CS57300: Homework 3

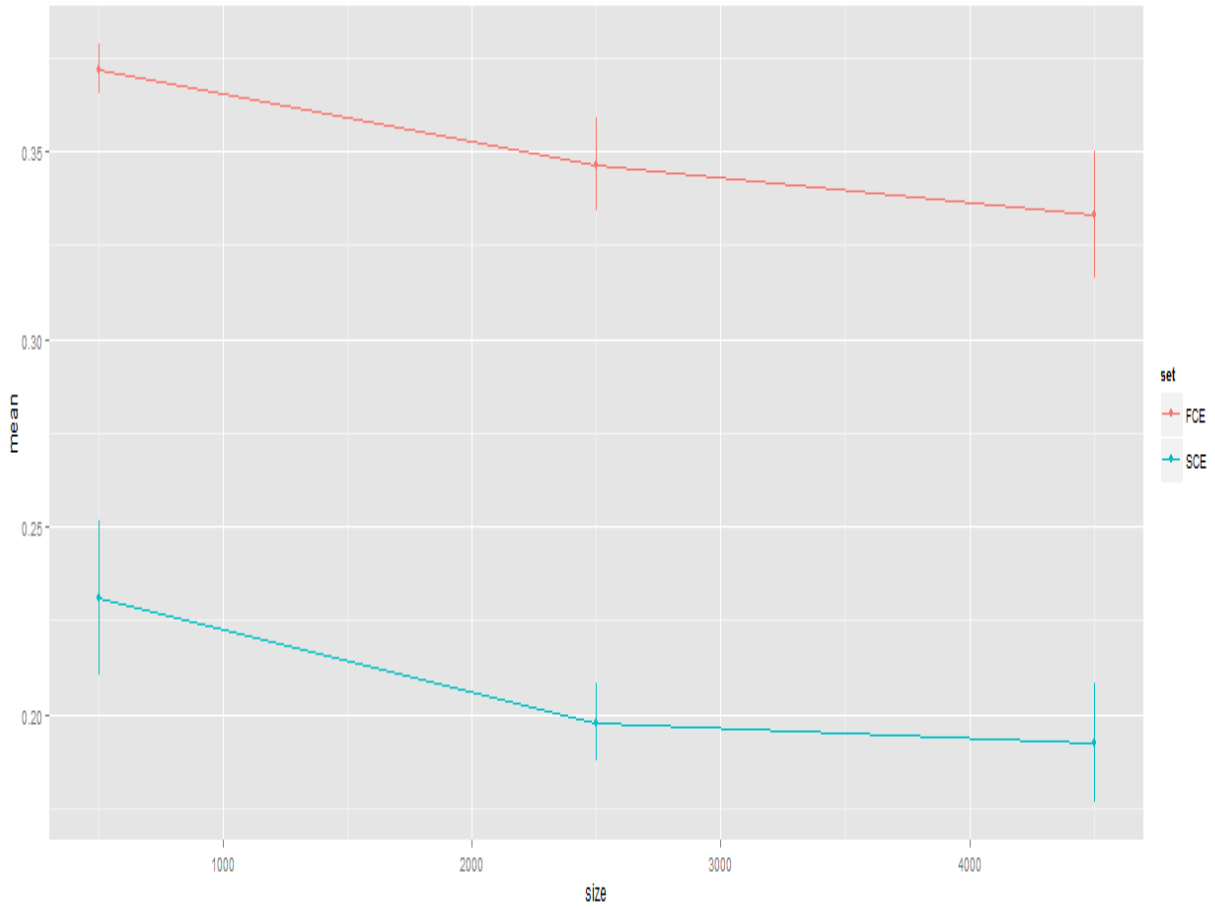
Hasini, Urala Liyanage Dona Gunasinghe

1 Analysis

Following table lists the mean and standard deviation of zero-one loss of naive base classifier when trained over data sets of different sizes from both funny_data and stars_data. Please note that SCE stands for star data (i.e: isPositive classification task) and FCE stands for funny data (i.e: isFunny classification task).

size	set	mean	sdv
500	FCE	0.3720667	0.006765234
2500	FCE	0.3466800	0.012620048
4500	FCE	0.3332000	0.016833300
500	SCE	0.2308889	0.020633665
2500	SCE	0.1978800	0.010315115
4500	SCE	0.1922000	0.015784803

Following graph illustrates the learning curves: training set size vs. mean of zero-one loss with standard deviation as error-bars for each data set (i.e: for each classification task).

