Goal: Given a dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

find a w with maximum width s.t all points in dataset D are classified correctly

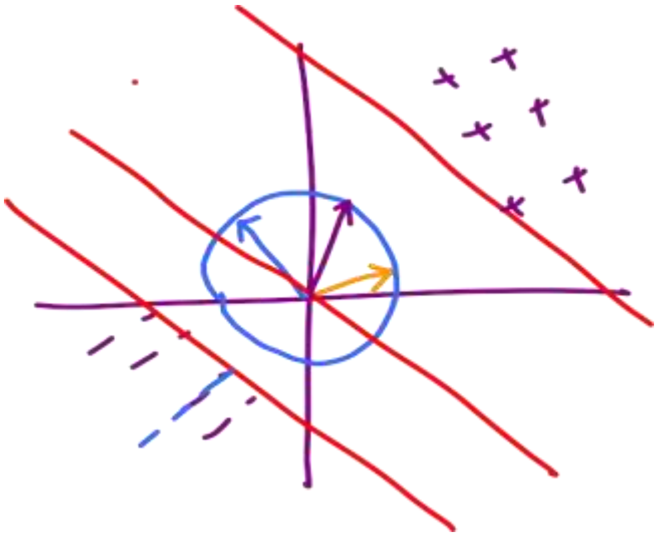
$$\max_{w, \gamma} \ \gamma$$

s.t

$$(w^T x_i) y_i \geq \gamma \quad \forall i$$

Issue:

Can scale w arbitrarily.



Possible fix

$$\max_{w,\gamma} \ \gamma$$

$$(w^T x_i) y_i \geq \gamma$$

$$\|w\|^2 = 1$$

Goal: Given a dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

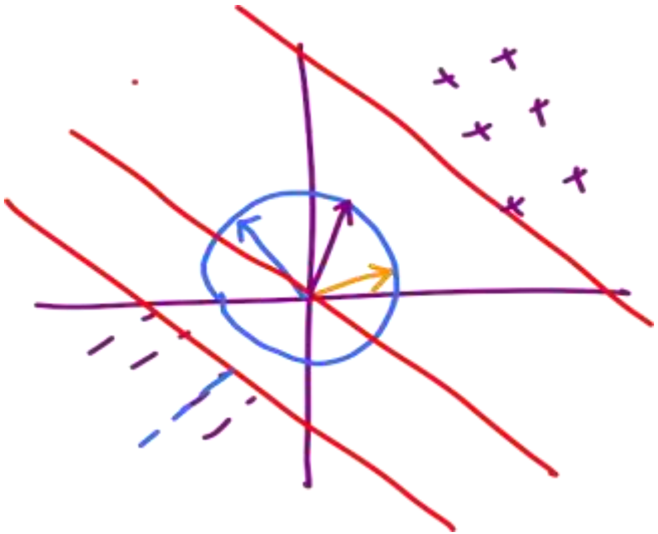
find a w with maximum width s.t all points in dataset D are classified correctly

$$\max_{w, \gamma} \gamma$$

s.t

$$(w^T x_i) y_i \geq \gamma \quad \forall i$$

Issue: Can scale w arbitrarily.

