Mehmet Mert Bezirgan

Engr 421 – homework 3

Problem: Naïve Bayes Classifier

Solution:

For this algorithm we have image of letters of 320 pixels (16x20) given us as a vector. Our task is to implement a naïve Bayes classifier for letter prediction. We have vectors consist of 1's and 0's and have 320 features. We are going to calculate sample means and prior probabilities and using them we are going to write score function.

$$\begin{aligned} g_i(\boldsymbol{x}) &= \log p(\boldsymbol{x}|C_i) + \log P(C_i) \\ &= \sum_j \left[x_j \log p_{ij} + (1 - x_j) \log(1 - p_{ij}) \right] + \log P(C_i) \end{aligned}$$

Figure 1 formula for score function for class I

When we calculate sample means for each letter we got values as:

	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0.04	0.04	0.04	0.16	0.2	0.16	0.12
1	0.04	0.24	0.24	0.2	0.12	0.08	0.12	0.16	0.24	0.32
2	0	0	0	0	0	0.12	0.2	0.24	0.4	0.56
3	0.12	0.44	0.4	0.16	0.12	0.08	0.08	0.08	0.12	0.08
4	0	0.12	0.12	0.08	0.12	0.16	0.12	0.04	0.12	0.12

310	311	312	313	314	315	316	317	318	319
0.6	0.68	0.72	0.72	0.68	0.56	0.72	0.68	0.68	0.64
0.12	0.4	0.72	0.8	0.8	0.88	0.88	0.8	0.56	0.24
0	0	0	0	0.08	0.16	0.2	0.6	0.88	0.8
0.96	0.88	0.8	0.84	0.8	0.64	0.56	0.28	0.16	0.08
0.24	0.2	0.08	0.04	0	0.04	0.28	0.32	0.48	0.44

Figure 2 calculated sample means

Our prior probabilities are equal and all of them are 0.2. Then we are going to calculate scores of all training set. Our score function is going to return an array consist of 5 score values and maximum of that scores index + 1 will be our class label. We make calculations and print confusion matrix of training data.

```
training performance
y_test       1.0  2.0  3.0  4.0  5.0
y_pedicted
1.0           25    0    0    0    0
2.0            0   24    1    0    1
3.0            0    0   24    0    0
4.0            0    1    0   25    0
5.0            0    0    0    0   24
```

Figure 3 Training performance

As we can see predictions for our training data is accurate enough to proceed. We have our test data that we separated before consist of 14 letter images of 5 classes (total 70). We calculate scores for our testing data and see performance of our classifier with confusion matrix.

test performance

y_test y_pedicted	1.0	2.0	3.0	4.0	5.0
1.0	7	0	0	0	0
2.0	0	11	3	2	4
3.0	0	0	7	0	0
4.0	7	3	3	12	0
5.0	0	0	1	0	10

Figure 4 confusion matrix for test data

As we can see from this confusion matrix the accuracy of our algorithm is not good and number of false predictions are high.