Report on

Fisher's Linear Discriminant

Team

1. Sai Satvik Vuppala (2017B4A71449H)
2. Shreya Kotagiri (2017B3AA0296H)
3. Dhanush Karupakula(2017B3A71011H)

Introduction

We implement Fisher's Linear Discriminant algorithm to classify data points into binary classes. The entire dataset is used for training the algorithm. The boundary is arrived at by reducing the data to a single dimension from D-dimensional space, by maximizing the difference of means and minimizing sum of variances of the classes. The threshold (in 1D)/hyperplane (in D-dimensional space) is calculated from the point of intersection of the normal distributions of the projections of the points of the two classes onto the discriminant vector.

Libraries Used

1. Numpy
2. Pandas
3. Statistics
4. Matplotlib

Implementation

We proceed through the whole process in five different steps:

1. Finding the means of the classes

The idea behind the algorithm is that the boundary has the maximum distance from the mean points of the two classes. We find the means M1 and M2 of the two classes in the 3-dimensional space, both of which are 3x1 vectors. We project the 3D points onto a vector w that is parallel to the vector joining the means M1 and M2. So there are infinitely many possibilities for the vector w.

$$w \propto (M1-M2)$$

2. Finding Sw

To ensure all the points are classified correctly, it is important to not just maximize the distance between the means from the boundary, but also make sure the within class variances of the two classes is minimized.

$$\sum_{y=1} ((x_n - M_1) . (x_n - M_1)^T) + \sum_{y=0} ((x_n - M_2) . (x_n - M_2)^T)$$

3. We find W (the weights/feature vector)

Solving this optimization problem yields a result that w should be a vector of dimensions 3x1, that is proportional to product of S_w^{-1} and (M1-M2), where S_W is a 3x3 matrix.

$$w \propto S_w^{-1}(M1-M2)$$

4. Projecting the points on to the 1D space using W

$$w^T.X$$

5. We plot the gaussian curves to find out the intersection point (the threshold)

The threshold in 1D space is the intersection of the normal curves of the collapsed points of the two clusters onto the vector w. The intersection point belongs to the roots of the quadratic equation **$\log(f_1(x))-\log(f_2(x))=0$** where $f_1(x)$ and $f_2(x)$ are the two normal distributions.

$$f_1(x) = f_2(x)$$
$$\implies \log f_1(x) = \log f_2(x)$$
$$\implies \log f_1(x) - \log f_2(x) = 0.$$

If f_i has mean μ_i and standard deviation σ_i $(i = 1, 2)$ then

$$\log(f_i(x)) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_i^2) - \frac{1}{2}(x - \mu_i)^2/\sigma_i^2.$$

So

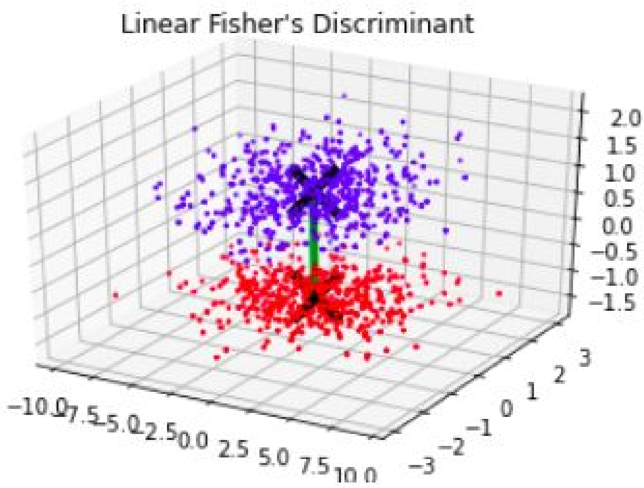
$$\log(f_1(x)) - \log(f_2(x)) = \frac{1}{2}[\log(\sigma_2^2) - \log(\sigma_1^2) + (x - \mu_2)^2/\sigma_2^2 - (x - \mu_1)^2/\sigma_1^2]$$
$$= \frac{1}{2}(Ax^2 + Bx + C)$$