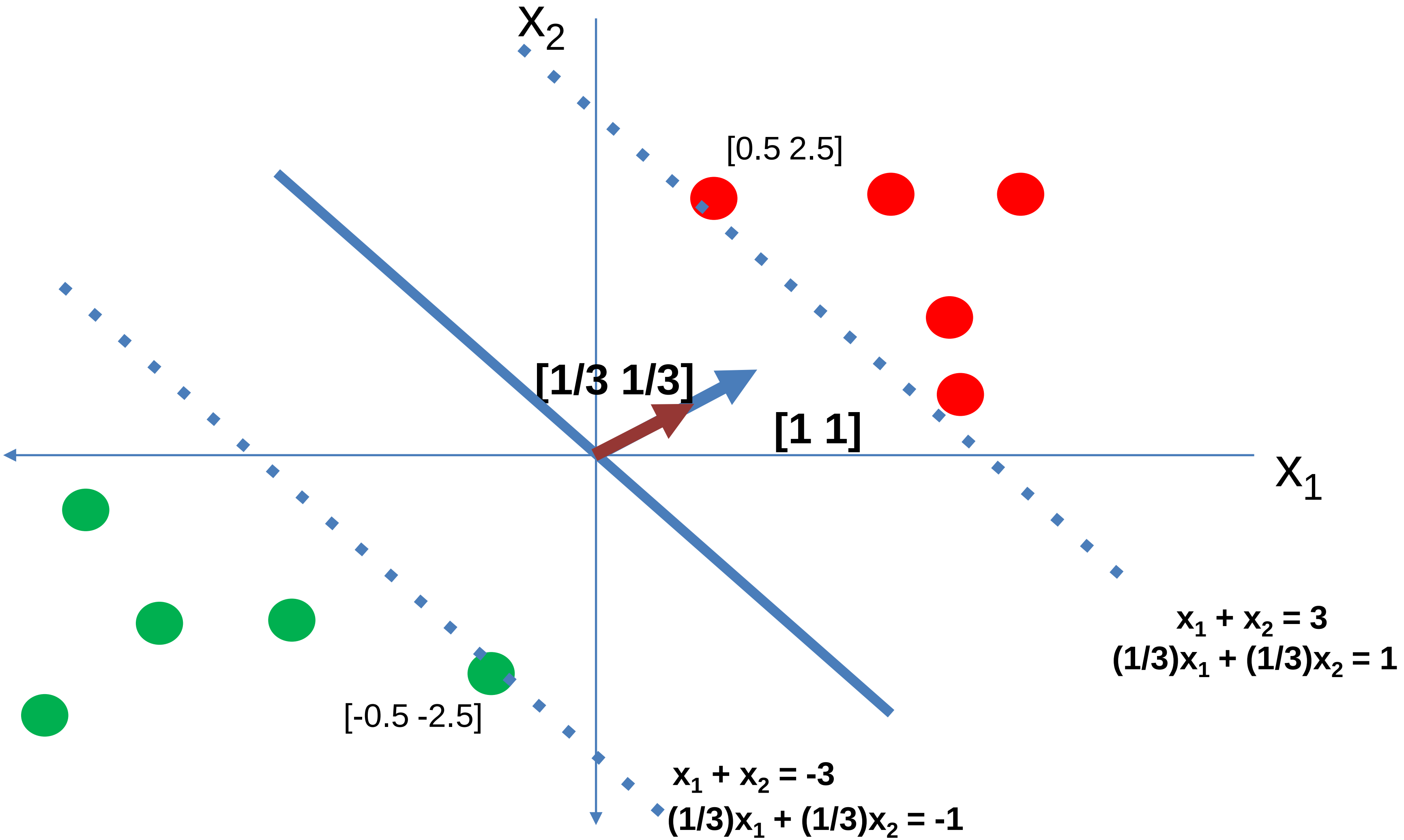


Possible fix

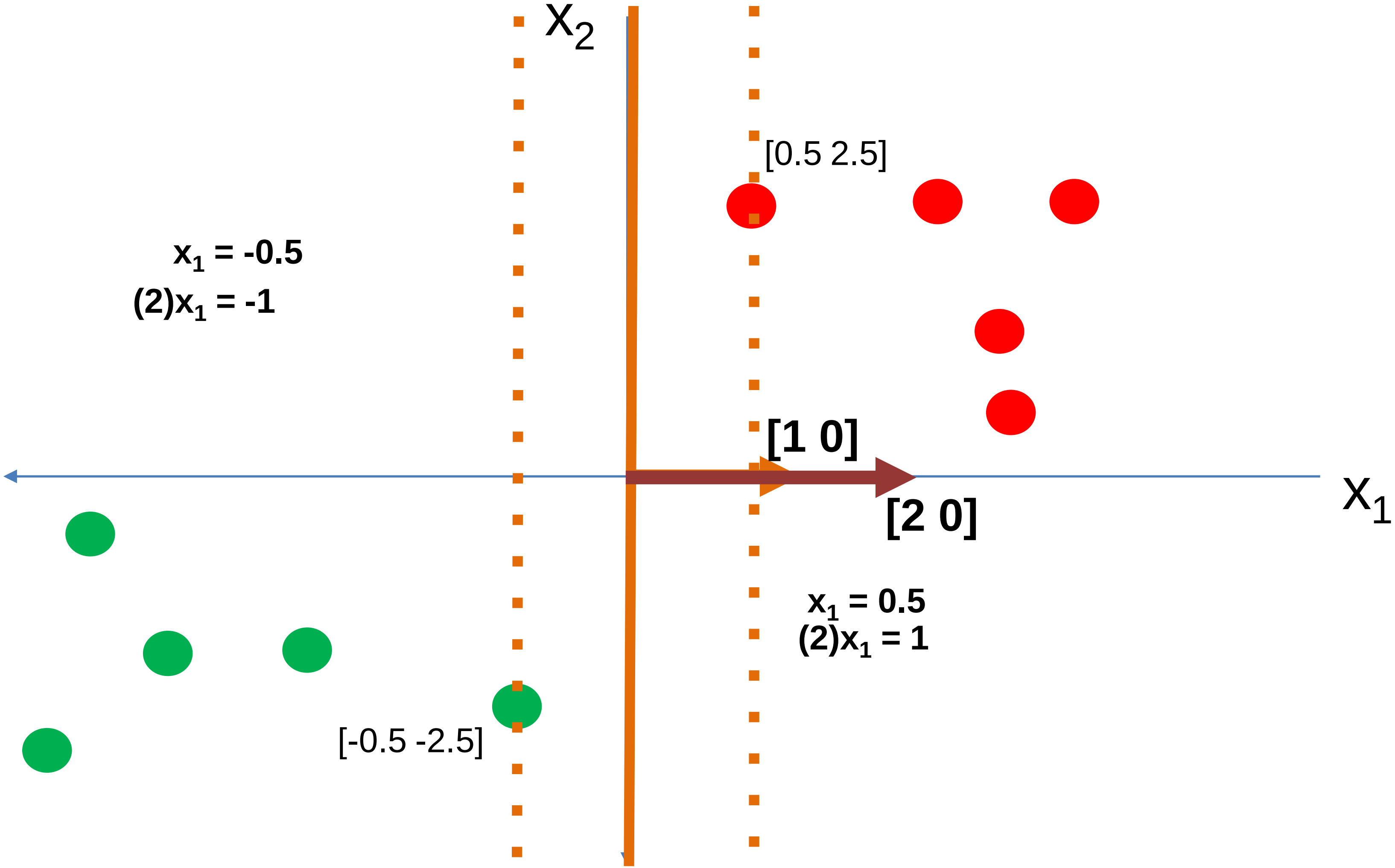
$$\max_{w,\gamma} \ \gamma$$

$$(w^T x_i) y_i \geq \gamma$$

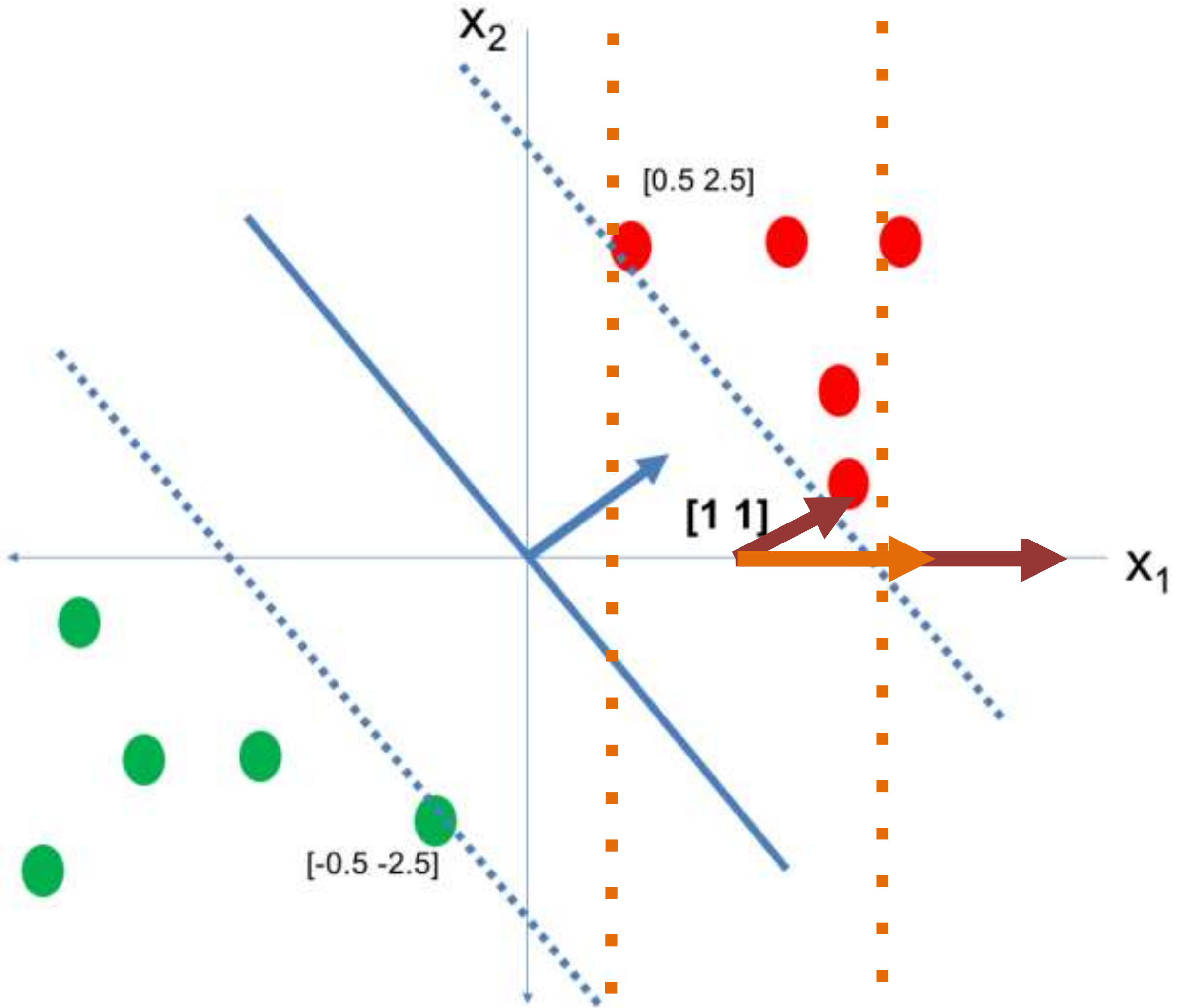
$$\|w\|^2 = 1$$



X_2
[0.5 2.5]
[1/3 1/3]
[1 1]
X_1
$x_1 + x_2 = 3$
$(1/3)x_1 + (1/3)x_2 = 1$
