ENGR 421 / Homework 1: Naive Bayes Classifier

Umur Berkay Karakaş

March 18, 2021

In homework 1, we are given 400 images which consist of 4096 pixels where each pixel is continuous between 0 and 1.

First, I created trainingset variable from the first 200 points in the data set and testset variable from the last 200 points in the data set. I assigned their corresponding class labels to ytraining and ytest. I also calculated number of classes by using np.max and prior probabilities by calculating the mean of each class label.

I found the means and deviations of each feature of data points for the corresponding class. In Naive Bayes Classifier, all features of a data point are assumed to be independent. Therefore, I didn't need to find the covariance matrix because the covariances of distinct features would be equal to zero. Instead, since they are independent, the likelihood function of x for class c would be equal to the multiplication of the probability of each feature of x given y = c times prior probability of class c. And then the loglikelihood function would be equal to:

$$g_c(x) = \sum_{i=1}^{4096} \log p(x_i|y=c) + \log P(y=c)$$

Since the features of the data points are continuous, I assumed that the data points are Gaussian RV and the score function has become:

$$g_c(x) = \sum_{i=1}^{4096} (-\frac{1}{2}\log(2\pi\sigma_c^2) - \frac{(x-\mu_c)^2}{2\sigma_c^2})_i + \log P(y=c)$$

Here, all x, σ and μ are 4096x1 vectors. And the score function is equal to the sum of the resulting vectors' features plus prior probability of class c. So, I created a function in my code accordingly for the score function which takes dataset as input and returns the scores of each data points for class 1 and class 2. Then, I plotted the confusion matrices for 200 points in the training set and 200 points in the test set. The results were:

y_hat	1	2
y_train		
1	18	2
2	24	156

Figure 1: Confusion matrix for training set

y_hat	1	2
y_test		
1	15	5
2	19	161

Figure 2: Confusion matrix for test set