Support vector machine

Xiangli Chen

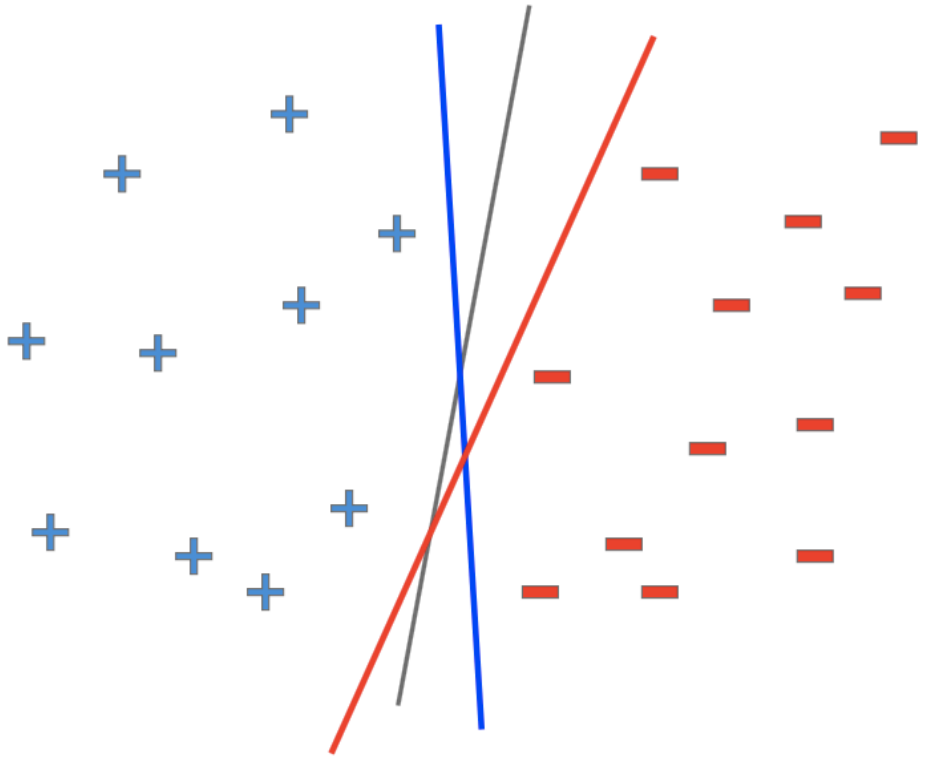
A Binary Linear Classifier

A binary classifier

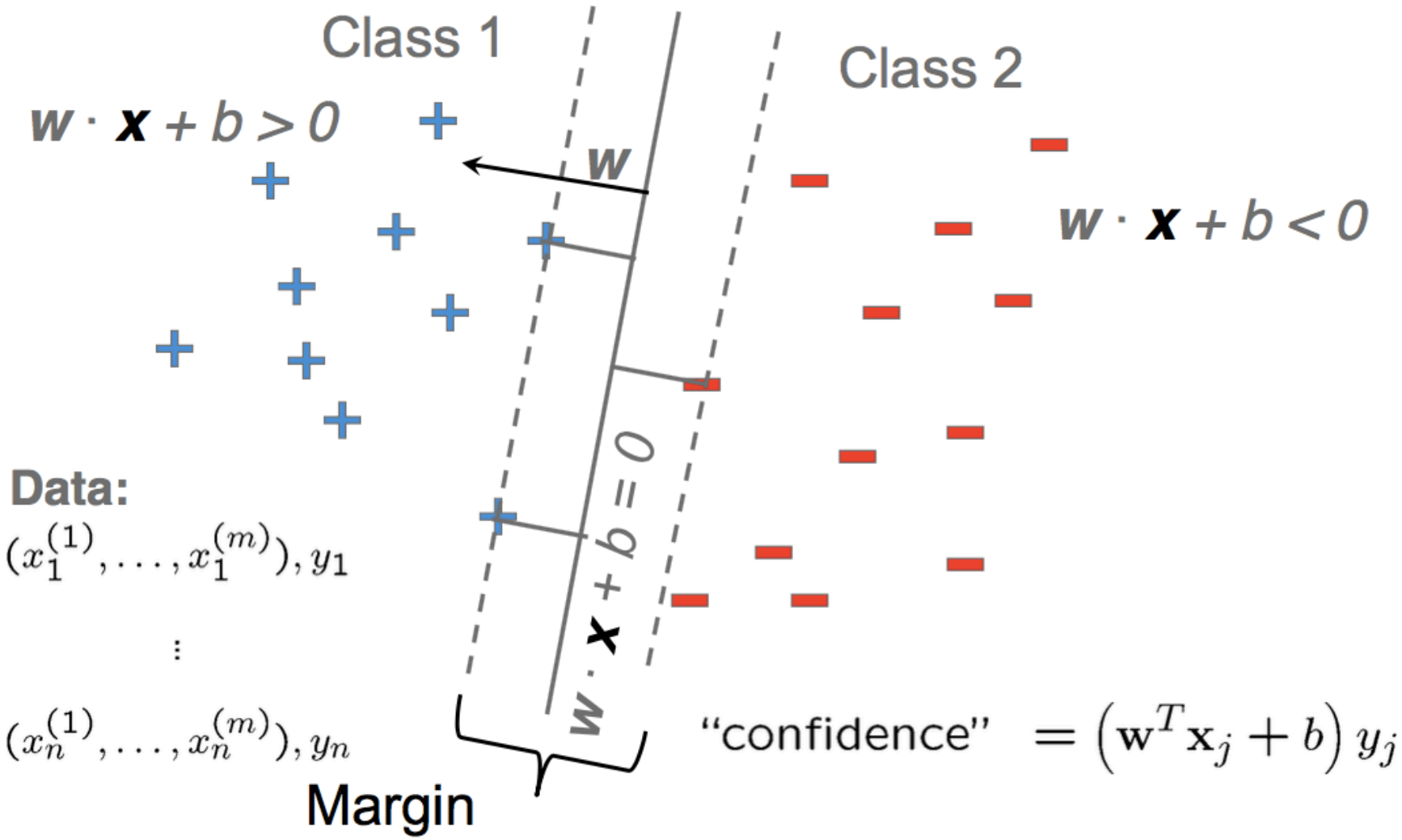
$f(x): X \to Y \quad x = \mathbb{R}^m \quad y \in \{-1, 1\}$