[-0.5 -2.5]
$x_1 + x_2 = -3$
$(1/3)x_1 + (1/3)x_2 = -1$



X_2
[0.5 2.5]
$x_1 = -0.5$
$(2)x_1 = -1$
[1 0]
[2 0]
X_1
$x_1 = 0.5$
$(2)x_1 = 1$
[-0.5 -2.5]



Notice what happens to the lengths of the w as we adjust it to have margin 1

OBSERVATIONS

-> Once a direction is fixed,
the width between the margin lines
is fixed

-> If the width is large, then the w that
achieves margin 1 in that direction
has smaller length

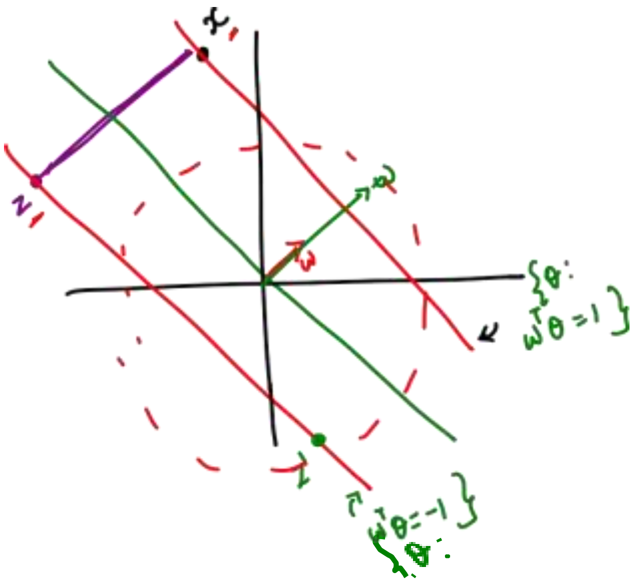
-> If the width is small, then the w that
achieves margin 1 in that direction
has larger length

