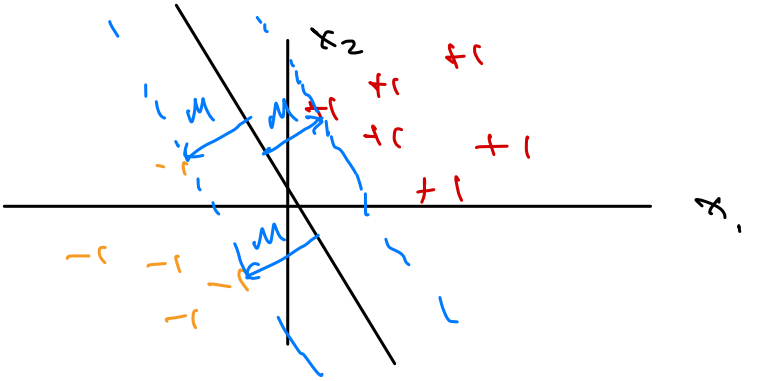
Recall

- Hyperplane is all $\underline{x} \in \mathbb{R}^p$ s.t. $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = 0$
- Note hyperplane is defined by $\beta_0, \beta_1, \ldots, \beta_p$, & splits space into halves
- Data $y_i \in \{-1, +1\}$ & $\underline{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ $i = 1, \ldots, n$

Suppose <u>linearly separable</u>



**Note** The maximum margin hyperplane is calculated by:

- Maximize $M$
  $\beta_0, \beta_1, \ldots, \beta_p$
- Subject to
  $\|\vec{\beta}\|_2^2 = \sum_{j=1}^{p} \beta_j^2 = 1$ and $y_i(\beta_0 + \vec{\beta}^T \vec{x}_i) \geq M \; \forall i = 1, \ldots, n$

$\vec{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \rightarrow$ no $\beta_0$, $p \times 1$

**Remark**

Restricting $\vec{\beta}$ to be length 1 is not a restriction bc any $\vec{x}$ that satisfies

$$\beta_0 + \vec{\beta}^T \vec{x} = 0$$

also satisfies

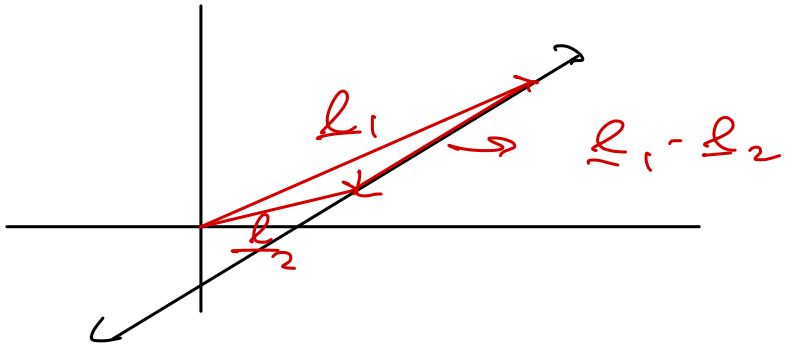
$$k\beta_0 + k\vec{\beta}^T \vec{x} = 0 \quad \text{for } k \neq 0$$

**Claim** $y_i(\beta_0 + \underline{\beta}^T \underline{x}_i)$ is the distance from $\underline{x}_i$ to the hyperplane. Thus, $M$ is the margin from the nearest data points to the plane.

Argument in 2 steps:

(1) Note $\underline{\beta}$ is normal to the hyperplane.

Any vector parallel to the hyperplane can be written as a difference of two vectors on the plane




Need to show

$\underline{\beta}^T(\underline{l}_1 - \underline{l}_2) = 0$

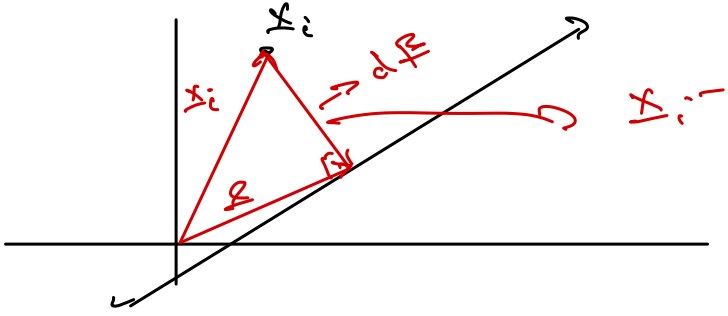
$\beta^T \underline{l}_1 - \beta^T \underline{l}_2$ $\quad$ [ $\underline{l}_1, \underline{l}_2$ on plane $\Rightarrow \beta_0 + \beta^T \underline{l}_1 = 0$ $\quad$ $\beta^T \underline{l}_1 = -\beta_0$ ]

$= -\beta_0 - (-\beta_0) = 0.$

(2) Due to the algorithm, $\|\underline{\beta}\|_2 = 1$ is a unit vector, so the vector from a candidate point $\underline{x}_i$ to the plane can be written

$$\underline{x}_i - \underline{l} = d\underline{\beta}$$

where $d$ is the signed distance to the plane.