Restrict the class of $f(x)$ to a linear classifier or hyperplane

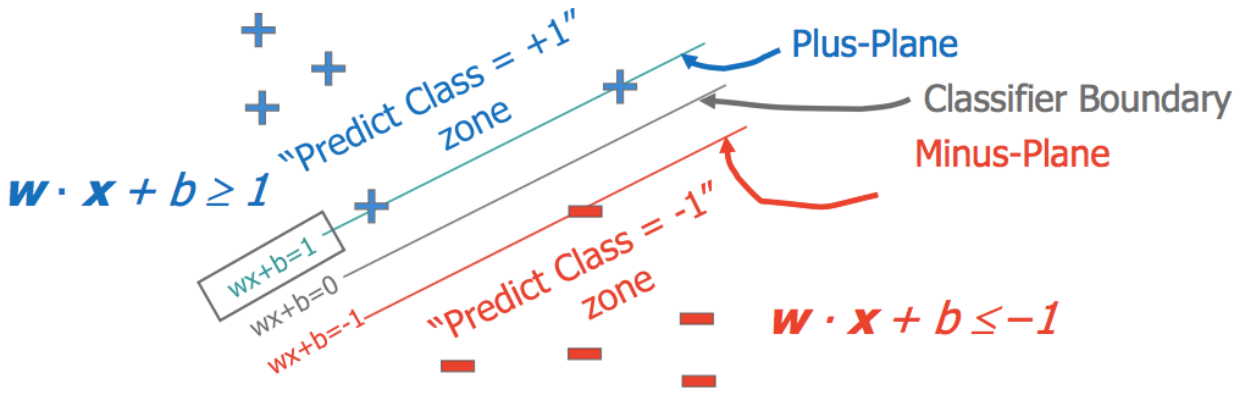
$$f(x) = \begin{cases} +1 & wx + b > 0 \\ -1 & wx + b < 0 \end{cases}$$



Largest Margin



Scaling



Classification rule:

Classify as..	+1	if	$\boldsymbol{w} \cdot \boldsymbol{x} + b \geq 1$
	−1	if	$\boldsymbol{w} \cdot \boldsymbol{x} + b \leq -1$
	Universe explodes	if	$-1 < \boldsymbol{w} \cdot \boldsymbol{x} + b < 1$