-> In general, width(w) seems to be
inversely proportional to length(w)

$$\max_{w} \quad \boxed{\text{width}(w)}$$

$$\text{s.t} \quad (w^T x_i) y_i \geq 1 \quad \forall i$$

What is width(w)?



$$\min_{z:} \quad \underbrace{\frac{1}{2}\|x - z\|^2}$$

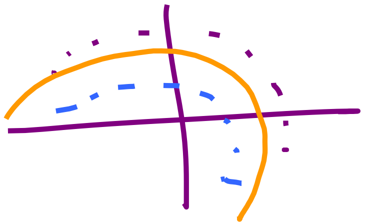
$$\text{s.t.} \quad \left.\begin{array}{l} w^T x = 1 \\ w^T z = -1 \end{array}\right\}$$

[Exercise]

$$\text{width}(w) = \frac{2}{\|w\|^2}$$

$$\max_{w} \frac{2}{\|w\|^2}$$

$$\text{s.t } \forall i \quad (w^T x_i) y_i \geq 1$$



Issues

- L.S is a strong assumption

$$\min_{w} \frac{1}{2} \|w\|^2$$

$$\text{s.t } \forall i \quad (w^T x_i) y_i \geq 1$$

- Non-linear structure?

DETOUR

$$\boxed{\begin{aligned} &\min_{\omega} \; f(\omega) \\ &\quad g(\omega) \le 0 \end{aligned}}$$

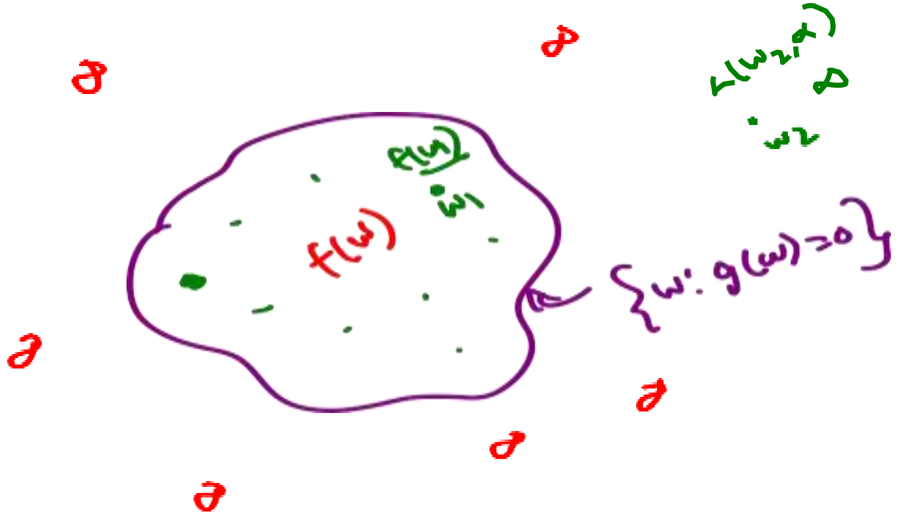
$$\underline{\mathcal{L}(\omega, \alpha)} = f(\omega) + \alpha \cdot g(\omega)$$

Fix some ω.

