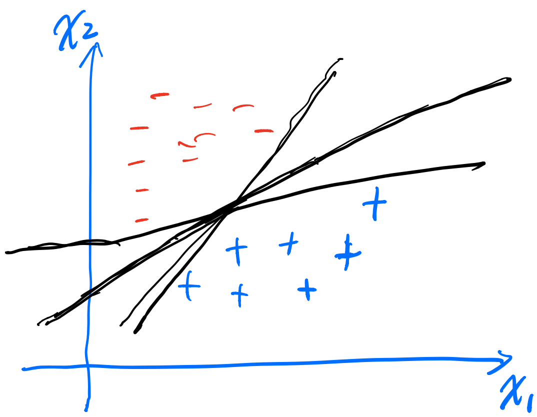
Logistic Loss ✓

Hinge loss.

Used by Support Vector Machine (SVM)

Main idea: Margin Maximization



Linearly separable

Infinite linear models w/ perfect accuracy.

Want to maximize the distance between decision boundary & examples.

$w \in \mathbb{R}^d$: weight vector

Decision boundary : $H = \{x \in \mathbb{R}^d \mid \langle w, x \rangle = 0\}$

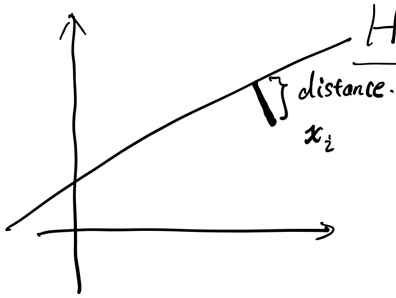
Hyperplane

Suppose w is perfectly accurate

for all examples $(x_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$

$$y_i w^T x_i > 0$$

Distance between x & H.



$H = \{x \mid \langle w, x\rangle = 0\}$ not changed if $w \cdot 2$.

$y_i w^T x_i > 0$.

$$\text{distance} = \underbrace{\frac{y_i w^T x_i}{\|w\|_2}}_{\text{invariant under multiplication of } w.}$$

Idea of margin / distance maximization.

$$\boxed{\max_w \; \underbrace{\min_i \frac{y_i w^T x_i}{\|w\|_2}}_{\text{closest distance}}}$$

optimization does not care about $\|w\|_2$

might as well consider w's

such that $\boxed{\min_i y_i w^T x_i = 1}$

$$\Longleftrightarrow \quad \boxed{\max_w \frac{1}{\|w\|_2}} \quad \text{such that} \quad \min_i y_i w^T x_i = 1. \qquad (2)$$

$\Updownarrow$

$$\Longleftrightarrow \quad \min_w \boxed{\frac{1}{2}\|w\|_2^2} \quad \text{such that} \quad \boxed{y_i(w^T x_i) \geq 1} \;\; \forall i \qquad (3)$$

$\min_i y_i(w^T x_i) = 1$ for all i

w

such that

$\min_i y_i(w^T x_i) = 2 \longrightarrow$ take $\frac{w}{2}$.

$\rightarrow$ margin maximization formulation

"In a perfect world"

What if not linealy separable?

$$\min_w \boxed{\tfrac{1}{2}\|w\|_2^2} \quad \text{such that} \quad \boxed{y_i(w^T x_i) \geq 1} \quad \underbrace{\forall i}_{\text{all}}$$

"Hard" Margin SVM (may not be feasible)

"Soft" Margin SVM.

$$\min_w \tfrac{1}{2}\|w\|_2^2 + C\sum_{i=1}^n \xi_i$$

such that

Constr. $\begin{cases} \forall i, \quad y_i(w^T x_i) \geq 1 - \underbrace{\boxed{\xi_i}}_{\text{slack variable}}, \quad \text{"}x_i\text{"} \\ \forall i, \quad \xi_i \geq 0. \end{cases}$

get rid of?