Scaling (w, b), the hyperplane maximizing the margin still exists

Hyper Plane and its Normal

Consider a hyper plane

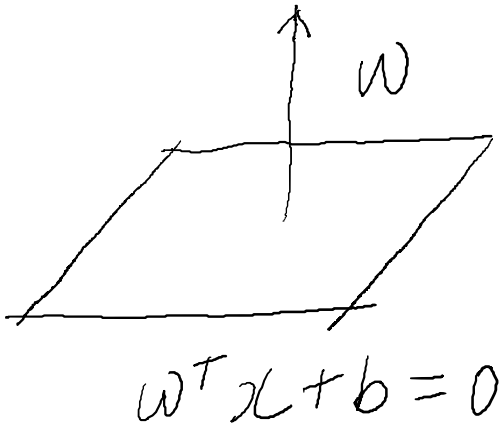
$$w^T x + b = 0$$

Let $w^T x^0 + b = 0$ for a particular x^0

$$w_1(x_1 - x_1^0) + w_2(x_2 - x_2^0) + \cdots + w_m(x_m - x_m^0) = 0$$

$$w^T(x - x_0) = \|w\|\|x - x_0\|\cos\theta = 0 \Rightarrow \theta = 90^\circ$$

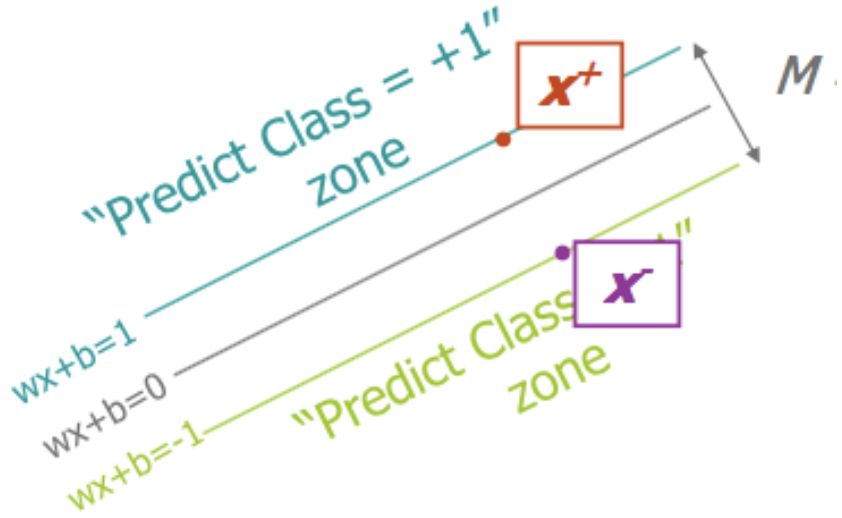
w is orthogonal to the hyperplane denoted by $w^T x + b = 0$



Computing the Margin

Let x^+ and x^- be such that

- $w \cdot x^+ + b = +1$
- $w \cdot x^- + b = -1$
- $x^+ = x^- + \lambda w$
- $|x^+ - x^-| = M$



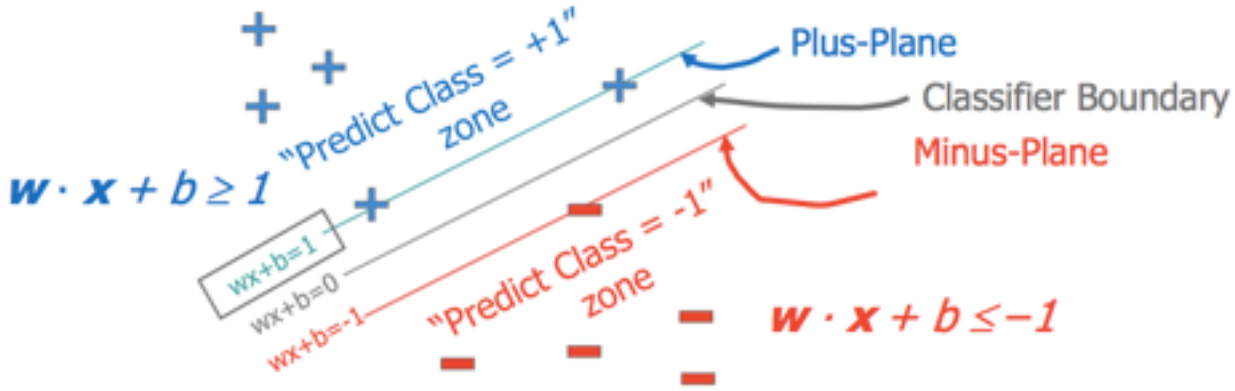
$w(x^+ - x^-) = 2 \Rightarrow \|w\|_2 \|x^+ - x^-\|_2 \cdot \cos\theta = 2$

Since $\cos\theta = 1$

$\|x^+ - x^-\|_2 = \frac{2}{\sqrt{w \cdot w}} = \frac{2}{\|w\|_2}$

So maximizing margin is minimizing $\|w\|_2^2$

Maximize Margin



The distance of any $x_0 \in \mathbb{R}^n$ to the hyperplane $wx+b=0$ is

$$\frac{|w^T x_0 + b|}{\|w\|_2}$$

If $w^T x_0 + b = \pm 1$ (points lie on the marginal hyperplanes) the distance is $\frac{1}{\|w\|_2}$

So maximizing margin is minimizing $\|w\|_2^2$

The Primal Hard SVM

The primal hard SVM

- Giving training data set $D = (x_i, y_i)_{i=1}^{n}$
- Assume that D is linear separable

$$\hat{w}, \hat{b} = \arg\min_{w,b} \frac{1}{2}\|w\|_2^2$$

subject to $y_i(w^T x_i + b) \geq 1 \quad \forall i = 1, \ldots, n$

Solve it as a quadratic programming problem

Lagrange Duality

The Lagrange form

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{n} \alpha_i \left(y_i(w^T x_i + b) - 1\right) \quad \alpha_i \geq 0 \text{ Lagrangian multipliers}$$

The primal problem given by Lagrange form

$$\min_{w,b} \max_{\alpha_i \geq 0} \mathcal{L}(w,b,\alpha)$$

The dual form

$$\max_{\alpha_i \geq 0} \min_{w,b} \mathcal{L}(w,b,\alpha)$$

Theorem (weak duality) $d^* = \max\min \mathcal{L} \leq \min\max \mathcal{L} = p^*$

Theorem (string duality) $d^* = p^*$

For convex optimization problem, such as SVM (if a saddle point exits), strong duality usually exists