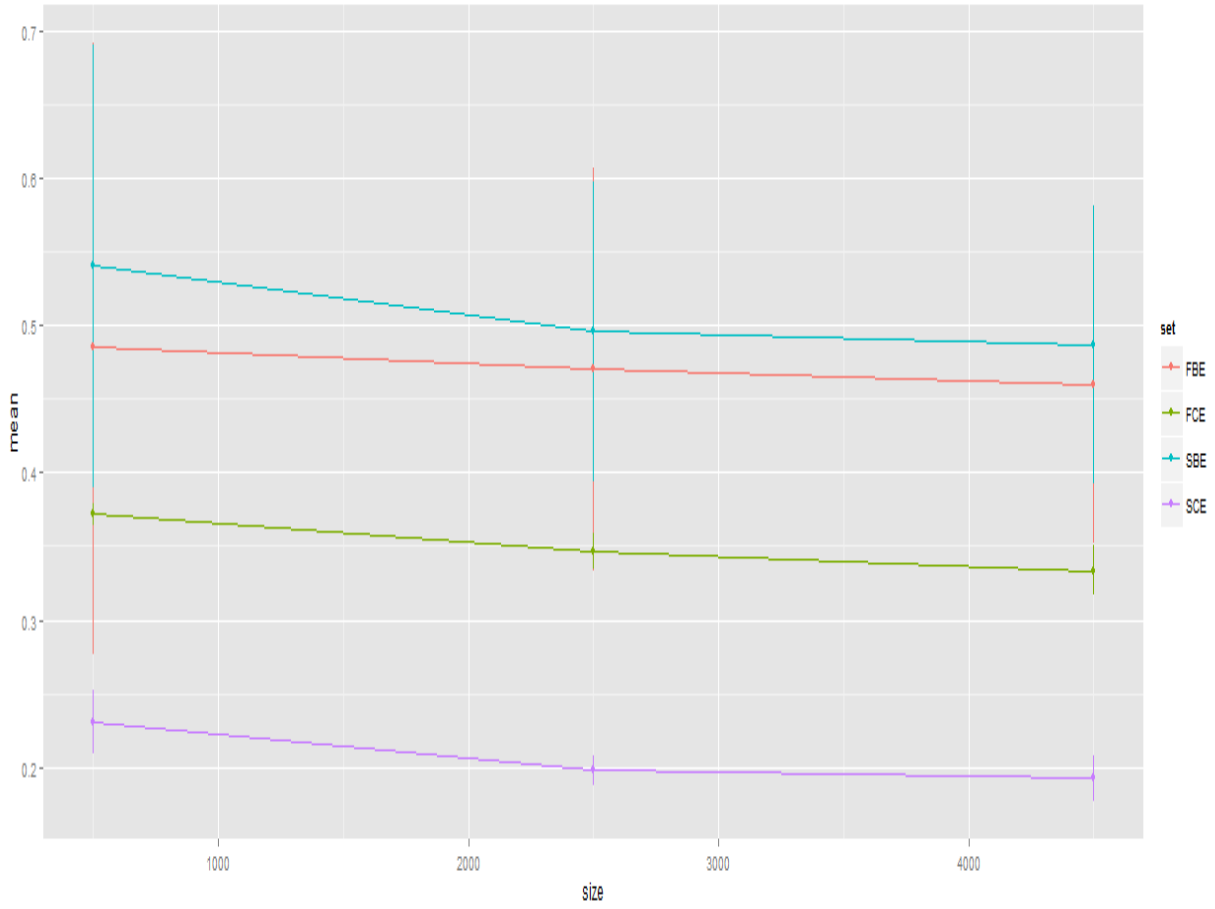


As we can observe, mean zero-one loss gets reduced as the training set size increases, for both the curves, as it is expected. This is because when the number of training samples increases, NBC learns the features and parameter estimates better. However, interestingly, the standard deviation increases as the training set size increases for both the curves. Furthermore, NBC learns and classifies better on star data (i.e: isPositive classification task) than on funny data.

In addition to the data in the previous table, following table also includes the mean and standard deviation of baseline default error over data sets of different sizes from both funny_data ans stars_data. In addition to the aforementioned notation, please note that SBE stands for baseline error of star data (i.e: isPositive classification task) and FBE stands for baseline error of funny data (i.e: isFunny classification task)

size	set	mean	sdv
500	FCE	0.3720667	0.006765234
500	FBE	0.4852444	0.207074537
2500	FCE	0.3466800	0.012620048
2500	FBE	0.4702000	0.136933298
4500	FCE	0.3332000	0.016833300
4500	FBE	0.4594000	0.108085337
500	SCE	0.2308889	0.020633665
500	SBE	0.5413778	0.149629000
2500	SCE	0.1978800	0.010315115
2500	SBE	0.4957600	0.101965359
4500	SCE	0.1922000	0.015784803
4500	SBE	0.4874000	0.094340023

Following graph plots the mean of baseline default error curves vs. training set size, along with the previously plotted zero-one loss curves of NBC, for both the data sets. Notation is the same as mentioned above.



As we can clearly see, learned NBC model performs better than the default prediction in both the cases. Furthermore, the gap between the baseline default error and the zero-one loss of NBC is higher w.r.t star data (i.e: isPositive classification task).