Equivalent form:

$$\min_w \tfrac{1}{2}\|w\|_2^2 + \boxed{C}\sum_{i=1}^n \underbrace{\max\{0, 1 - y_i w^T x_i\}}_{\text{"Hinge loss"}}$$

for some λ

$$\Longleftrightarrow \quad \min_w \underbrace{\boxed{\lambda}\|w\|_2^2}_{\text{Regularization.}} + \underbrace{\sum_{i=1}^n \max\{0, 1 - y_i w^T x_i\}}_{\text{Hinge loss ERM}}$$

Constrained OPT $\Longrightarrow$ Unconstrained OPT

$$\min_w \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \xi_i$$

such that

$$\forall i, \quad y_i(w^T x_i) \geq 1 - \boxed{\xi_i}.$$

slack variable

$$\forall i, \quad \xi_i \geq 0.$$

$$\min_w \frac{1}{2}\|w\|_2^2 + \boxed{C}\sum_{i=1}^{n} \underbrace{\max\{0, 1 - y_i w^T x_i\}}_{\text{"Hinge loss"}}$$

Detour : Lagrange Duality.

$$\min_w F(w) \quad \text{such that} \quad \underline{h_j(w) \leq 0}, \quad \forall j \in [m]$$

Constrained OPT

$[m] \rightarrow \{1, 2, \dots m\}$

"Put constraints into obj" $\quad w \leftarrow$ "primal ____"

For each j, introduce $\lambda_j \geq 0 \leftarrow$ "dual variable"

Lagrangian : $\quad L(w, \lambda) = F(w) + \sum_{j=1}^{m} \lambda_j h_j(w)$

a game.

primal

control w

$\min L(w,\lambda)$

$\downarrow$

forced to satisfy $h_j(w) \leq 0 \quad \forall j$

Dual.

control λ

$\max L(w,\lambda)$

Suppose w is not a feasible solution

$h_j(w) > 0$

$\lambda_j \leftarrow$ "infinity"

$L(w,\lambda) \rightarrow \infty$

Zero-sum game view

$$\underline{\min_{w} \max_{\lambda}} \; L(w,\lambda) \quad \text{"minmax"}$$

$$\underline{\max_{\lambda} \min_{w}} \; L(w,\lambda) \quad \text{"max min"}$$

left $\longrightarrow$ right

minmax : minimization goes first
maximization "best response"

maxmin : opposite

Play second is "usualy" better.

Weak Duality: $\underbrace{\text{maxmin}}_{\text{favors min player.}} L(w,\lambda) \leq \underbrace{\text{minmax}}_{\text{favors max player.}} L(w,\lambda)$

Strong Duality : $\max_{\lambda} \min_{w} L(w,\lambda) = \min_{w} \max_{\lambda} L(w,\lambda)$

under "mild" condition (e.g., soft SVM, Slaters Condition)

$$\lambda^* = \arg\max_{\lambda} \left(\min_{w} L(w,\lambda) \right)$$

$$w^* = \arg\min_{w} \left(\max_{\lambda} L(w,\lambda) \right)$$

Strong Duality $\Longrightarrow$

$$\begin{aligned}
F(w^*) &= \min_{w} \max_{\lambda} L(w,\lambda) \\
&= \max_{\lambda} \min_{w} L(w,\lambda) && \text{(Strong Duality)} \\
&= \min_{w} L(w,\lambda^*) && \text{(Defn } \lambda^*) \\
&\leq L(w^*,\lambda^*) && \text{(Defn min)} \\
&= F(w^*) + \sum_{j} \lambda_j^* h_j(w^*) && \text{(Defn of } L)
\end{aligned}$$

$$\underbrace{\sum_j \lambda_j^* h_j(w^*)} \geq 0$$

$$\min_w F(w) \quad \text{such that } \boxed{h_j(w) \leq 0}, \quad \forall j \in [m]$$

w^* is feasible

$\lambda_j^* \geq 0$

$$\sum_j \lambda_j^* h_j(w^*) \leq 0$$

$$\sum_j \lambda_j^* h_j(w^*) = 0$$

Each $\lambda_j^* h_j(w^*) \leq 0$

$\Rightarrow \forall j, \ \lambda_j^* h_j(w^*) = 0$

$\boxed{\text{KKT conditions}}$

(Characterization of w^*, λ^*)

- $\forall j, \ \lambda_j^* h_j(w^*) = 0$ (complementary slackness)
- w^* is the minimizer of $L(w, \lambda^*)$
  $\nabla_w L(w^*, \lambda^*) \longrightarrow 0$ (Stationarity)
- $\lambda_j \geq 0, \ h_j(w^*) \leq 0$ for all j (Feasibility)

Dated back Kuhn-Tucker '1951

Karush '1939 in unpublished master thesis