have also provided that information in the tables below.

q4)

p0	x0^2 score	Number of related features	Test accuracy
baseline	0	32846	0.72
0.001	13.816	2401	0.75
0.01	9.210	3895	0.75
0.025	7.378	5223	0.756666666667
0.05	5.991	7484	0.733333333333
0.1	4.605	8499	0.746666666667

q5)

p0	x0^2 score	Number of related features	Test accuracy
baseline	0	32846	0.816666666667
0.001	13.816	2401	0.856666666667
0.01	9.210	3895	0.85
0.025	7.378	5223	0.846666666667
0.05	5.991	7484	0.82
0.1	4.605	8499	0.853333333333

q6)

Conclusions

Feature selection using chi-square considerably improves the accuracy of the algorithm - the baseline (where no features were selected) was the lowest performing among both binarized and non-binarized features. Chi-square makes a good judgement of usefulness of the features for classification.
Increasing or decreasing p0 does not seem to have a direct influence on performance. For instance, performance accuracy decreased for the non-binarized features selected with the significance level from 0.025 to 0.05, but increased from 0.01 to 0.025, while for binarized features the accuracy dropped over the same change of significance level: 0.01 to 0.025. This means that for text classification it may not matter if a few features are added to the set or removed from it.
Binarizing the features has a positive impact on performance of knn.