KKT Optimality Conditions

For any optimization problem

- Differentiable objective and constraint functions
- Strong duality obtains

Any pair of primal and dual optimal points must satisfy KKT condition

KKT condition of SVM

$$\alpha_i \geq 0$$

$$\nabla_w L = \hat{w} - \sum_{i=1}^{n} \hat{\alpha}_i y_i x_i = 0 \Rightarrow \hat{w} = \sum_{i=1}^{n} \hat{\alpha}_i y_i x_i$$

$$\nabla_b L = -\sum_{i=1}^{n} y_i = 0 \Rightarrow \sum_{i=1}^{n} \hat{\alpha}_i y_i = 0$$

$$\forall i, \hat{\alpha}_i (y_i (\hat{w}^T x_i + b) - 1) = 0 \Rightarrow \hat{\alpha}_i = 0 \text{ or } y_i (\hat{w}^T x_i + b) = 1$$

The Dual Hard SVM

The dual problem

$$\hat{\alpha} = \arg\max_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

$$= \arg\max \alpha^T 1_n - \frac{1}{2} \alpha^T y G y \alpha$$

where $y = \mathrm{diag}(y_1, \cdots, y_n)$, $y \in \{-1, 1\}$ $G = \{G_{ij}\}_{ij}$ where $G_{ij} = \langle x_i, x_j \rangle$

subject to $\alpha_i \geq 0$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$, $\forall i = 1, \cdots, n$

$\langle x_i, x_j \rangle$ → kernel trick $K(\langle x_i, x_j \rangle)$

Solve as a quadratic programming problem

Why dual form

- More efficient QP algorithm especially in high dimension $m \gg n$
- Kernel trick

Support Vector

Support vectors are the points where
$\hat{\alpha}_i \neq 0$ ($\hat{\alpha}_i > 0$) $i = 1, \cdots, n$

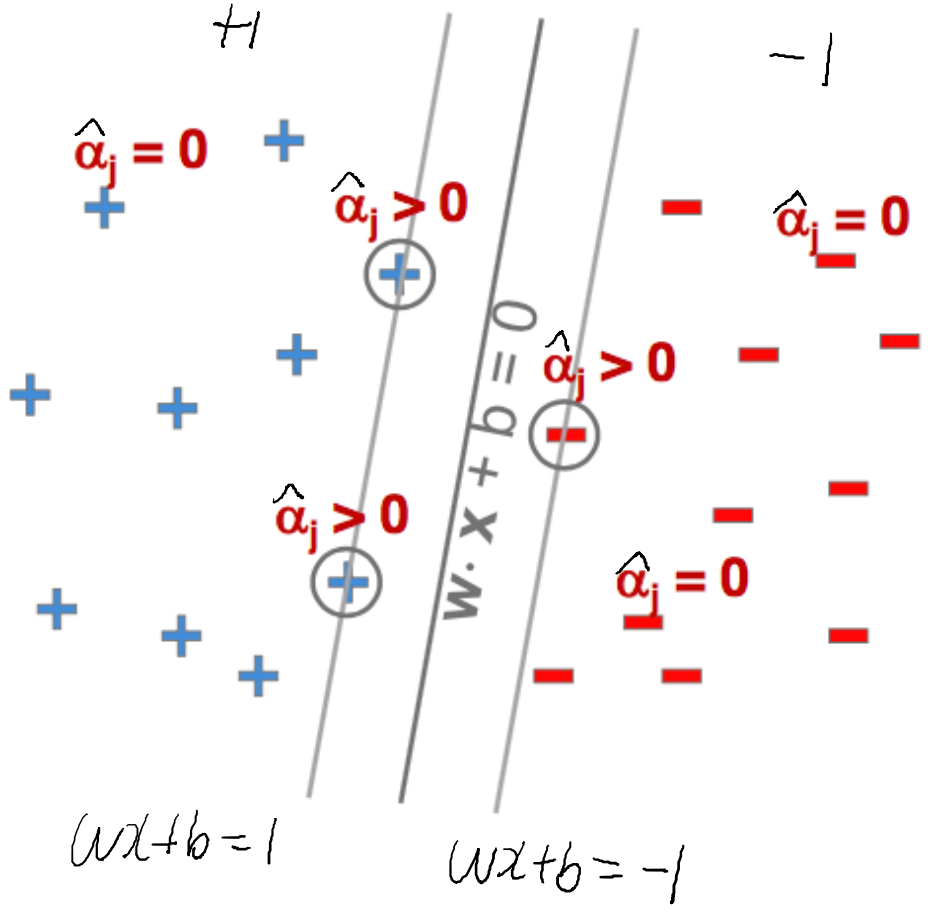
KKT condition shows that

when $\hat{\alpha}_i \neq 0$

$y_i(\hat{w} \cdot x_i + b) = 1$

Support vectors are the points $i \in SV$

lie on the margin hyperplanes



SVM Predictor

Direction

$$\hat{w} = \sum_{i \in SV} \hat{\alpha}_i y_i x_i$$

Bias estimator

For any $\hat{\alpha}_i > 0$ i.e. $i \in SV$ $y_i(\hat{w}^T x_i + \hat{b}) = 1$

Note that $\|y_i\|_2^2 = 1$, $\hat{b} = y_i - \hat{w}^T x_i = y_i - \sum_{j \in SV} \hat{\alpha}_j y_i x_j^T x_i$

