Q4

February 11, 2018

0.1 Gaussian Discriminant Analysis

In this problem, we implement GDA for separating out salmons from Alaska and Canada. Alaska is assigned identifier 0 and Canada is assigned 1.

Firstly we assume that both normal distributions have same covariance matrices and calculate the poisson parameter, means and the covariance matrix. Following are the parameters learned, first value corresponds to the first feature.

```
Phi =  0.5


mu_0 =  [-0.75529433  0.68509431]


mu_1 =  [ 0.75529433 -0.68509431]


Covariance matrix =  [[ 0.42953048 -0.02247228]
 [-0.02247228  0.53064579]]
```

Expanding decision boundary expression to an explicit form gives a complicated expression in terms of covariances, means and determinants of covariances. For same covariance matrix assumption, expression is linear, let it be $k_1 * x_1 + k_2 * x_2 + k_3 = 0$

$$det = np.linalg.det(cov)$$

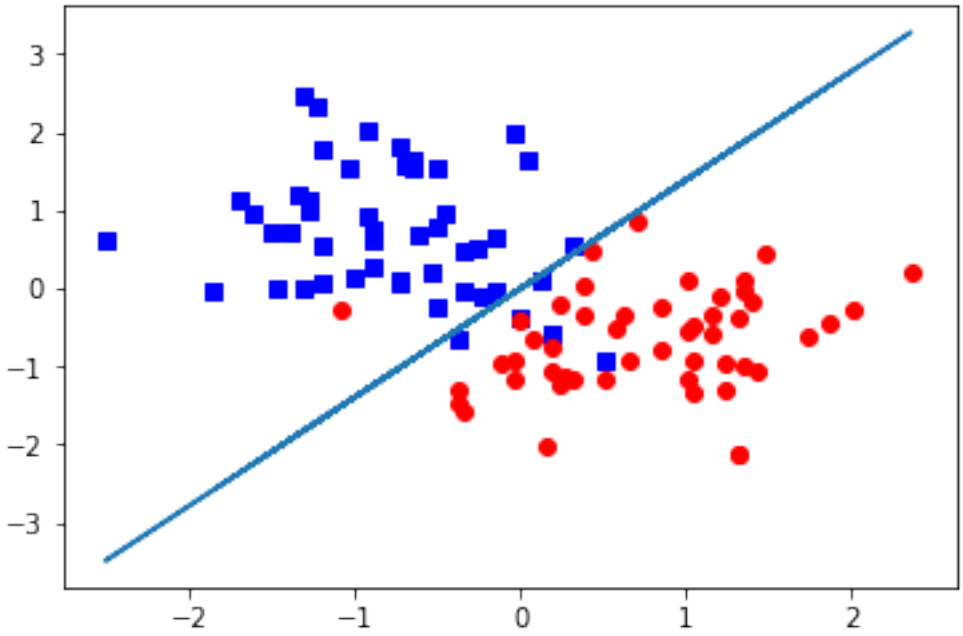
$$k_1 = (1/np.abs(det)) * (-2 * mu_1[0] * cov[1,1] + 2 * mu_1[1] * cov[0,1] + 2 * mu_0[0] * cov[1,1] - 2 * mu_0[1] * cov[0,1])$$

$$k_2 = (1/np.abs(det)) * (-2 * mu_1[1] * cov[0,0] + 2 * mu_1[0] * cov[0,1] + 2 * mu_0[1] * cov[0,0] - 2 * mu_0[0] * cov[0,1])$$

$$k_3 = (1/np.abs(det)) * (mu_1[0] * mu_1[0] * cov[1,1] - 2 * mu_1[0] * mu_1[1] * cov[0,1] + mu_1[1] * mu_1[1] * cov[0,0] - mu$$

Following is a plot of the data and the linear decision boundary. Blue points correspond to Alaska and redpoints correspond to Canada.

```
Red = Canada , Blue = Alaska
```



After this we relax the assumption of same covariance matrices and allow them to be different. The new covariance matrices are as follows (other parameters are unchanged).

```
Cov_0 is  [[ 0.38158978 -0.15486516]
 [-0.15486516  0.64773717]]


Cov_1 is  [[ 0.47747117  0.1099206 ]
 [ 0.1099206   0.41355441]]
```

The expression for decision boundary is now quadratic, let it be $l_1 * x_1^2 + l_2 * x_2^2 + l_3 * x_1 * x_2 + l_4 * x_1 + l_5 * x_2 + l_6 = 0$

The exact expressions are given below.

