FYS-STK4155, Nov 11, 2022

Gradient boosting

$$f_M(x) = \sum_{i=1}^{M} \beta_i b_i(x; \gamma_i)$$

$$f_m(x) = f_{m-1}(x) + \beta_m b_m(x; \gamma_m)$$

$$f_m(x) = f_{m-1}(x) + \gamma_m r_m(x)$$

$$\tilde{y}_i = f(x_i) = f_M(x_i)$$

Regression as example

$$C(f) = \frac{1}{n}\sum_{i=0}^{n-1}\left(y_i - f(x_i)\right)^2$$

$$\tilde{y}_i = f_m(x_i) = f_{m-1}(x_i) + r_{im}\gamma_m$$

$$r_{im} = y_i - f_m(x_i)$$

$$-\frac{\partial C(f)}{\partial f(x_i)} = \frac{2}{n}\left(y_i - f(x_i)\right)$$

$$= \frac{2}{n} r_{im}$$

Algorithm for gradient boost

Define $D = \{ (x_0, y_0), (x_1, y_1), \ldots (x_{n-1}, y_{n-1}) \}$

Define M

Define differentiable cost function $C(f)$

$$C(f) = \sum_{i=0}^{n-1} \mathcal{L}(y_i, f(x_i))$$

initialize $f_0(x)$ by optimizing $f_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=0}^{n-1} \mathcal{L}(y_i, f(x_i))$

for $m = 1 : M$

(1) compute

$$r_{im} = - \frac{\partial \mathcal{L}(y_i, f(x_i))}{\partial f(x_i)}$$

at $f(x_i) = f_{m-1}(x_i)$

for all $i = 0, 1, \ldots n-1$

(ii) Fit a base (weak) learner (= Tree) using on training set for all $i = 0, 1, \dots n-1$

(iii) compute the multiplier γ_m by optimizing

$$\hat{\gamma}_m = \underset{\gamma}{\arg\min} \sum_{i=0}^{n-1} \mathcal{L}(y_i, f_{m-1}(x_i) + \gamma h_m(x_i))$$

(iv) update

$$f_m(x) = f_{m-1}(x) + \gamma_m h_m(x)$$

end

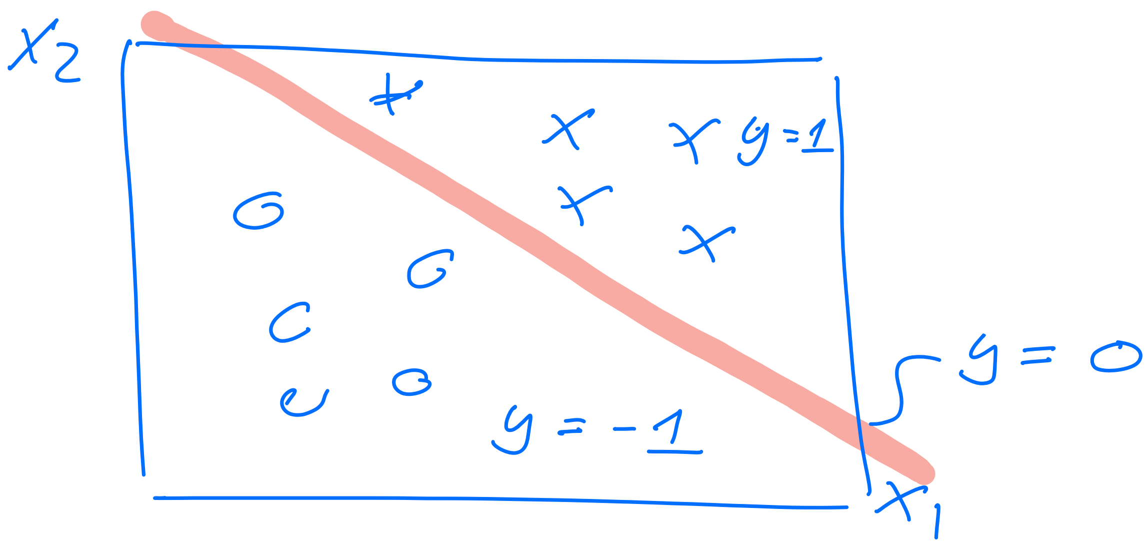
Return $f_M(x)$

— Decision trees
Binary split, CART
For classification we use gini factor or

the entropy.