Taking average to have a numerical more stable solution

$$\hat{b} = \frac{1}{N_{SV}} \sum_{i \in SV} \left(y_i - \sum_{j \in SV} \hat{\alpha}_j y_i x_j^T x_i \right)$$

For taking a new data x

$$\hat{w}^T x + \hat{b} = \sum_{i \in SV} \alpha_i y_i x_i^T x + \hat{b}$$

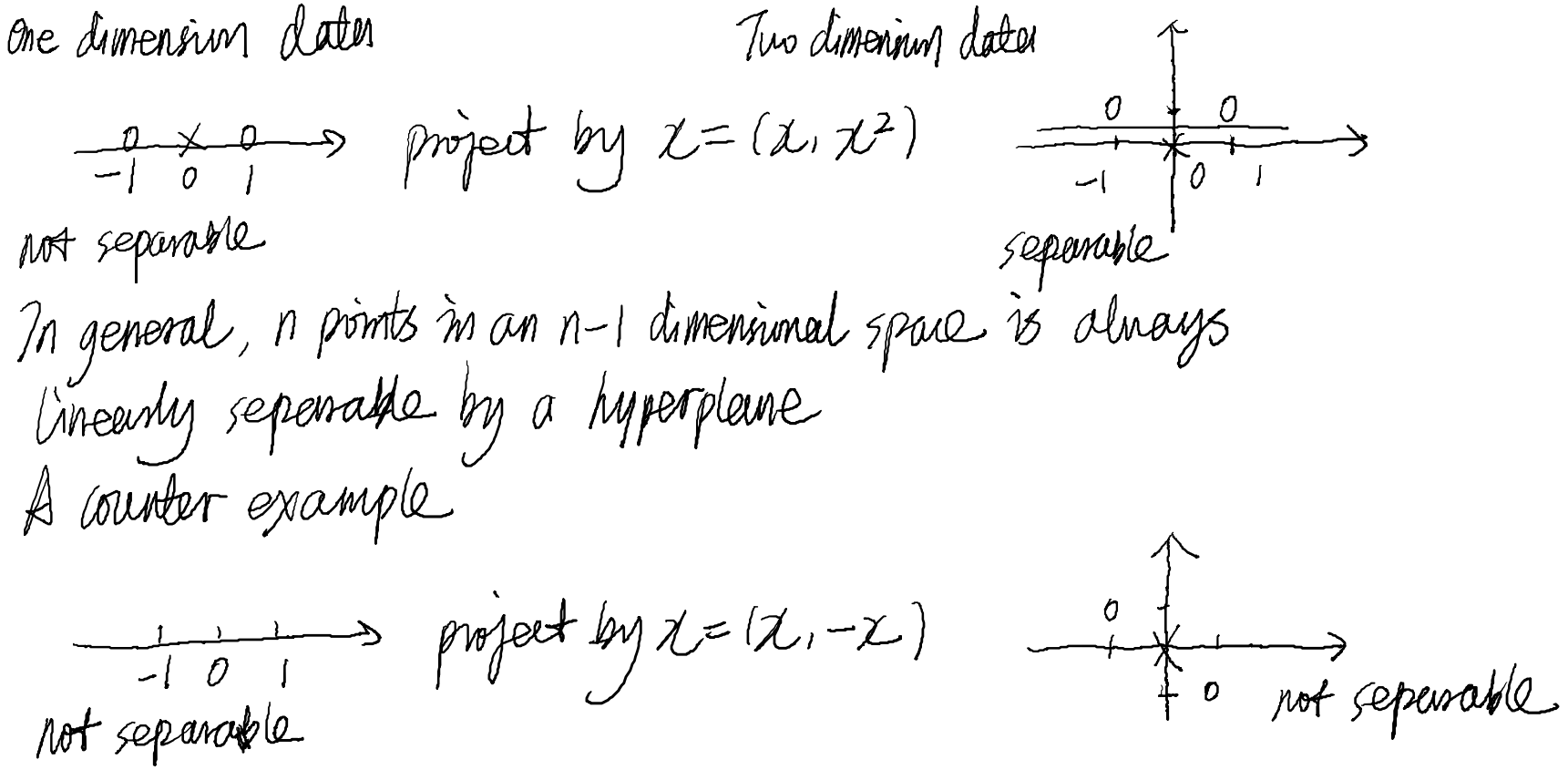
kernel trick $K(x_i, x_j)$

↓

- Making decision only by the support vectors $i \in SV$
- To train α_i and test new example, only need to specify the inner product $\langle x_i, x_j \rangle$

Non-linearly Separable

Feature Projection



Feature projection will result in overfitting

Soft Constraint

Recall the hard SVM

$$\hat{w}_{hard} = \arg\min_{w \in \mathbb{R}^n} \sum_{i=1}^{n} l_{0-\infty}(y_i f(x_i) - 1) + \frac{1}{2}\|w\|_2^2 \quad \text{where } f(x) = w^T x + b$$

