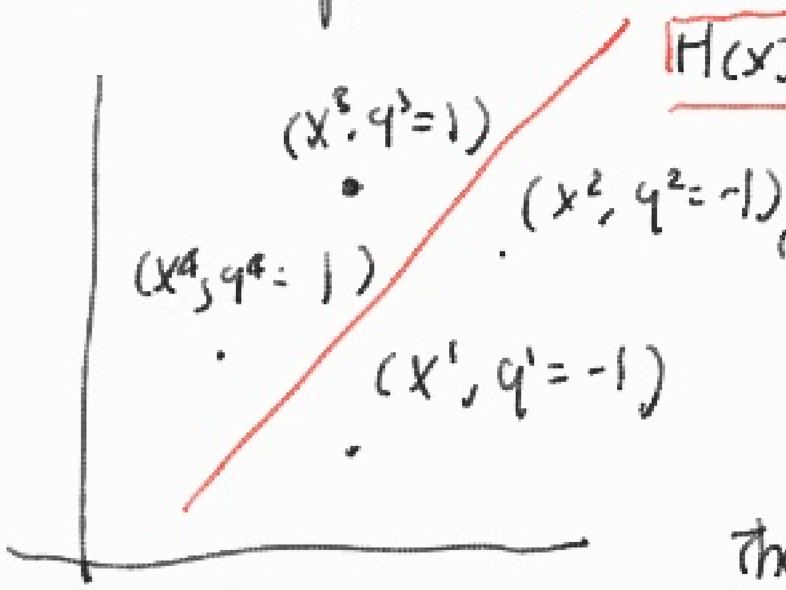
Support Vector Machine.

- given instance vector $X \in \mathbb{R}^n$, $y \in \{-1, 1\}$, there exists a decision boundary $H = \Theta^T X + b$ $\left(\begin{array}{l}\text{where } \Theta \in \mathbb{R}^n \\ b \text{ is bias}\end{array}\right)$ that seperates the domain where $y=-1$ and the domain where $y=1$.



$\boxed{H(x) = \Theta^T x + b}$

- $\Theta = [\theta_1\ \theta_2\ \theta_3 \cdots \theta_n]$
- $X = [x_1\ x_2\ x_3 \cdots x_n]$
- $H(x) = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n + b$

then the distance $d_H(X^{(i)})$ is the distance between a certain datapoint $(X^{(i)}, y^{(i)})$ and $H(x)$

$$d_H(x^{(i)}) = \frac{|\Theta^T X^{(i)} + b|}{\|\Theta\|_2}.$$

The objective is to find an optimal Θ that results in a $H(x)$ with largest margin $(d_H(x^{(i)}))$ to the closest datapoint $(X^{(i)}, y^{(i)})$

- $\underset{\Theta}{\operatorname{argmax}} \left(\min_i \left(\frac{|\Theta^T X^{(i)} + b|}{\|\Theta\|_2} \right) \right)$

- $\underset{\Theta}{\arg\max}\left(\min_i\left(\frac{|\Theta^T x^{(i)}+b|}{\|\Theta\|_2}\right)\right)$

we can rewrite

$|\Theta^T x^{(i)}+b| \Longleftrightarrow y^{(i)}(\Theta^T x^{(i)}+b) \begin{cases} \geq 0 & (y \text{ and } H(x) \text{ have the same sign} \Leftrightarrow \text{correctly classified}) \\ < 0 & (\text{else}) \end{cases}$

$$\underset{\Theta}{\arg\max}\left(\min_i\left(\frac{y^{(i)}(\Theta^T x^{(i)}+b)}{\|\Theta\|_2}\right)\right) = \underset{\Theta}{\arg\max}\left(\frac{1}{\|\Theta\|_2}\min_i\left(y^{(i)}(\Theta^T x^{(i)}+b)\right)\right)$$

this results in the following optimization problem

objective : $\underset{\Theta}{\arg\max}\ \frac{1}{\|\Theta\|_2}$

condition : $\min_i\ y^{(i)}(\Theta^T x^{(i)}+b)$

$y^{(i)}(\Theta^T x^{(i)}+b)\begin{cases}\geq 1 \\ < 1\end{cases} \leftarrow$ normalize the minimum margin into 1 by multiplying some constant

then

obj : $\underset{\Theta}{\arg\max}\ \frac{1}{\|\Theta\|_2}$

condition : $\min_i\ y^{(i)}(\Theta^T x^{(i)}+b) = 1,\ \forall i$

this indicates that our boundary H can correctly classify EVERY data points $(x^{(i)}, y^{(i)})$ (Hard Margin SVM)

convert the problem into minimization problem
thus constructing the final form of
Hard Margin SVM

obj : $\min_{\Theta} \frac{1}{2} \frac{1}{\|\Theta\|_2}$

condition : $y^{(i)}(\Theta^T x^{(i)} + b) \geq 1, \forall i$

But in case where it's impossible to correctly classify every data points, (like most cases) we allow errors.

