ENSF 619 Machine Learning Project Release 1

David Tang, Waley Chen, Karmveer Sidhu

1 Data Preparation

For the IMDb data set, the images were spread across 9 different archives of 30GB each. The meta data was created using Matlab and we used pandas and numpy to convert the mat file into a data frame. From there, we chose the first 10,000 pictures in the data frame. Each photo contained as string of the directory path which was cleaned and loaded into scikit image as grayscale. Afterwards, it was processed and cropped using the face_location coordinates. For the purposes of this release, we resized all the images to the same size without considering how this would downscale or upscale each image. It was necessary to load the pictures as grayscale and flatten the numpy data to ensure no issues arise when using the SVM model.

2 Data Analysis

We created an SVM model in order to predict gender. We ran our model for different sample sizes of 1000, 3000, 5000, 7000, 10000 images. Afterwards, we determined the testing and training accuracy, as well as the precision and recall values for both genders.

To our surprise, the training accuracy decreased with sample size. It is possible that the first 1000 images were easier to predict or could be part of the training error. However, as sample size increases, we expect the training accuracy to converge to a certain point. This appears to be the case as illustrated in figure 1.

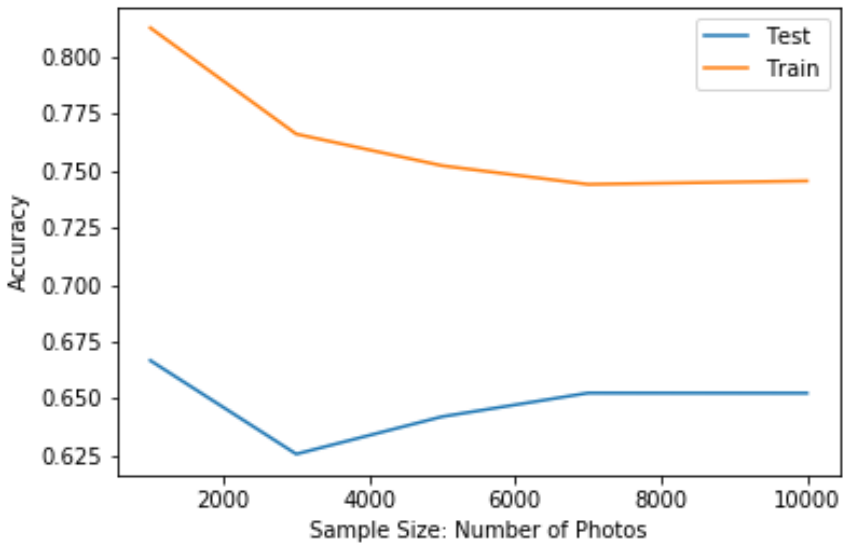

Figure 1: Accuracy of SVM Model

From figure 2, precision decreases for male and increases for female. It is possible that these values will converge as sample size increases. Also, it is worth noting that there is almost a 2:1 ratio of male photos compared to female which could affect the result of the model. The recall values are similar for male and female.

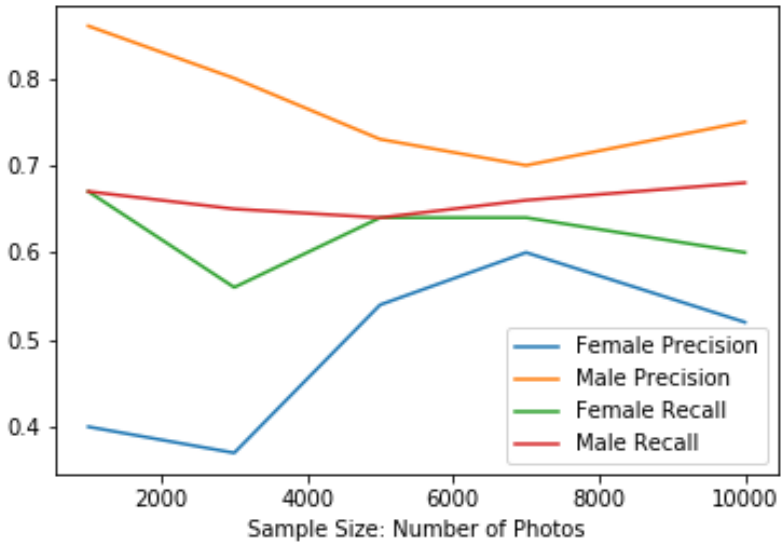

Figure 2: Precision and Recall of SVM Model