Consider

$$\boxed{\max_{\alpha \geq 0} \; L(w, \alpha)}$$

$$= \max_{\alpha \geq 0} \; \underline{f(w)} + \underline{\alpha \, g(w)}$$



$$\begin{cases} \infty & \text{if } \; g(w) > 0 \\ f(w) & \text{if } \; g(w) \leq 0 \end{cases}$$

$$\min_{w} \Bigg[\underbrace{\overset{\theta(w)}{\overline{\max_{\alpha \geq 0} \; \mathcal{L}(w,\alpha)}}} \Bigg] \equiv \text{equivalent} \quad \left. \begin{array}{l} \min_{w} \; f(w) \\ \text{s.t} \;\; g(w) \leq 0. \end{array} \right\}$$

- Can we swap min and max?

mi
w

Multiple constraints

→ Same idea

$$\min_{w} f(w) \quad \text{s.t.} \quad g_i(w) \le 0 \quad \forall i = 1 \ldots k$$

$$= \min_{w} \left[\max_{\{\alpha_1, \ldots, \alpha_k \ge 0\}} \left[f(w) + \underbrace{\sum_{i=1}^{k} \alpha_i \underline{g_i(w)}} \right] \right]$$

||| Strong duality for convex f, g_i

$$\max_{\alpha_1, \ldots, \alpha_k \ge 0} \; \min_{w} \; f(w) + \sum_{i=1}^{k} \alpha_i g_i(w)$$

$$\min_{w} \underbrace{\frac{1}{2}\|w\|^2}_{} \leftarrow f(w)$$

$$\text{s.t.} \quad \underbrace{(w^T x_i) y_i \geq 1}_{} \quad \forall i$$

$$\underbrace{1-(w^T x_i) y_i}_{} \leq 0$$

$$g_i(w) = 1 - (w^T x_i) y_i$$

$$\mathcal{L}(w, \underset{\uparrow \in \mathbb{R}^n}{\alpha}) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right)$$

$$\min_{w} \left[\max_{\alpha \geq 0} \; \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right) \right]$$

$$\to \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} \geq 0$$

$$|||$$

$$\underline{\max_{\alpha \geq 0}} \left[\min_{w} \; \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right) \right]$$

Fix $\alpha \geq 0$

$$\min_{w} \left[\underbrace{\frac{1}{2}\|w\|^2} + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right) \right]$$

Grad w.r.t w

$$w^* + \sum_{i=1}^{n} -\alpha_i x_i y_i = 0$$

$$\boxed{w^* = \sum_{i=1}^{n} \alpha_i x_i y_i}$$

$\in \mathbb{R}^d$

$\{+1, -1\}$

Fixed choice

In matrix notation

$$\boxed{w^* = X Y \alpha}$$

$$X = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix}_{d \times n} \quad Y = \begin{bmatrix} y_1 & & & \\ & \ddots & & \\ & & & y_n \end{bmatrix}_{n \times n} \quad \alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}_{n \times 1}$$

Substituting $\underline{\text{soln}}$ back in the objective.

$$\frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right)$$

$$= \frac{1}{2} \underline{w^T w} + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i (w^T x_i) y_i$$

$$\underbrace{\frac{1}{2}(XY\alpha)^T(XY\alpha)} + \alpha^T \mathbf{1} - \underbrace{\sum_{i=1}^{n} (XY\alpha)^T x_i \, y_i \, \alpha_i}$$

$$\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \begin{bmatrix} \vdots \\ \vdots \end{bmatrix}$$

On simplification [please do this]

$$\alpha^T \mathbf{1} - \frac{1}{2}(XY\alpha)^T(XY\alpha)$$

DUAL PROBLEM

$$\max_{\boxed{\alpha \geq 0}} \; \alpha^T \mathbf{1} - \frac{1}{2}\alpha^T Y^T X^T X Y \alpha$$

Solving in n instead of d.

easy constraints

can be KERNELIZED!

n×n.

Revisiting the Lagrangian

$$\min_{w} \left[\overbrace{\max_{\alpha \geq 0} \; f(w) + \alpha g(w)} \right] \equiv \max_{\alpha \geq 0} \left[\min_{w} \; f(w) + \alpha g(w) \right]$$