$l_{0-\infty}(z)$ is 0 if $z \geq 0$ otherwise ∞ $\Leftrightarrow$ inequality constraints

Relax the hard version to a soft version

$$\hat{w}_{soft} = \arg\min_{w \in \mathbb{R}^n} l_{0-1}(y_i f(x_i) - 1) + \frac{\lambda}{2}\|w\|_2^2$$

$l_{0-1}(z)$ is 0 if $z + 1 \geq 0$ otherwise 1 i.e. $l_{0-1} = \begin{cases} 1 & \text{misclassified} \\ 0 & \text{correct classified} \end{cases}$

λ controls the trade-off between minimizing loss and Maximizing margin

Hinge Loss

0-1 loss is nonconvex in w

Approximate 0-1 loss by convex functions

Hinge loss

$$L_{hin}(f(x), y) = \max(1 - yf(x), 0)$$

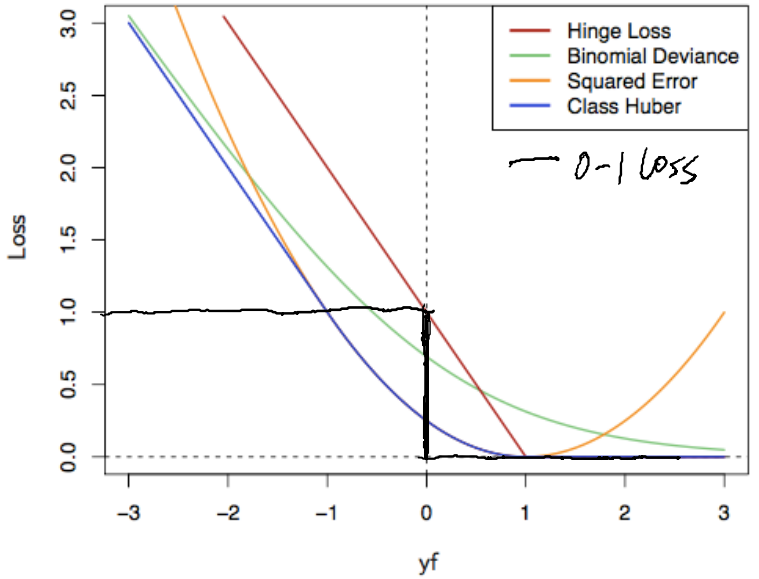
log loss (logistic regression) (Binomial Deviance)

we can define log loss for $y \in \{-1, 1\}$

so log loss $-\log p(y|x) = \log(1 + e^{-yf(x)})$

Square loss

$$L_{sq}(f(x), y) = (f(x) - y)^2$$



Loss Function	$L[y, f(x)]$	Minimizing Function
Binomial Deviance	$\log[1 + e^{-yf(x)}]$	$f(x) = \log \frac{\Pr(Y = +1\|x)}{\Pr(Y = -1\|x)}$
SVM Hinge Loss	$[1 - yf(x)]_+$	$f(x) = \text{sign}[\Pr(Y = +1\|x) - \frac{1}{2}]$
Squared Error	$[y - f(x)]^2 = [1 - yf(x)]^2$	$f(x) = 2\Pr(Y = +1\|x) - 1$
"Huberised" Square Hinge Loss	$-4yf(x)$, $yf(x) < -1$ $[1 - yf(x)]_+^2$ otherwise	$f(x) = 2\Pr(Y = +1\|x) - 1$

Equivalent Problem

$$\eta = \max(y, 0) = \min \xi \quad \text{s.t } y \le \xi \quad \xi \ge 0$$

Let $\eta(w) = \max(y(w), 0)$

$$\arg\min_w (\eta(w) + g(w)) = \arg\min_w \left(\min_\xi \xi + g(w) \right)$$

$$\text{s.t } y(w) \le \xi, \quad \xi \ge 0$$

$\eta(w) = \hat{\xi}(w)$ $\eta(\hat{w}) = \hat{\xi}$

Let $\eta_i(w) = \max(y_i(w), 0)$ for $\forall i = 1, \dots, n$

$$\arg\min_w \sum_{i=1}^n (\eta_i(w) + g(w)) = \arg\min_w \left(\min_{\xi_i} \sum_{i=1}^n \xi_i + g(w) \right)$$

$$\text{s.t } y_i(w) \le \xi_i, \quad \xi_i \ge 0$$

$\eta_i(w) = \hat{\xi}_i(w)$, $\eta_i(w) = \hat{\xi}_i$ $\forall i = 1, \dots, n$

The Primal Soft SVM

The primal soft SVM

$$\hat{w}_{soft}, \hat{b}_{soft} = \arg\min_{w,b} \sum_{i=1}^{n} L_{hinge}(w^T x_i + b, y_i) + \frac{\lambda}{2}\|w\|^2$$

Equals to ($C = \frac{1}{\lambda}$)

