DASC521: Introduction to

Machine Learning

Fall 2021

Instructor: Mehmet Gönen

Homework 01- Multivariate Parametric Classification

Gamze Keçibaş- 60211

22.10.2021

There are four main steps as data generation, parameter estimation, confusion matrix creating and plot decision boundaries and misclassified points. In first step, three different dataset is generated from bivariate Gaussian densities (1) by provided size, mean array and covariance matrix.

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \qquad (1)$$

These datasets are presented below:

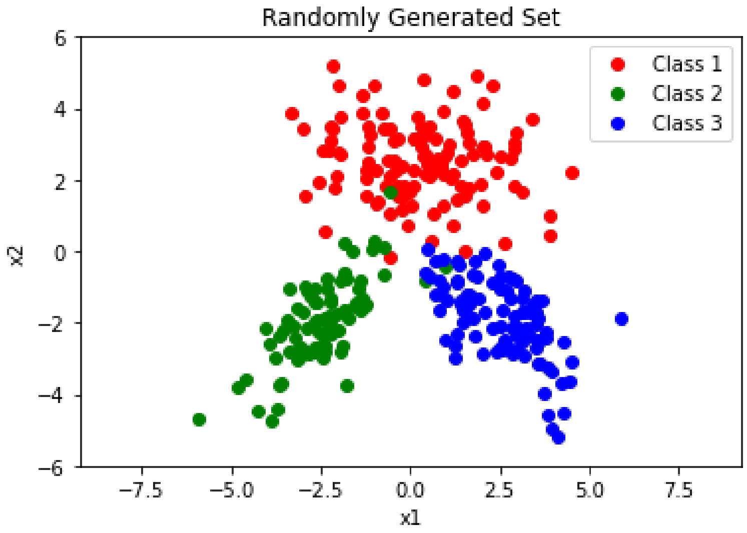


Figure 1: Randomly generated data points

After data points and their labels are generated; means, covariances and priors of these samples are calculated. Sample mean (2) and covariance formulas (3) are given below:

$$\widehat{\mu_i} = \sum_{j=1}^{N_i} x_j/N \qquad (2)$$

$$\widehat{\Sigma_i} = \frac{\sum_{j=1}^{N_i}(X_i-\bar{x})(Y_i-\bar{y})}{N-1} \qquad (3)$$

According to Equation 1, 2 and 3, the results are obtained:

```
Sample means

Class [1]:
[0.26563078 2.52716835]
Class [2]:
[-2.47809683 -1.95300192]
Class [3]:
[ 2.56009664 -1.90556717]
```

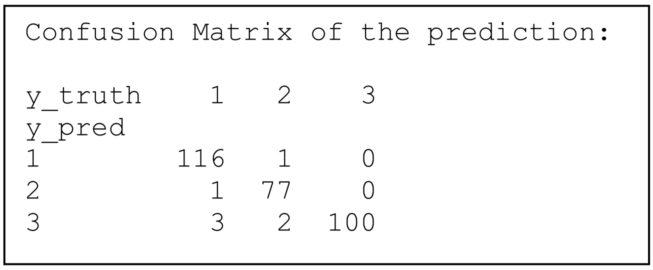
```
Sample covariances

Class [1]:
[[ 2.66148669 -0.23424653]
 [-0.23424653  1.16477455]]
Class [2]:
[[1.19581994 0.92481706]
 [0.92481706 1.39635535]]
Class [3]:
[[ 1.23731467 -0.70749062]
 [-0.70749062  1.13882277]]
```

```
Sample priors

Prior probability of
Class [1]:  0.4
Prior probability of
Class [2]:  0.2666666
6666666666
Prior probability of
Class [3]:  0.3333333
333333333

Total probability=1.0
```

When priors are calculated, the total of the priors is checked for validation. The result should be equal zero for a consistent solution. After these parameters are defined, score functions are calculated by Equation 4:

$$g_c(x) = -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log\left(\left|\widehat{\Sigma_c}\right|\right) - \frac{1}{2}(x-\widehat{\mu_c})^T \widehat{\Sigma}_c^{\,-1}(x-\widehat{\mu_c}) + \log\hat{P}(y=c) \quad (4)$$

These predictions are used to create a confusion matrix. The matrix is created by *pandas* library where y_pred is predictions and y_truth is generated data points:



```
Confusion Matrix of the prediction:

y_truth     1    2     3
y_pred
1         116    1     0
2           1   77     0
3           3    2   100
```

In final step, posteriors are calculated and decision boundaries are determined using estimated parameters. Quadratic discriminant method is applied to draw the boundaries.

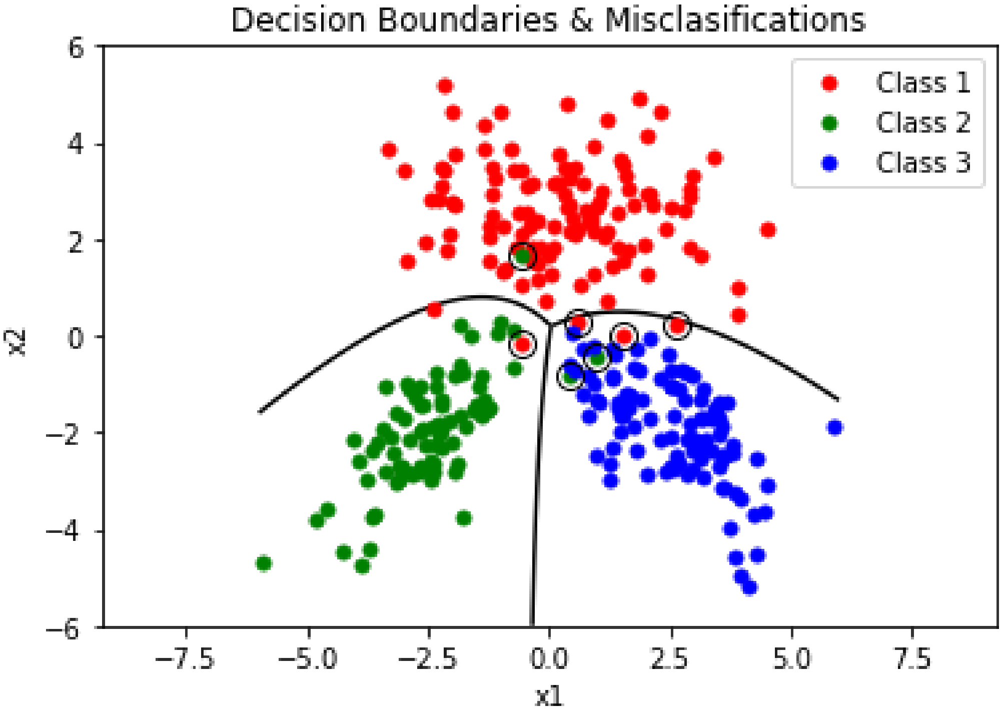


Figure 2: Decision Boundaries and Misclassified Points

Black lines shows decision boundaries for the model. Misclassified points are also marked black circles in the Figure 2.