where

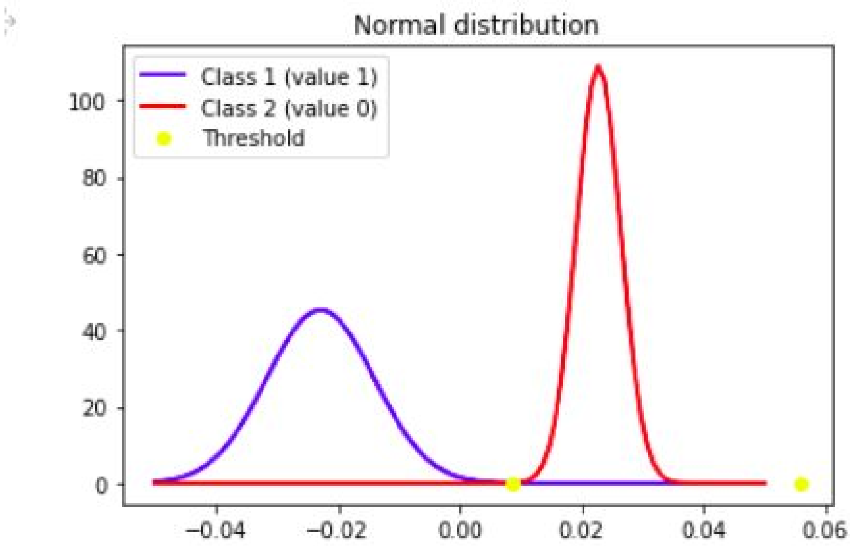
$$A = -1/\sigma_1^2 + 1/\sigma_2^2$$
$$B = 2(-\mu_2/\sigma_2^2 + \mu_1/\sigma_1^2)$$
$$C = \mu_2^2/\sigma_2^2 - \mu_1^2/\sigma_1^2 + \log(\sigma_2^2/\sigma_1^2).$$

source:https://stats.stackexchange.com/questions/311592/how-to-find-the-point-where-two-normal-distributions-intersect#:~:text=The%20two%20intersections%20are%20easiest,2(x)%3D0.

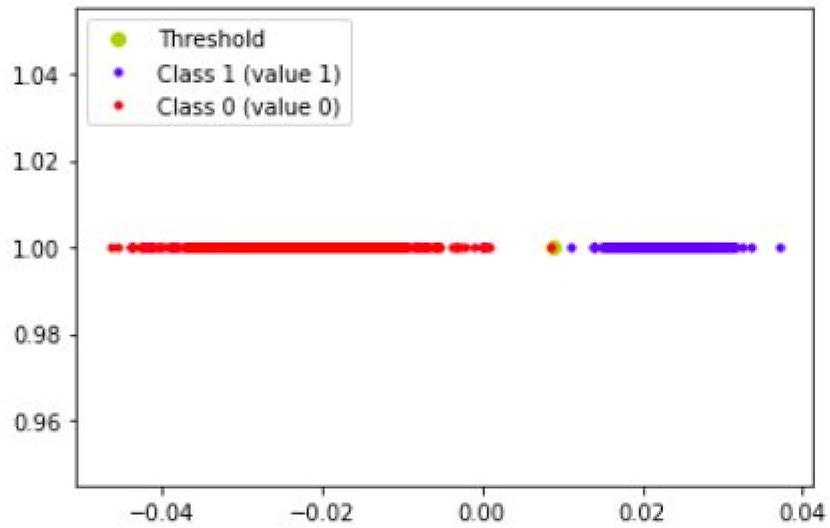
Plot of the higher dimensional data



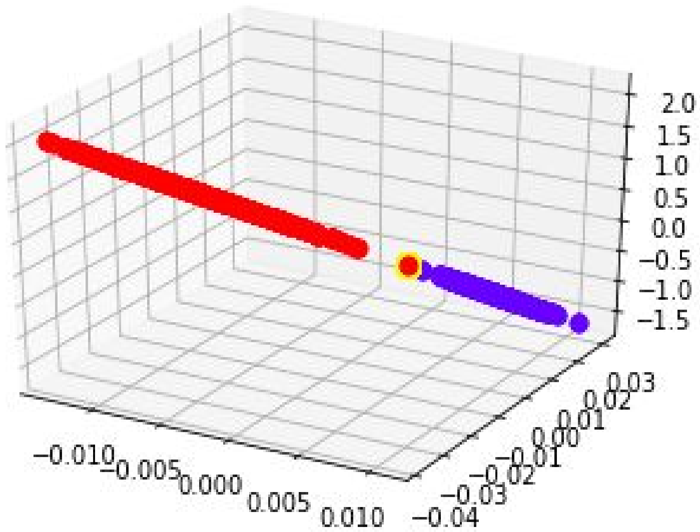
Normal distributions of the collapsed classes



Unit vector in 1D space



Unit vector 3D space



Hyperplane/ classifier boundary