PRIMAL DUAL

w^* α^*

$$\boxed{\underbrace{\max_{\alpha \geq 0} \; f(w^*) + \alpha \, g(w^*)}} = \min_{w} \; f(w) + \alpha^* g(w)$$

$$f(w^*) = f(w') + \alpha^* g(w')$$

$$f(w^*) \leq f(w^*) + \alpha^* g(w^*)$$

$$\Rightarrow \boxed{\alpha^* g(w^*) \geq 0}$$

But we know $\alpha^* g(w^*) \leq 0$

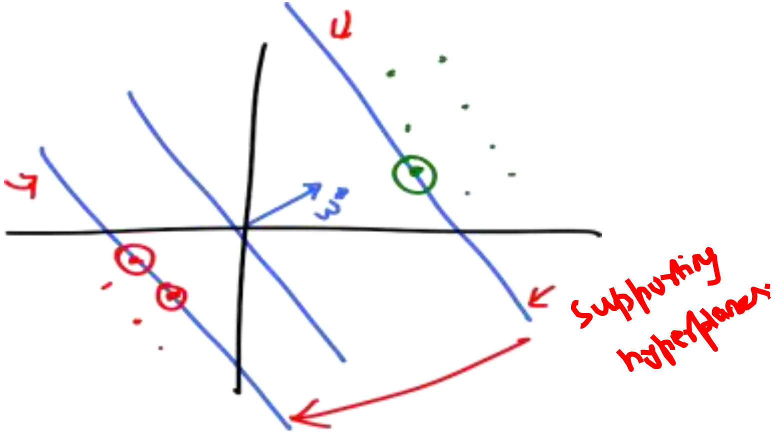
$$\Rightarrow \boxed{\alpha^* g(w^*) = 0} \rightarrow \text{COMPLEMENTARY SLACKNESS}$$

For multiple constraints

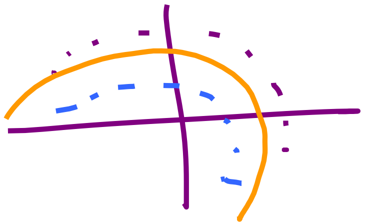
$$\alpha_i^* \, g_i(w^*) = 0 \quad \forall i$$

For our problem

$$\boxed{\alpha_i^* \left(1 - (w^{*T} x_i) y_i\right) = 0 \quad \forall i}$$



$$\boxed{\min_{\omega} \quad \frac{1}{2}\|\omega\|^2 \qquad \text{s.t.} \; \forall i \; (\omega^T x_i) y_i \geq 1}$$



Issues

- L.S is a strong assumption

$$\boxed{\bullet \text{ Non-linear structure?}}$$

So far

Support Vector Machines

Primal Problem – Margin Maximization

Dual Problem

- Kernel Version

Now

- What if there are **outliers** in the problem?

Idea (to deal with outliers):

Fix any w. w classifies some points correct and some incorrectly. Let the incorrect points pay "bribe" to get to the correct side.

Modified formulation

$c > 0$ [hyper parameter]

$$\min_{\boxed{w, \xi}} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

$\rightarrow (w^T x_i) y_i + \underline{\xi_i} \geq 1 \leftarrow \forall i$

$\rightarrow \underline{\xi_i \geq 0} \leftarrow \forall i$

If $C = 0 \Rightarrow$ Bribes don't cost $\Rightarrow$ $w = 0$ is solution

$C \rightarrow \infty \Rightarrow$ Bribes are too costly $\Rightarrow$ Linear separable case.

$$\mathcal{L}(\omega, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + \underbrace{c\left(\sum_{i=1}^{n} \xi_i\right)}_{\uparrow} + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i)\, y_i - \xi_i\right) + \sum_{i=1}^{n} \beta_i \left(-\xi_i\right)$$

Dual:

$$\max_{\substack{\alpha \geq 0 \\ \beta \geq 0}} \; \min_{\omega} \; \mathcal{L}(w, \xi, \alpha, \beta)$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w^* = \sum_{i=1}^{n} \alpha_i x_i y_i$$

$$\boxed{w^* = XY\alpha}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \boxed{C - \alpha_i - \beta_i = 0}$$

$$\boxed{\alpha_i + \beta_i = C} \quad \forall i$$

Substitute $w^* = XY\alpha$ in the original objective

$$\frac{1}{2}(XY\alpha)^T(XY\alpha) + \sum_{i=1}^{n}(C - \underline{\alpha_i - \beta_i)\xi_i} + \alpha^T 1$$

