Separating hyperplanes

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p = 0$$

- For a given observation that satisfy the above equation, the observation lies on the hyperplane
- Now, if the LHS > 0, then the observation lies to one side of the hyperplane
- If LHS < 0, then the observation lies on the other side of the hyperplane
- We can therefore use the hyperplane to classified the observation

Imagine data with 2 classes ; -1 , +1

$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} > 0$, if $y_i = +1$

$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} > 0$, if $y_i = -1$

Both equation above is equivalent to :

$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}) > 0$

The maximal margin classifier

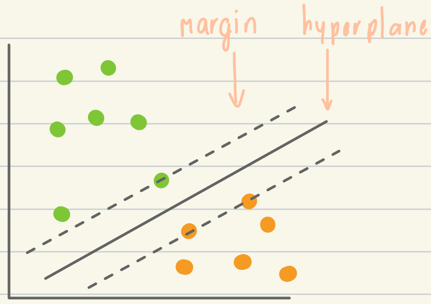
Optimization :

$$\underset{\beta_0, \beta_1, \dots, \beta_p, M}{\text{maximize}} \quad M$$

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

$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}) \geq M \quad \forall i = 1, \dots, n$

- when data is perfectly separated by hyperplane
- largest margin



The soft margin classifier

Optimization :

$$\underset{\beta_0, \beta_1, \dots \beta_p, \epsilon_1, \dots, \epsilon_n, M}{\text{maximize}} \quad M$$

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

$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}) \geq M(1 - \epsilon_i)$,

$\epsilon_i \geq 0$, $\sum_{i=1}^{n} \epsilon_i \leq C$

- C is a nonnegative tuning parameter
- ϵ_i tells us where the observation ith located, relative to hyperplane and margin
  - $\epsilon_i = 0$, correct side of the margin
  - $\epsilon_i > 0$, wrong side of the margin
  - $\epsilon_i > 1$, wrong side of the hyperplane
- this mean that if $C = 0$, then it is a maximal margin classifier

  if $C > 0$, then, no more than C observation can be on the wrong side of the hyperplane

- C controls the bias-variance trade-off
  when C is small, narrow margin, highly fit to the data, ↓ bias ↑ variance
  when C is large, wider margin, fit to the data less hard, ↑ bias ↓ variance
- only the observations that $\underbrace{\text{lie on the margin or violate the margin}}$ will affect the hyperplane
  ↳ support vectors

Support Vector Machine

- we may want to enlarge our feature space to accommodate a non-linear boundary
- basically, soft margin classifier with kernel trick
- using kernel approach is simply for computational efficiency

$f(x) = \beta_0 + \sum_{i \in S} a_i K(x, x_i)$, where number of parameter α equal to number of support point

Common kernel function :
- polynomial : $(1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d$
- radial : $\exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2)$