$\underline{x}_i - \underline{l}$ = orthogonal to plane & of length $d = d\beta$

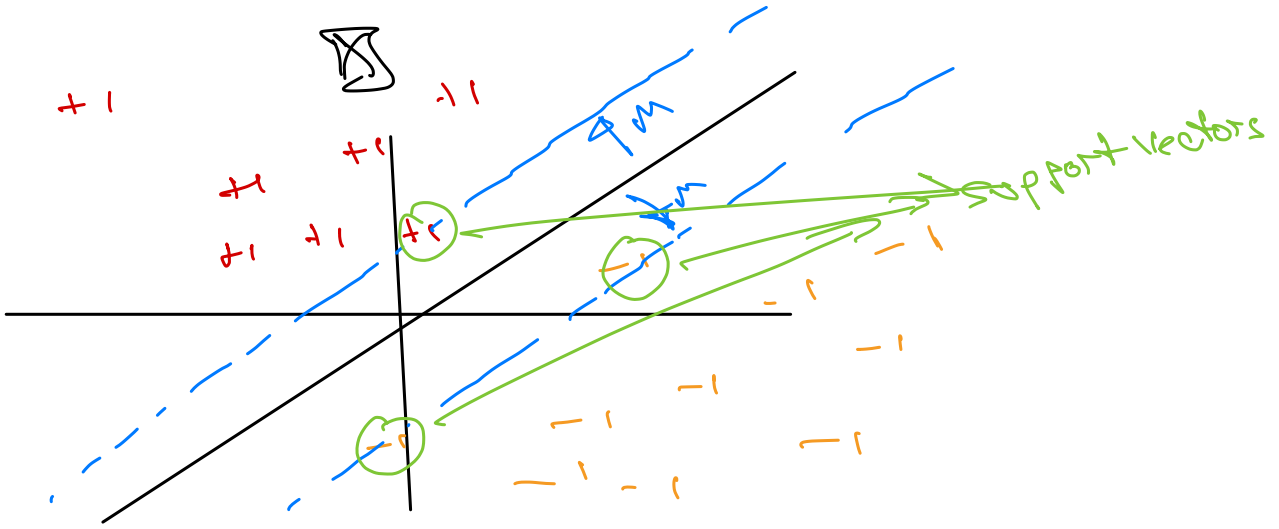
multiply by $\beta^T$:

$$\beta^T x_i - \beta^T \ell = d\, \beta^T \beta \qquad \left| \quad \beta^T \beta = \|\beta\|_2^2 = 1 \right.$$

$$\Leftrightarrow \quad \beta^T x_i - (-\beta_0) = d$$

$$\Leftrightarrow \quad \beta_0 + \beta^T x_i = d$$

$$\Leftrightarrow \quad y_i(\beta_0 + \beta^T x_i) = y_i d = \text{unsigned distance.}$$



Algorithm

- Maximize $M$
  $\beta_0, \underline{\beta}$
- subject to $\|\underline{\beta}\|_2 = 1$ & $y_i(\beta_0 + \underline{\beta}^T \underline{x}_i) \geq M \quad \forall i$

---

Try to remove $\|\underline{\beta}\|_2 = 1$ condition

Let $\underline{\beta}_d$ be parallel to $\underline{\beta}$ but of length $d$, so $\underline{\beta}_d = d\underline{\beta}$

$$y_i(\beta_0 + \underline{\beta}^T \underline{x}_i) \geq M$$

$$y_i(\beta_0 + \underline{\beta}_d^T \underline{x}_i / d) \geq M$$

$$y_i(d\beta_0 + \underline{\beta}_d^T \underline{x}_i) \geq dM$$

The plane defined by $\beta_0$ & $\underline{\beta}$ is the same as that defined by $d\beta_0 = \gamma_0$ & $\underline{\beta}_d$, choose $d = \frac{1}{M}$

The algorithm can be rewritten as

$$\min_{\beta_0, \underline{\beta}} \frac{1}{2}\|\underline{\beta}\|_2^2$$

subject to $y_i(\beta_0 + \underline{\beta}^T \underline{x}_i) \geq 1 \quad \forall i$

Can convert this to a Lagrangian formulation

[go back $\beta_0 + \underline{\beta}$ notation]

$$\min_{\beta_0, \underline{\beta}} \frac{1}{2}\|\underline{\beta}\|_2^2 - \sum_{j=1}^{n} \alpha_j (y_j(\beta_0 + \underline{\beta}^T \underline{x}_j) - 1)$$

where $\alpha_1, \ldots, \alpha_n$ are Lagrange multipliers

$$\frac{d}{d\underline{\beta}} \Rightarrow \underline{\beta} - \sum_{i=1}^{n} y_i \alpha_i \underline{x}_i \overset{\text{set}}{=} 0$$

$$\Rightarrow \quad \underline{\beta} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \underline{x}_i$$

$$\frac{\partial}{\partial \beta_0} \Big| \Rightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

Plug back in

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \underline{x}_i^T \underline{x}_j$$

maximize over $\underline{\alpha}$ subject to $\alpha_i \geq 0$.

At solution it turns out that:

- $\alpha_i = 0$ happens if $y_i(\beta_0 + \underline{\beta}^T \underline{x}_i) > 1$
- $\alpha_i > 0$ " " $y_i(\beta_0 + \underline{\beta}^T \underline{x}_i) = 1$

↳ these points are the <u>support vectors</u>, it is only these points that define the hyperplane.

The classification function is still

$$f(x) = \beta_0 + \beta^T x = \beta_0 + \sum_{i=1}^{n} \alpha_i y_i \, x_i^T x$$

only nonzero terms are support vectors

& classification rule is still

$$\hat{y} = \begin{cases} +1 & f(x) > 0 \\ -1 & f(x) < 0 \end{cases}$$