- Ensemble (weak learner)
  - Bootstrap aggregate = Bagging (Homogeneous)
  - Random forests (Heterogeneous)
  - Adaptive boosting
  - ~~Gradient boosting~~

Support Vector Machines



$y_i \in \{-1, +1\}$

$X^T = [x_1 \ x_2]$

w = weight vector

w_0 = bias (intercept)

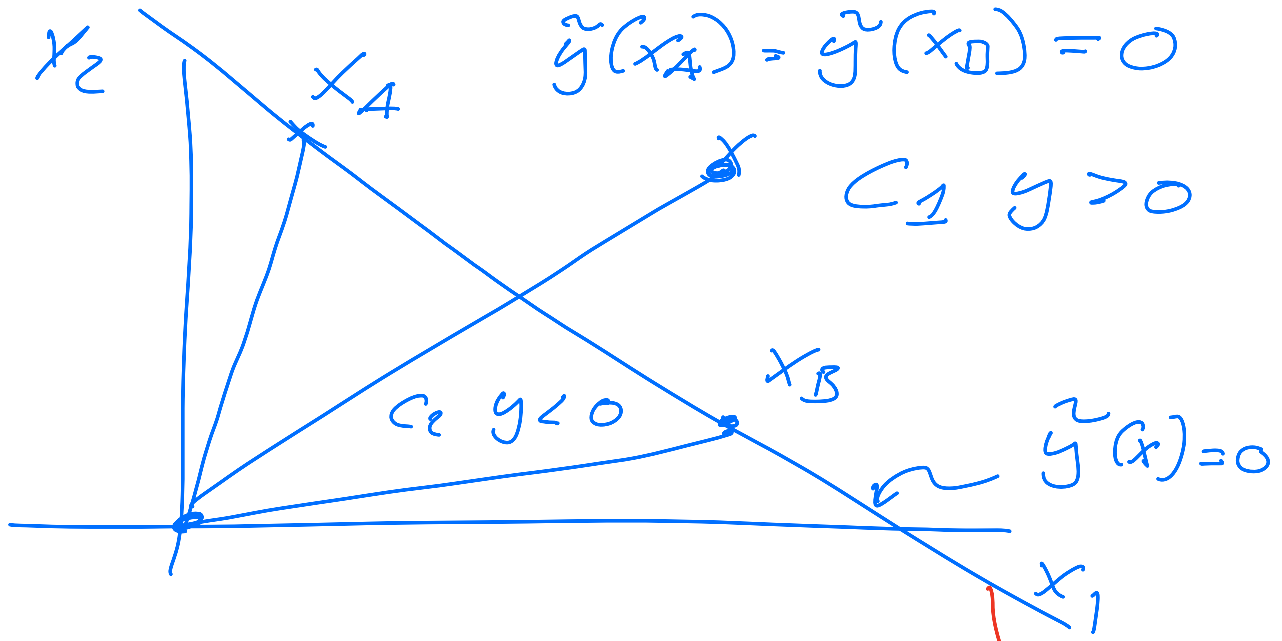
$\tilde{y}(x) = w^T x + w_0$

$= w_1 x_1 + w_2 x_2 + w_0$

if $\tilde{y}(x) > 0$, then the output belongs to c_1 $(y_i = +1)$

else $\tilde{y}(x) < 0$, then c_2 $(y_i = -1)$

$\tilde{y}(x) = 0$ defines the boundary.



$$\tilde{y}(x_A) = \tilde{y}(x_B) = 0$$

$$w^T(x_A - x_B) = 0$$

Decision surface

w is orthogonal to every vector which lies within the decision surface.

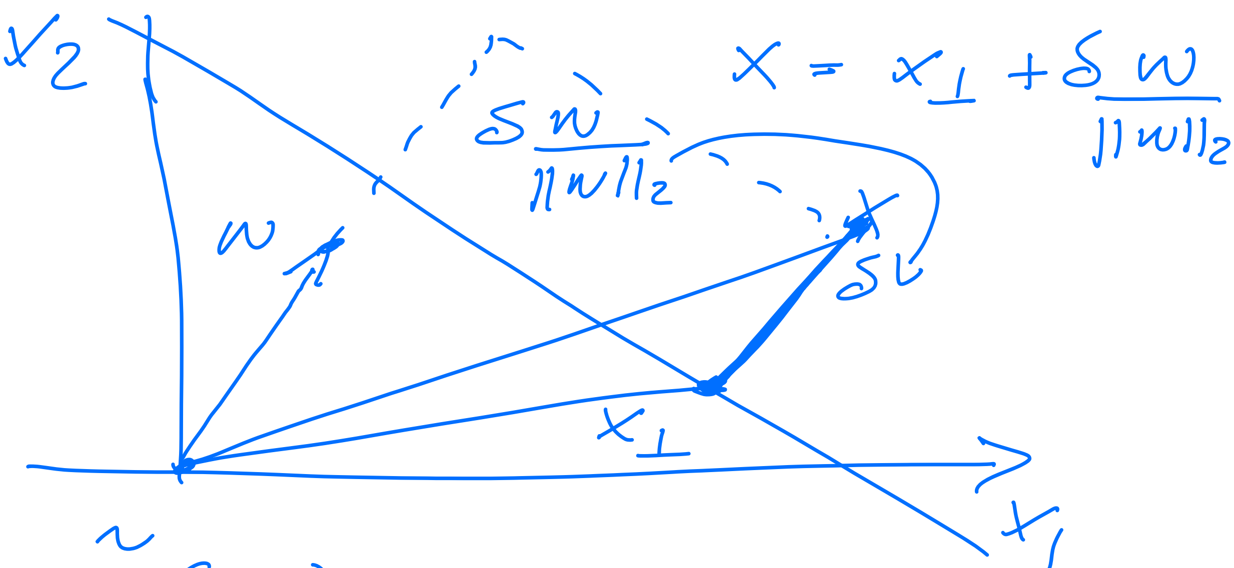
w determines the orientation of the decision surface (line) from a chosen reference.

