M339G, April 19th, 2024.

Margins and Separating Hyperplanes.

Linear classifiers can be described geometrically as separating hyperplanes.

Any affine function $x \longmapsto \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ determines a hyperplane in $\mathbb{R}^p$ our feature space

More precisely, $\{x : \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = 0\}$ is a hyperplane splitting the space $\mathbb{R}^p$ into two half spaces:

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p > 0$$

and

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p < 0$$

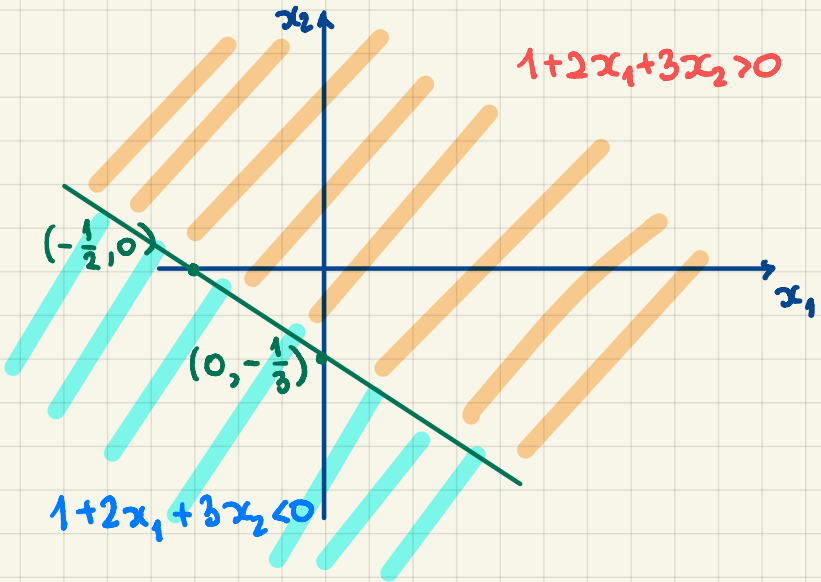
The vector $\vec{n} = (\beta_1, \beta_2, \ldots, \beta_p)$ is the normal vector of our hyperplane.

For a given hyperplane, we can always choose $\vec{n}$ so that

$$\|\vec{n}\| = 1$$

Of course, the β_0 coefficient will also need to be scaled.

Example. Consider $x \longmapsto 1 + 2x_1 + 3x_2$



If $x_1 = 0$, then $1 + 3x_2 = 0$

$x_2 = -\frac{1}{3}$

If $x_2 = 0$, then $1 + 2x_1 = 0$

$x_1 = -\frac{1}{2}$

Note: • If the hyperplane goes through the origin, then $\beta_0 = 0$.

For any point in the plane, the deviation between the point

$$x = (x_1, x_2, \dots, x_p)$$

and our hyperplane is w/ $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ equal to

$$x \cdot \beta = x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p$$

As we have seen in the previous example, the sign of the dot product specifies the direction (or the side).

• If $\beta_0 \neq 0$, the hyperplane does not go through the origin.

The deviation of the point x from the hyperplane

$$\beta_0 + x \cdot \beta$$

The sign tells us on which side of the hyperplane x lies.

Maximal margin classifier.

Suppose that we have a classification problem w/ two classes. We choose to encode one class as $Y = -1$ and the other class as $Y = 1$.

Our criterion for the best among all separating hyperplanes (if such exist) is to find the one w/ the largest possible margin around the hyperplane.

Optimization Problem.

The above task can be expressed as

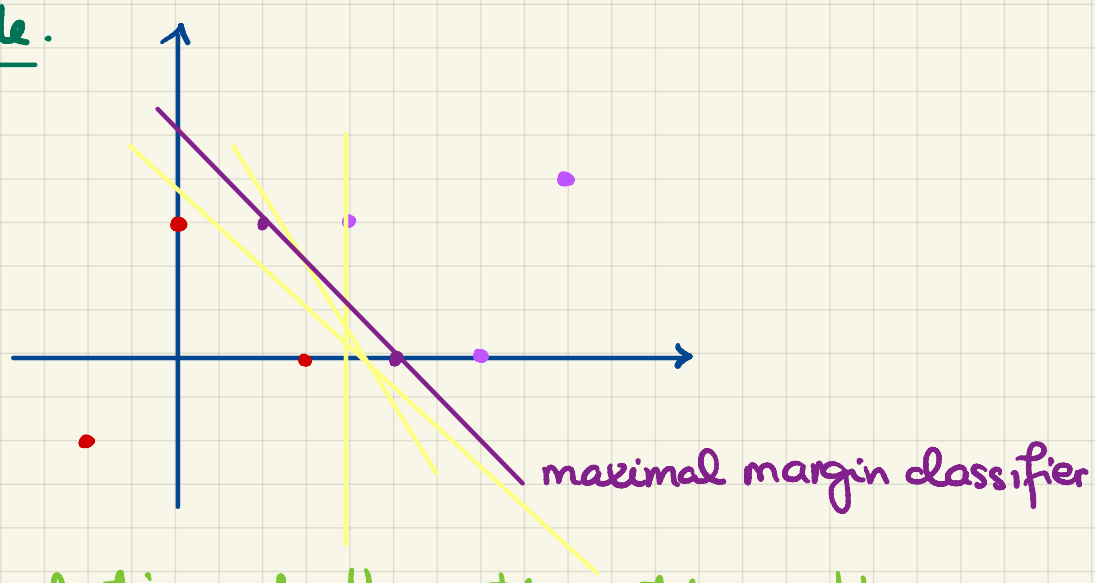
$$\max_{\beta_0, \beta_1, \dots, \beta_p} M$$

subject to $\sum_{j=1}^{p} \beta_j^2 = 1$

and $y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M$

for all $i = 1, \dots, n$

Example.



Reformulation of the optimization problem

Define a vector:

$$w = (w_1, w_2, \dots, w_p) = \frac{\beta}{M}$$

$$\min_{\beta_0, w} \frac{1}{2}\|w\|^2$$

subject to $y_i(\beta_0 + w_1 x_{i1} + w_2 x_{i2} + \cdots + w_p x_{ip}) \geq 1$

for all $i = 1, 2, \dots, n$

This is a quadratic optimization problem.

We introduce Karush-Kuhn-Tucker (KKT) multipliers

$\lambda_1, \lambda_2, \dots, \lambda_n$.

Now, our optimization problem is equivalent to

$$\max_{\lambda} \min_{\beta_0, w} \left(\frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \lambda_i \left(y_i(\beta_0 + w_1 x_{i1} + \cdots + w_p x_{ip}) - 1 \right) \right)$$

subject to $\lambda_i \geq 0$.