MSML 603 Project 1

Colleen Aulton
Prashant Swarnapuri

January 25, 2021

1 Methods

Both datasets (Expressions and Poses) were labeled with their expression type (0 for neutral, 1 for expression) or subject (data set limited to 5 subjects out of 68 possible).

In order to classify expressions (smiling vs. neutral) all 200 samples were loaded from the data.mat files for each expression. For each expression, the data was split into test and train data sets using a test-train split ratio for each run of the program.

For the identification of subjects problem, the 5 out of the 68 subjects were chosen from the pose.mat data set. Depending on the ratio of test-train split, each subjects pictures were divided in to test and train data set.

For all methods of classification the data was processed down to lower dimensions using the Principal Component Analysis (PCA) method. The exact amount of dimensions were tweaked across runs to observe the impact on the accuracy.

Two classification methods were used for both problems - a) Naive Bayes classifier and b) K-Nearest Neighbor (K-NN) Method

For the Naive-Bayes method, accuracy of the model was measured against various values of PCA dimensions and test-train split ratio.

For the K-NN method, accuracy of the model was measured against various values of K and test-train split ratio.

Both authors developed code independently and observations were combined subsequently. Implementation differences are expected. Differences where applicable are highlighted in the observations.

Colleen Aulton's observations will be referred to as Model 1, and Prashant Swarnapuri's observations will be referred to as Model 2.

2 Code Repository

Colleen Aulton (Model 1)
https://drive.google.com/drive/folders/1YNGqDDswn73m4CADTKvTQrNmtKa9i477
Prashant Swarnapuri (Model 2)
https://github.com/spprashant/msml603 (Refer to README.md for details)

Both sets of code are expected to be run in a Jupyter Notebook environment using a Python 3 kernel.

3 Observations

Expression Classifier - data.mat

Naive Bayes For the Naive Bayes classifier, the Expressions dataset (data.mat) performed the best at an 80%/20% train/test split and lower numbers of PCA components.

Both set of codes, confirmed that the model was optimal when training split was higher than 80% and PCA dimensions were limited to 40 or less.

In Model 1, the maximum accuracy achieved was 95% at an 80%/20% train/test split and 10 PCA components.

Naive Bayes Expressions Dataset		Train/Test Split					
		95/5	90/10	80/20	70/30	60/40	50/50
PCA Components	10	90%	90%	95%	91%	88%	89%
	20	90%	90%	93%	89%	89%	89%
	30	90%	90%	92%	87%	87%	88%
	40	90%	93%	90%	85%	86%	84%
	50	90%	88%	89%	83%	84%	84%

Figure 1: Naive Bayes Classifier Accuracy (Expressions Data set) - Model 1

In Model 2, the maximum accuracy achieved was 88.75% at an 80%/20% train/test split and 20 PCA components.

Naive Bayes Expressions Dataset		Train/Test Split					
		95/5	90/10	80/20	70/30	60/40	50/50
PCA Components	10	80%	77.5%	77.5%	78.33%	80%	78.35%
	20	85%	87.5%	**88.75%**	85%	88.12%	87.5%
	30	75%	80%	82.5%	83.33%	83.75%	83.5%
	40	75%	77.5%	77.5%	80%	83.12%	80.5%
	50	70%	80%	80%	79.17%	79.375%	80.5%

Figure 2: Naive Bayes Classifier Accuracy (Expressions Data set) - Model 2

K-NN For the KNN classifier, the Expressions dataset performed the best at a higher number of neighbors, with a maximum accuracy achieved of 90%.

In Model 1, the maximum accuracy was achieved at both a 95%/5% train/test split and an 80%/20% train/test split, though the 90%/10% train/test split experienced a slight decrease in accuracy (88%), which could be due to the random state of train/test split. The number of PCA components used was determined to be the minimum number such that 95% of the variance in the data is retained.

KNN Expressions Dataset		Train/Test Split			
		95/5	90/10	80/20	75/25
Number Neighbors	1	70%	58%	60%	57%
	2	55%	65%	70%	69%
	3	75%	80%	83%	77%
	4	70%	78%	81%	74%
	5	80%	90%	88%	82%
	10	80%	85%	85%	81%
	15	90%	88%	90%	88%

PCA: Using minimum number of components such that 95% of variance is retained

Figure 3: K-NN Classifier Accuracy (Expressions Data set) - Model 2

In Model 2, the maximum accuracy of 87.5% was achieved at both a 80%/20% train/test split, however changes were not significant between 0.95 training split and 0.75 training split. Number of PCA components was hard coded at 10 for this analysis.