$\tilde{y}(x) = 0 \qquad w^T x + w_0$

we normalize w by

$$\sqrt{w^T w} = \|w\|_2$$

$$\frac{w^T x}{\|w\|_2} = -\frac{w_0}{\|w\|_2}$$



$$\tilde{y}(x_\perp) = 0$$

$$w^T x = w^T x_\perp + \delta \frac{w^T w}{\|w\|_2}$$

add w_0 to both sides

$$\tilde{y}(x) = w^T x + w_0$$

$$= w^T \underline{x_1} + w_0 + \delta \|w\|$$

$$\tilde{y}(\underline{x_1}) = 0 = w^T \underline{x_1} + w_0$$

$$-\frac{\tilde{y}(x)}{\|w\|} = \delta$$

<u>intermediate step</u>

could we define a cost function which contains all misclassification points (set M) and we want to minimize it.

$$C(w, w_0) = -\sum_{i \in M} y_i (w^T x_i + w_0)$$

if $y_i = 1$, misclassified if $w^T x_i + w_0 < 0$

opposite if $y_i = -1$

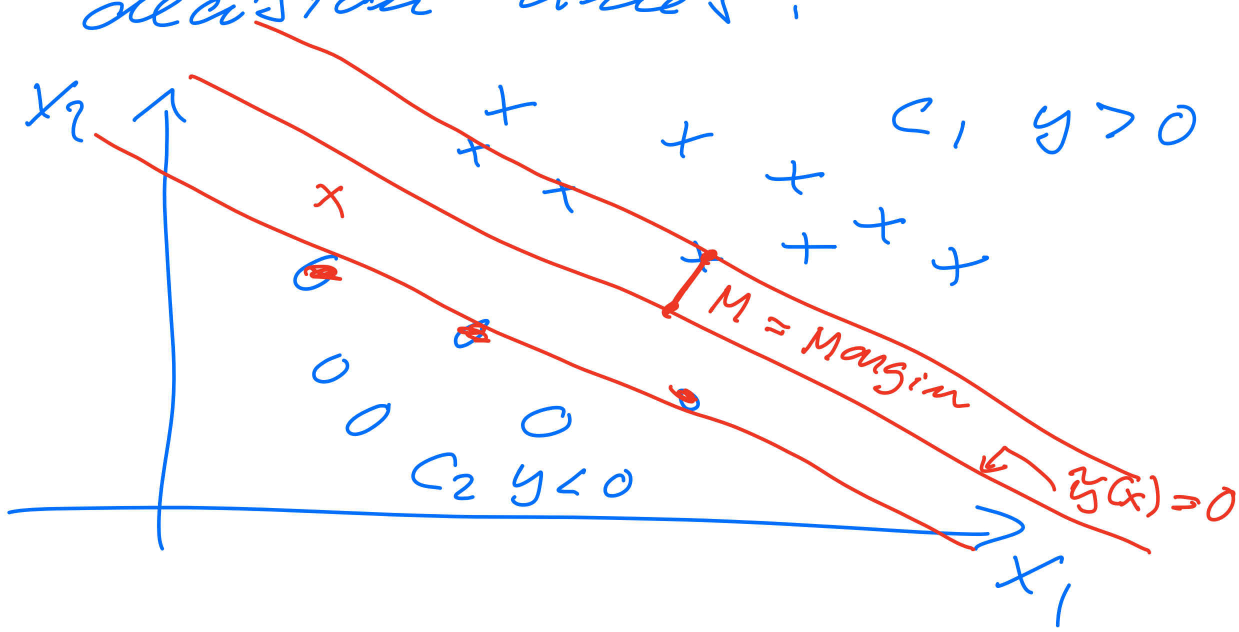
$$\frac{\partial C}{\partial w_0} = 0 = -\sum_{i \in M} y_i$$

$$\frac{\partial C}{\partial w} = 0 = -\sum_{i \in M} y_i x_i$$

$$w_0 \leftarrow w_0 - \eta \frac{\partial C}{\partial w_0}$$

$$w \leftarrow w - \eta \frac{\partial C}{\partial w}$$

Can lead to different decision lines.



$$\delta = \frac{f(x)}{\|w\|_2}$$

we seek a margin M defined by

$$w^T(x - x_c) = \frac{1}{\|w\|_2}(w^T x + w_0)$$

$$= \frac{f(x)}{\|w\|_2}$$

$$y_i \frac{1}{\|w\|_2}(w^T x_i + w_0) \geq M$$

$$(y_i \tilde{y}_i \geq 0)$$

$$y_i (w^T x_i + w_0) \geq M \|w\|_2$$

$$M \|w\|_2 = 1 \quad \Rightarrow$$

$$M = \frac{1}{\|w\|_2}$$

$$\min_{w, w_0} \|w\|_2$$

with the constraint

$$y_i (w^T x_i + w_0) \geq 1$$