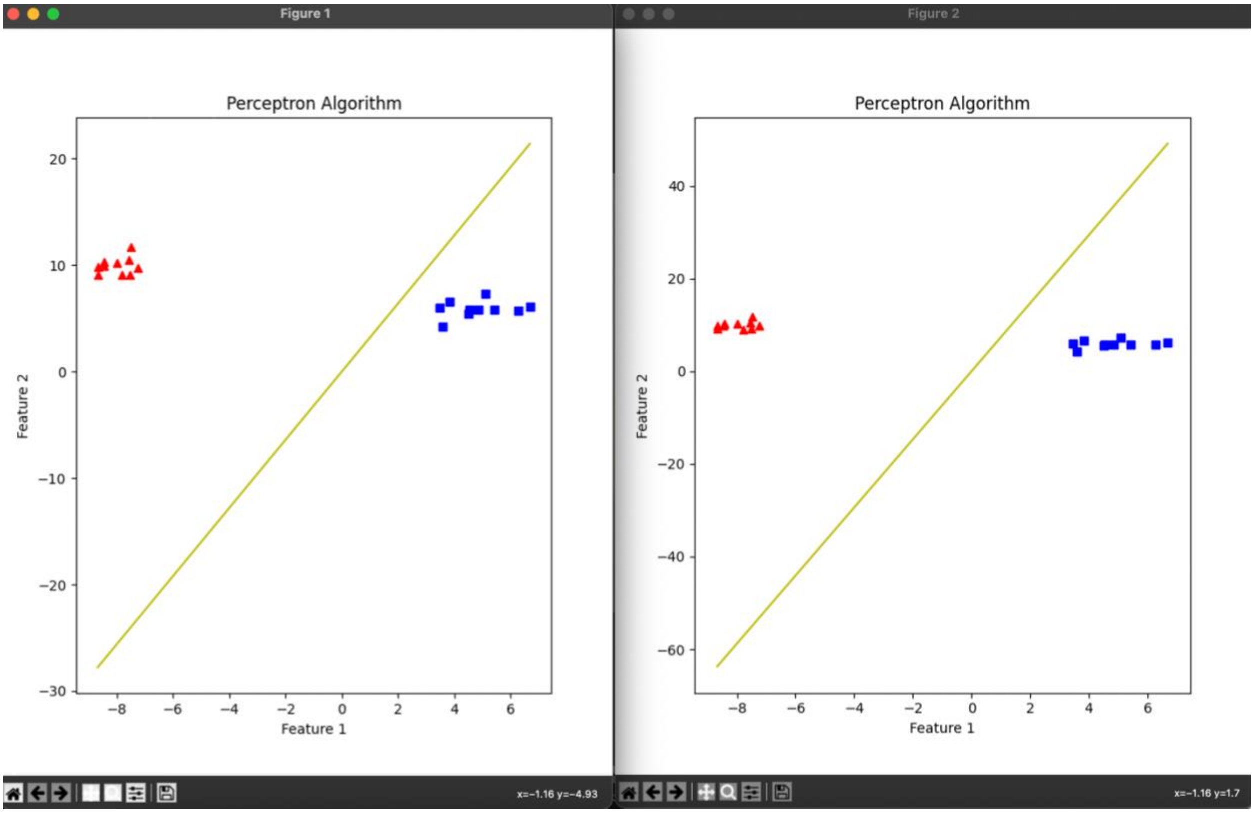
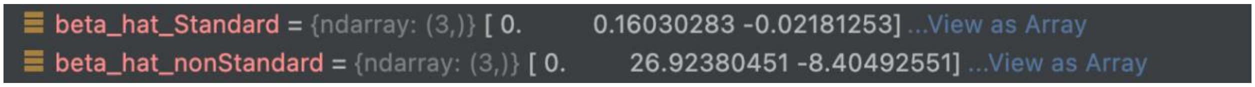
Please see code in the .py files.



If the data are separable and the two clusters are far enough, the two algorithms both work but will give different lines.
If we standardize the beta, the gradient will be zero if the norm of beta is large and the gradient will be extremely large if the norm is small enough. If the norm is close to zero the gradient will be NaN. The algorithm will no longer work.



```
beta_hat_Standard = {ndarray: (3,)} [ 0.          0.16030283 -0.02181253] ...View as Array
beta_hat_nonStandard = {ndarray: (3,)} [ 0.          26.92380451 -8.40492551] ...View as Array
```

Derivative calculation:

$$\min D(\beta,\beta_0) = -\sum_{i\in Mis} \frac{y_i(\beta^T x_i+\beta_0)}{\|\beta\|} = -\frac{\sum_{i\in Mis} y_i(\beta_1 x_1+\beta_2 x_2+\beta_0)}{\sqrt{\beta_1^2+\beta_2^2}}$$

$$\frac{\partial D}{\partial \beta_k} = -\sum_{i\in Mis} \frac{y_i x_k \cdot \sqrt{\beta_1^2+\beta_2^2} - y_i(\beta_1 x_1+\beta_2 x_2+\beta_0)\frac{\beta_k}{\sqrt{\beta_1^2+\beta_2^2}}}{\beta_1^2+\beta_2^2}$$

$$= -\sum_{i\in Mis} \frac{y_i x_k \cdot (\beta_1^2+\beta_2^2) - y_i(\beta^T x_i+\beta_0)\cdot \beta_k}{(\beta_1^2+\beta_2^2)\sqrt{\beta_1^2+\beta_2^2}}$$

$k=1,2$

$$\frac{\partial (\beta_1^2+\beta_2^2)^{\frac{1}{2}}}{\partial \beta_1} = \frac{\beta_1}{\sqrt{\beta_1^2+\beta_2^2}}$$

$$\frac{\partial D}{\partial \beta_0} = \sum_{i\in Mis} -y_i$$