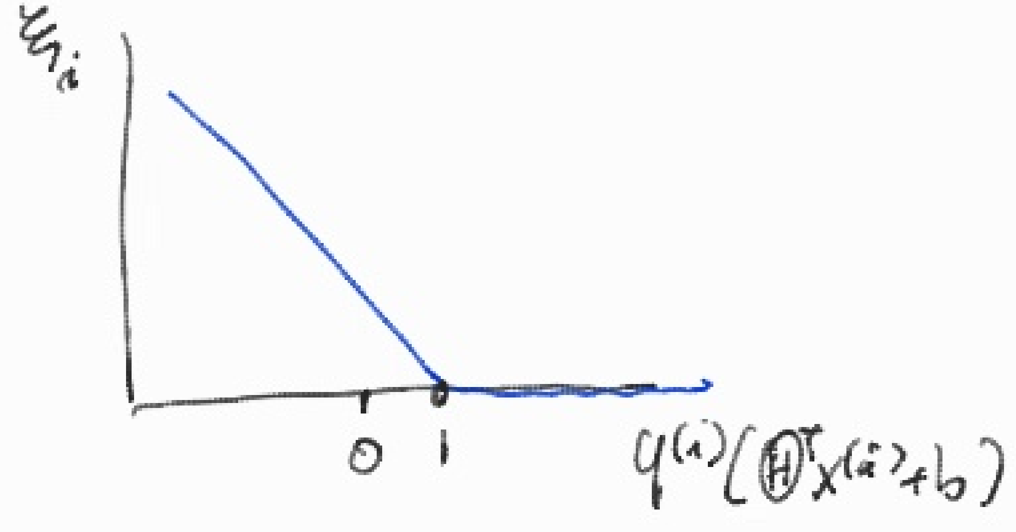
$$y^{(i)}[\Theta^T x^{(i)} + b] \leq 0, \ \exists i$$ for some i

and introduce an error measurement / slack variable

$$\xi_i = \max(0, 1 - y^{(i)}(\Theta^T x^{(i)} + b)).$$

$$\xi_i \begin{cases} 0 & (y^{(i)}(\Theta^T x^{(i)} + b) \geq 1, \text{ the data point } (x^{(i)}, y^{(i)}) \text{ is correctly classified}) \\ > 1 & (\text{else}) \end{cases}$$

* the slack variable ξ_i gets larger as the incorrect margin increases.



Then we can rewrite our problem.

obj: $\min_{\Theta, \xi} \left(\frac{1}{2} \|\Theta\|_2 + \frac{\lambda}{m} \sum^{m} \xi_i \right)$

where λ is regalarization hyper parameter

m is # of instances.

* when λ is small enough, the slack variable loses influence. Else it's the other way around

condition: $y^{(i)}(\Theta^T x^{(i)} + b) \geq 1 - \xi_i, \forall i$

$\xi_i \geq 0, \forall i$

* since ξ_i is defined to be $\max(0, 1 - y^{(i)}(\Theta^T x^{(i)} + b))$

Hence, the final form of the SVM

obj: $\min_{\Theta, \xi} \frac{1}{\|\Theta\|_2} + \frac{\lambda}{m} \sum^{m} \xi_i$

condition: $y^{(i)}(\Theta^T x^{(i)} + b) \geq 1 - \xi_i, \forall i$

$\xi_i \geq 0, \forall i$

In order to solve this problem,
We need to transform our problem into a convex optimization problem.

obj: $\min_{\Theta, b, \xi} \frac{1}{\|\Theta\|_2} + \frac{\lambda}{m}\sum^{m} \xi_i$

condition: $(1 - y^{(i)}(\Theta^T x^{(i)} + b)) - \xi_i \leq 0, \ \forall i$

$-\xi_i \leq 0, \ \forall i$

Once we get a convex op. form, we then proceed to determine Lagrangian Dual Form

$L(\Theta, b, \xi, \alpha, \beta)$

$= \frac{1}{2}\|\Theta\|_2 + \frac{\lambda}{m}\sum \xi_i + \sum \alpha[1 - y^{(i)}(\Theta^T x^{(i)} + b) - \xi_i] \sum \beta_i(-\xi_i)$

$= \frac{1}{2}\Theta^T\Theta + \sum \xi_i(\frac{\lambda}{m} - \alpha_i - \beta_i) + \sum \alpha_i(1 - y^{(i)}(\Theta^T x^{(i)} + b))$

* since both conditions are affine,
the problem has strong duality with an appropriate Θ, b, ξ

then

$g(\alpha, \beta) = \inf_{\Theta, b, \xi} L(\Theta, b, \xi, \alpha, \beta)$

calculate partial derivative of L

$\partial_{\Theta} L = 0 \iff \omega - \sum \alpha_i y^{(i)} x^{(i)} = 0 \quad \therefore \Theta = \sum \alpha_i y^{(i)} x^{(i)}$

$\partial_{b} L = 0 \iff -\sum \alpha_i y^{(i)} = 0 \quad \therefore \sum \alpha_i y^{(i)} = 0$

$\partial_{\xi} L = 0 \iff \frac{\lambda}{m} - \alpha_i - \beta_i = 0 \quad \therefore \alpha_i + \beta_i = \frac{\lambda}{m}$

substitute each component with the above result

in $g(\alpha, \beta)$

then the final form of our SUM would be

obj : $\max_{\alpha} \sum \alpha_i - \frac{1}{2} \sum_{i\,j}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)T} x^{(j)}$

cond : $\sum \alpha_i y^{(i)} = 0 \,, \quad \alpha \in \left[0, \frac{\lambda}{m}\right]$) → regularization

the corresponding solution would be

1. solve the above problem to obtain the optimal $\alpha = (\alpha_1\ \alpha_2 \cdots \alpha_i \cdots \alpha_m)$ / dual solution

2. obtain optimal Θ with $\Theta = \sum \alpha_i y^{(i)} x^{(i)}$