$$det_0 = np.linalg.det(cov_0)$$

$$det_1 = np.linalg.det(cov_1)$$

$$l_1 = (1/np.abs(det_1)) * cov_1[1,1] - (1/np.abs(det_0)) * cov_0[1,1]$$

$$l_2 = (1/np.abs(det_1)) * cov_1[0,0] - (1/np.abs(det_0)) * cov_0[0,0]$$

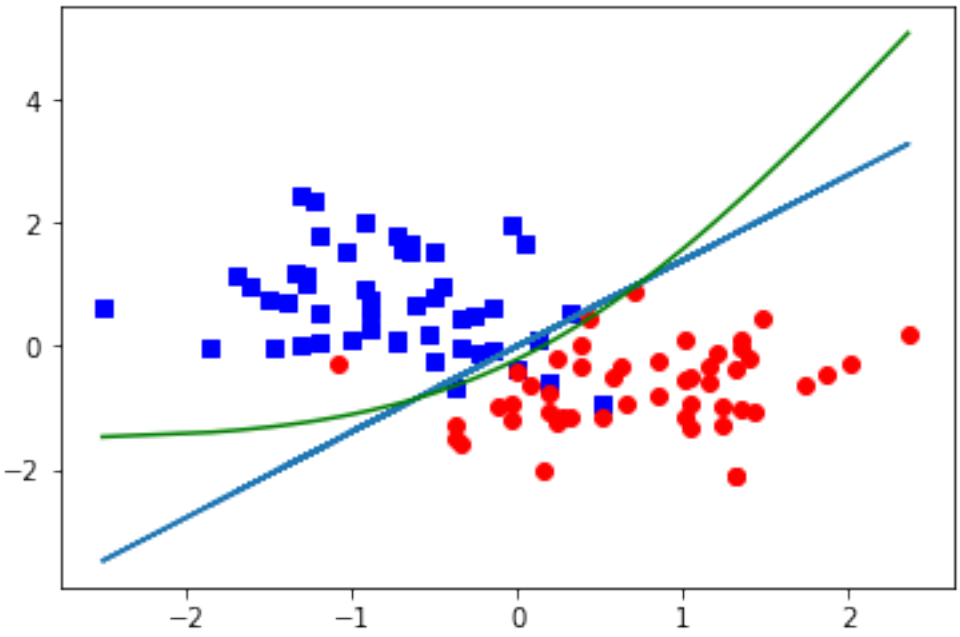
$$l_3 = -2 * (1/np.abs(det_1)) * cov_1[0,1] + 2 * (1/np.abs(det_0)) * cov_0[0,1]$$

$l_4 = (-2 * (1/np.abs(det_1)) * mu_1[0] * cov_1[1,1] + 2 * (1/np.abs(det_1)) * mu_1[1] * cov_1[0,1] + 2 * (1/np.abs(det_0)) * mu_0[0] * cov_0[1,1] - 2 * (1/np.abs(det_0)) * mu_0[1] * cov_0[0,1])$
$l_5 = (-2 * (1/np.abs(det_1)) * mu_1[1] * cov_1[0,0] + 2 * (1/np.abs(det_1)) * mu_1[0] * cov_1[0,1] + 2 * (1/np.abs(det_0)) * mu_0[1] * cov_0[0,0] - 2 * (1/np.abs(det_0)) * mu_0[0] * cov_0[0,1])$

$l_6 = ((1/np.abs(det_1)) * mu_1[0] * mu_1[0] * cov_1[1,1] - (1/np.abs(det_1)) * 2 * mu_1[0] * mu_1[1] * cov_1[0,1] + (1/np.abs(det_1)) * mu_1[1] * mu_1[1] * cov_1[0,0] - (1/np.abs(det_0)) * mu_0[0] * mu_0[0] * cov_0[1,1] + (1/np.abs(det_0)) * 2 * mu_0[0] * mu_0[1] * cov_0[0,1] - (1/np.abs(det_0)) * mu_1[1] * mu_1[1] * cov_0[0,0] + np.log(np.abs(det_1/det_0)))$

Finally we plot the data, linear decision boundary and quadratic decision boundary on the same plot.

```
Red = Canada , Blue = Alaska
```



It is clear from the graph that the quadratic boundary seems to be a better fit as it takes care of a few points which protrude out of the linear boundary. We can also observe that the different spreading out of the two classes has an effect on the quadratic boundary. The blue points seem more vertically spread out than the red points and hence the boundary bends to accomodate those points.