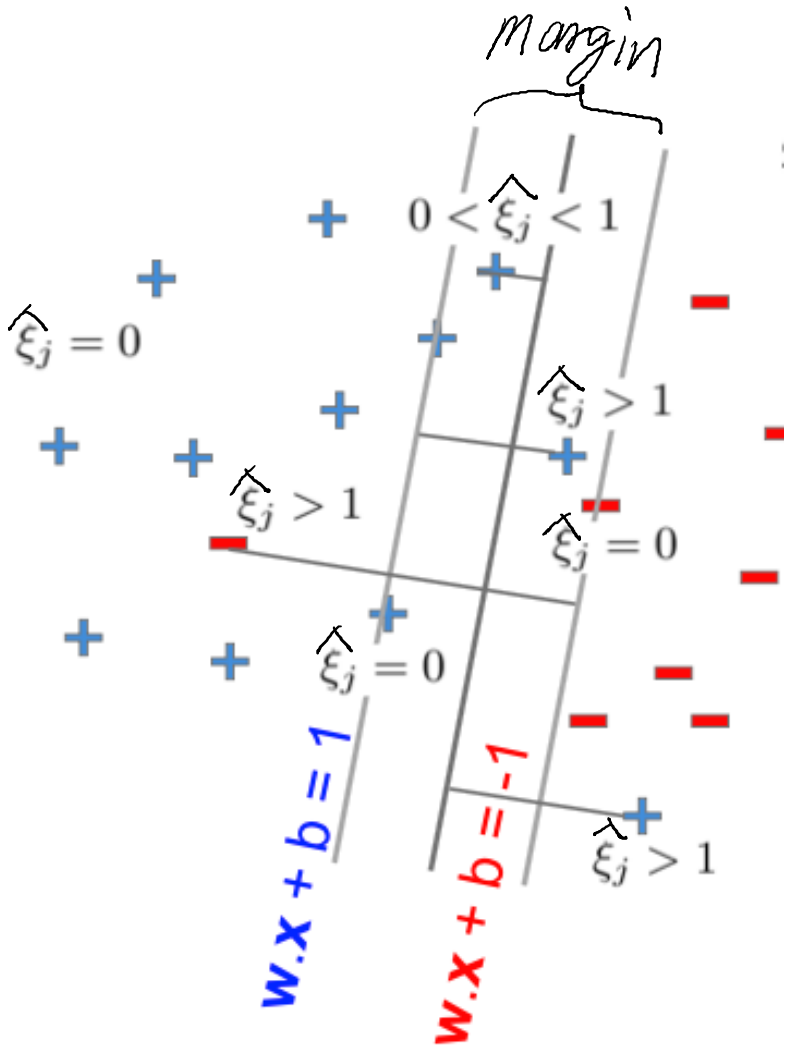
$$\hat{w}_{soft}, \hat{b}_{soft}, \hat{\xi} = \arg\min_{w,b,\xi} C\sum_{i=1}^{n} \xi_i + \frac{1}{2}\|w\|_2^2$$

Subject to $y_i(w^T x_i + b) \geq 1 - \xi_i, \forall i = 1, \dots, n$

ξ_i: slack variables

Given a (w, b), $\hat{\xi}_i$ has the following properties

- $\hat{\xi}_i = 0$ correct classified and out of margin
- $0 < \hat{\xi}_i \leq 1$ x_i is within the margins and correct classified
- $\hat{\xi}_i > 1$ misclassified



Lagrange Form and KTT

Lagrange form

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\xi_i - (y_i(w^T x_i + b) - 1 + \xi_i) - \sum_{i=1}^{n}\beta_i\xi_i$$

strong duality holds

$$\min_{w,b,\xi}\ \max_{\alpha_i \geq 0, \beta_i \geq 0} L = \max_{\alpha_i \geq 0, \beta_i \geq 0}\ \min_{w,b,\xi} L$$

KKT condition of soft SVM

$$\alpha_i \geq 0 \quad \beta_i \geq 0$$

$$\nabla_w L = 0 \Rightarrow \hat{w} = \sum_{i=1}^{n}\hat{\alpha}_i y_i x_i$$

$$\nabla_b L = 0 \Rightarrow \sum_{i=1}^{n}\hat{\alpha}_i y_i = 0$$

$$\nabla_{\xi_i} L = 0 \Rightarrow C - \hat{\alpha}_i - \hat{\beta}_i = 0 \Rightarrow \hat{\alpha}_i + \hat{\beta}_i = C \text{ for } \forall i = 1, \ldots, n$$

$$\forall i,\ \hat{\alpha}_i(y_i(\hat{w}^T x_i + \hat{b}) - 1 + \hat{\xi}_i) = 0 \Rightarrow \hat{\alpha}_i = 0 \text{ or } y_i(\hat{w}^T x_i + \hat{b}) = 1 - \hat{\xi}_i$$

$$\forall i,\ \hat{\beta}_i\hat{\xi}_i = 0 \Rightarrow \hat{\beta}_i = 0 \text{ or } \hat{\xi}_i = 0$$

The Dual Hard SVM

Simplify the lagrange form

$$L = L(\xi, \alpha, \beta) = \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} \xi_i (C - \alpha_i - \beta_i) - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