KNN Expressions Dataset		Train/Test Split			
		95/5	90/10	80/20	75/25
Number Neighbors	1	75%	80%	81.25%	85%
	2	70%	75%	81.25%	77%
	3	70%	80%	86.25%	77%
	4	80%	80%	86.25%	81%
	5	80%	85%	86.25%	85%
	10	80%	85%	**87.5%**	87%
	15	85%	82.5%	85%	81%

Figure 4: K-NN Classifier Accuracy (Expressions Data set) - Model 2

Face Identification - poses.mat

Naive Bayes In Model 1, the Poses dataset (limited to 5 subject classification labels) performed with 100% accuracy when at least 90% of the dataset was used for training with the Naive Bayes classifier. This held true at all numbers of PCA components tested. Accuracy was generally higher with fewer PCA components (20 or less) and a larger training set. Given the small dataset (65 observations when limited to 5 subjects), a high accuracy result is expected with a very small testing set.

Naive Bayes Poses Dataset (5 subjects)		Train/Test Split					
		95/5	90/10	80/20	70/30	60/40	50/50
PCA Components	10	100%	100%	85%	75%	92%	73%
	20	100%	100%	92%	75%	92%	73%
	30	100%	100%	85%	70%	85%	52%
	40	100%	100%	85%	70%	NA	NA
	50	100%	100%	69%	NA	NA	NA

Figure 5: Naive Bayes Classifier Accuracy (Poses Data set) - Model 1

In Model 2, the Poses dataset (limited to 5 subject classification labels) performed with 100% accuracy when train split was higher than 80 with PCA components greater than 30. The model performed poorly with a training split of below 70.

Naive Bayes Poses Dataset (5 subjects)		Train/Test Split					
		95/5	90/10	80/20	70/30	60/40	50/50
PCA Components	10	80%	80%	86.66%	70%	20%	20%
	20	80%	80%	93.33%	85%	20%	20%
	30	**100%**	**100%**	**100%**	70%	20%	20%
	40	**100%**	**100%**	**100%**	70%	20%	20%
	50	**100%**	**100%**	**100%**	65%	20%	20%

Figure 6: Naive Bayes Classifier Accuracy (Poses Data set) - Model 2

K-NN In Model 1, The Poses dataset (again limited to 5 subject classification labels) performed the best at a lower number of neighbors. The maximum accuracy of 86% was achieved with 3 or less neighbors and a 90%/10% train/test split. The accuracy score decreased significantly as the number of neighbors was increased, reaching only 29% with 15 neighbors at the same 90%/10% train/test split. The number of PCA components used was determined to be the minimum number such that 95% of the variance in the data is retained.

KNN Poses Dataset (5 subjects)		Train/Test Split			
		95/5	90/10	80/20	75/25
Number Neighbors	1	75%	86%	69%	71%
	2	75%	86%	69%	71%
	3	50%	86%	62%	71%
	4	50%	71%	69%	65%
	5	50%	71%	69%	76%
	10	25%	43%	54%	65%
	15	50%	29%	46%	53%

PCA: Using minimum number of components such that 95% of variance is retained

Figure 7: K-NN Classifier Accuracy (Poses Dataset) - Model 1

In Model 2, accuracy was generally on the lower side compared to the Bayes model. Accuracy was highest when the train data split was around the 75-80 range and optimal when number of neighbors was equal to 10.

KNN Poses Dataset PCA_DIM=10		Train/Test Split			
		95/5	90/10	80/20	75/25
Number Neighbors	1	**80%**	**80%**	73.33%	73.33%
	2	60%	60%	73.33%	73.33%
	3	40%	40%	60.0%	60.0%
	4	40%	40%	60.0%	60.0%
	5	40%	40%	66.66%	66.66%
	10	60%	60%	**80%**	**80%**
	15	40%	40%	66.66%	66.66%

Figure 8: K-NN Classifier Accuracy (Poses Dataset) - Model 2