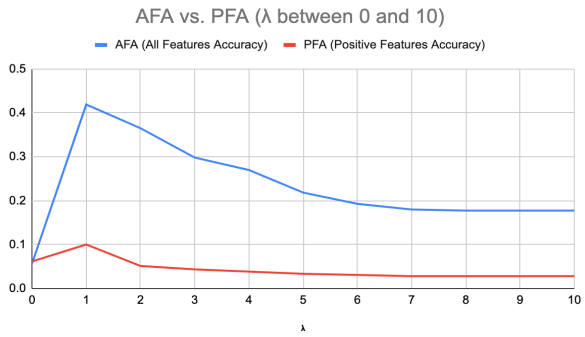
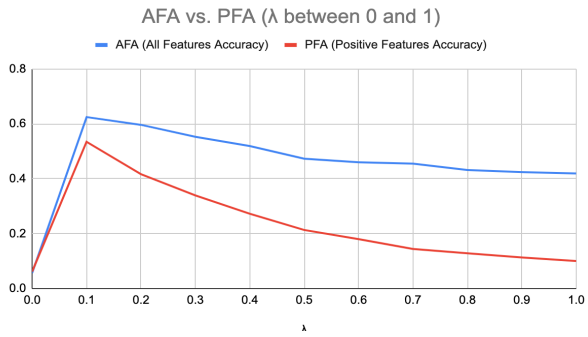
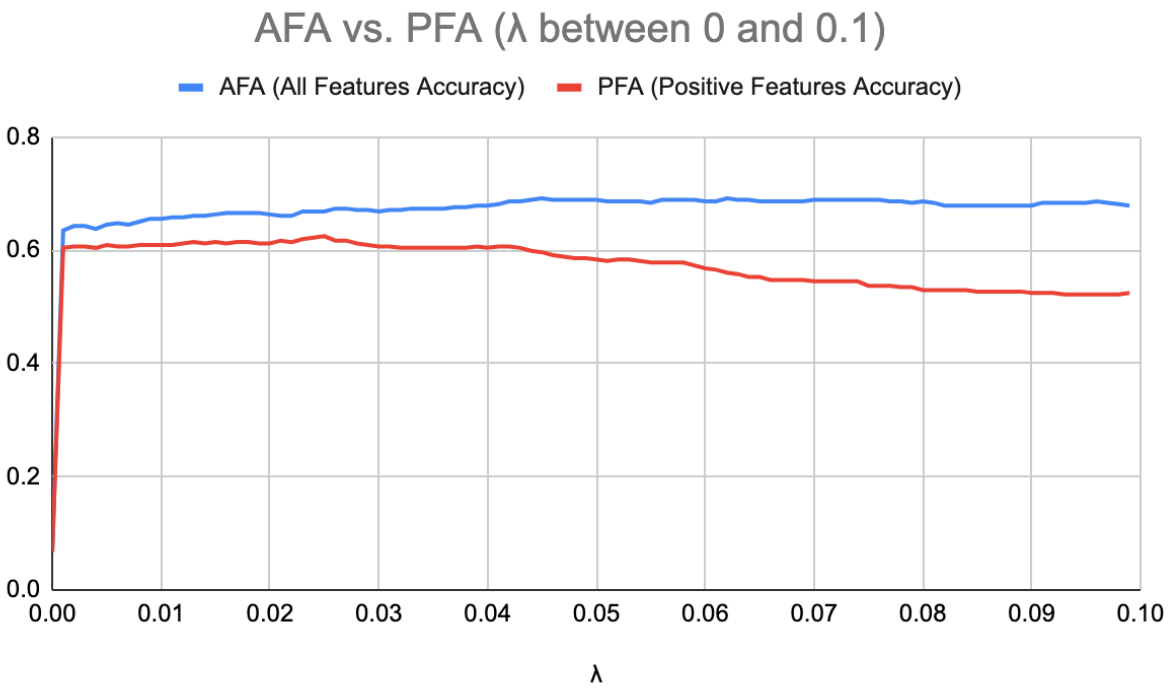
David D'Attile
Prof. Kauchak
CS158
11 March 2022

Assignment 07b: Naive Bayes Classifier

1 and 2. *What is the optimal lambda for the NB classifier using all features and the NB classifier using only positive features? Show your experimental data for how you found this:*







The raw data used to generate the above charts can be found at https://docs.google.com/spreadsheets/d/1huGG-YEDL5_x7QvvTPhFWB8ndHJhRgGF9pQShTn5QAE/edit?usp=sharing.

I found the optimal λ value for our AFA classifier approach to be 0.045 and the optimal λ value for the PFA approach to be 0.025 for the wine dataset.

To do this, I first divided the dataset into an 80/20 train/test split, then I measured the accuracy performance of the classifiers with integer λ values between 0 and 10 and found 1 to be the best for both classifiers (top left chart). Next, I measured the accuracy performance of the classifiers with λ values between 0 and 1 at a 0.1 step with the same data and found 0.1 to be the best for both classifiers (top right chart). Finally, since I had narrowed down the optimal λ range to between 0 and 0.1, I measured the accuracy of the classifiers for λ values in this range at a step of 0.001 for a total of 100 experiments (bottom middle chart).

We can see that small λ values (between 0.001 and 0.1) yield the best accuracy results for both Naive Bayes classifier approaches on the wine dataset. Interestingly, especially with the chart in the top left, we see that larger λ values degrade the performance of the PFA approach more rapidly than that of the AFA approach.

3. *Which version of the NB classifier is better (all features vs. positive features)? Include one or two sentences that describe how you came to this conclusion. If you use a performance argument, make sure to use statistical tests to justify your answer:*

fold	defaultAllFeatsAccuracy	defaultPosFeatsAccuracy	bestAllFeatsAccuracy	bestPosFeatsAccuracy
0	0.664948	0.608247	0.654639	0.603093
1	0.623711	0.608247	0.608247	0.623711
2	0.654639	0.634021	0.701031	0.634021
3	0.57732	0.592784	0.57732	0.597938
4	0.603093	0.56701	0.623711	0.561856
5	0.57732	0.556701	0.592784	0.57732
6	0.64433	0.582474	0.680412	0.56701
7	0.582474	0.520619	0.57732	0.494845
8	0.623711	0.551546	0.639175	0.56701
9	0.61809	0.603015	0.673367	0.61809
Average	0.6169636	0.5824664	0.6328006	0.5844894

T-Test	P-Value
default λ AFA vs. PFA:	0.003604820074
best λ AFA vs. PFA:	0.006004665745

After performing two 10-fold cross validation paired t-tests between the AFA and PFA Naive Bayes approach (one test with the default λ value of 0.01 for both classifiers and one with λAFA = 0.045 and λPFA = 0.025), we can see that the AFA approach is significantly better performing than the PFA approach. In both cases, the p-value for each t-test is less than 0.01, an indicator of strong performance differences between the two approaches. Further, we can see that the average accuracies are higher for the AFA approaches compared to their counterpart PFA approaches.