Note that $\beta_i = C - \alpha_i$, but we require $\beta_i \geq 0$ that $C - \alpha_i \geq 0$ if we eliminate β_i

$$\hat{\alpha} = \underset{\alpha \in R^n}{\operatorname{argmax}} \ \alpha^T 1_n - \frac{1}{2} \alpha^T y G y \alpha$$

$$s.t \quad 0 \leq \alpha \leq C \quad \alpha^T y = 0$$

where $y = \operatorname{diag}(y_1, \cdots, y_n)$ $G = \{G_{ij}\}_{i,j}^{n,n}$ where $G_{ij} = x_i^T x_j$

Solve it as a quadratic programming problem

$x_i^T x_j$ → kernel trick $K(\langle x_i, x_j \rangle)$

Support Vector

A support vector(sv) is a point where $\hat{\alpha}_i \neq 0$

By KTT condition.

$\hat{\alpha}_i \neq 0 \Rightarrow y_i(\hat{w}^T x_i + \hat{b}) = 1 - \hat{\xi}_i$

Under the equation above

If $\hat{\xi}_i = 0$, x_i is on the margin plane

(not the ones correct classified and out of margin)

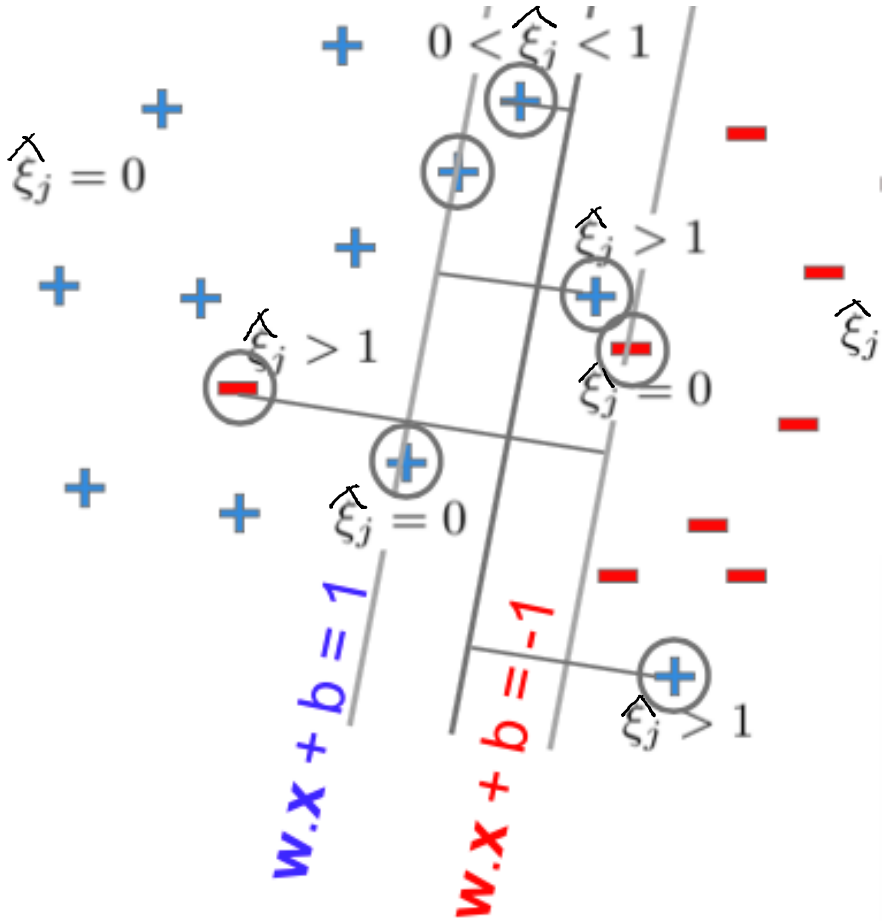
If $0 < \hat{\xi}_i \leq 1$ x_i is within the margin and correct classified

If $\hat{\xi}_i > 1$ misclassified points

Note that when $\hat{\xi}_i \neq 0 \Rightarrow \hat{\beta}_i = 0 \Rightarrow \hat{\alpha}_i = C$ i.e.

when $0 < \hat{\alpha}_i < C \Rightarrow \hat{\beta}_i > 0 \Rightarrow \hat{\xi}_i = 0$

support vectors : data points with circles



Soft SVM Predictor

Direction

$$\hat{w} = \sum_{i \in SV} \alpha_i y_i x_i$$

Bias parameter

For any $0 < \alpha_i < C$ i.e. $\hat{\xi}_i = 0$ $y_i(\hat{w}^T x_i + \hat{b}) = 1$. x_i lies on the margin

Taking average $\hat{b} = \frac{1}{N_{margin}} \sum_{i \in margin} (y_i - \sum_{j \in SV} \hat{\alpha}_j y_j x_j^T x_i)$

For testing $\hat{w}^T x + \hat{b} = \sum_{i \in SV} \alpha_i y_i x_i^T x + \hat{b}$

$x_i^T x$ → kernel trick $K(\langle x_i, x \rangle)$

SVM for Regression

Multi-class SVM