$$- (XY\alpha)^T(XY\alpha)$$

SOFT - MARGIN

SUPPORT

VECTOR

MACHINE

$$\max_{\substack{\alpha \geq 0 \\ \beta \geq 0 \\ \alpha_i + \beta_i = C}} \quad \alpha^T 1 - \frac{1}{2} (X y \alpha)^T (X y \alpha)$$

$$\equiv \quad \boxed{\max_{\underbrace{0 \leq \alpha \leq C}_{\text{BOX CONSTRAINT.}}} \quad \underbrace{\alpha^T 1 - \frac{1}{2} \alpha^T y^T (X^T X) y \alpha}}$$

HARD-MARGIN SVM

PRIMAL

$$\min_{w} \frac{1}{2}\|w\|^2$$

s.t. $\underbrace{(w^T x_i) y_i \geq 1}_{1 - w^T x_i y_i \leq 0} \quad \forall i$

DUAL

$$\max_{\alpha \geq 0} \alpha^T 1 - \alpha^T y^T x^T x y \alpha$$

$\alpha_i^*(1 - w^{*T} x_i y_i) = 0 \quad \forall i$

$$\boxed{w^* = \sum_{i=1}^{n} \alpha_i^* x_i y_i}$$

SOFT-MARGIN SVM

PRIMAL ✓

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

s.t. $(w^T x_i) y_i + \xi_i \geq 1 \quad \forall i = 1, \ldots, n$ — α

$\xi_i \geq 0 \quad \forall i = 1, \ldots, n$ — β

DUAL ✓

$$\max_{\substack{\boxed{\alpha + \beta = C} \\ \alpha \geq 0 \\ \beta \geq 0}} \alpha^T 1 - \alpha^T y^T x^T x y \alpha$$

$0 \leq \alpha \leq C$

- Let $(\underline{w}^*, \underline{\xi}^*)$ be the primal optimal solution
- Let $(\underline{\alpha}^*, \underline{\beta}^*)$ be the dual optimal solution

COMPLEMENTARY SLACKNESS

$$\underline{\alpha_i^*}\left(\underbrace{1 - (w^{*T}x_i)y_i - \xi_i^*}\right) = 0 \quad \forall i$$

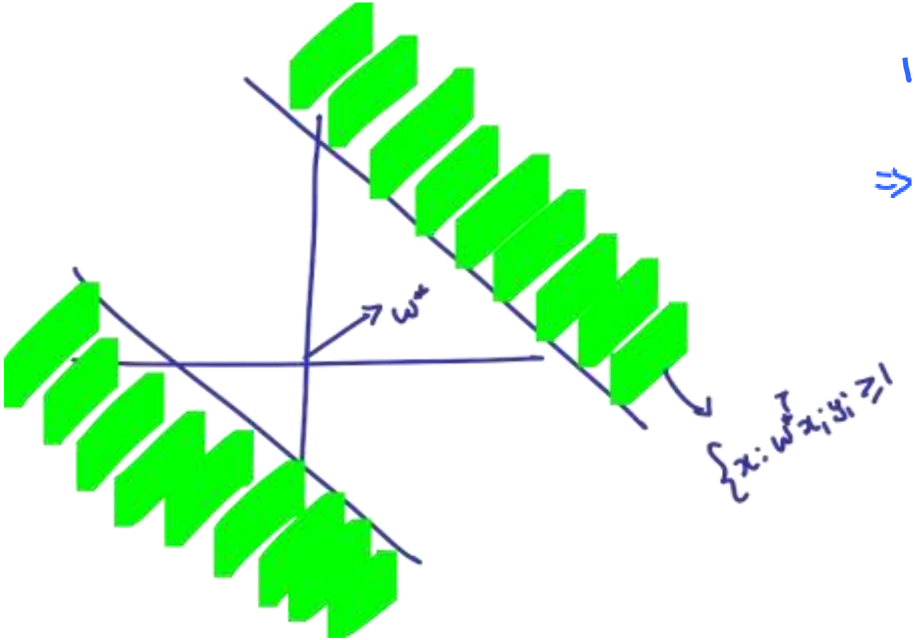
$$\beta_i^* \xi_i^* = 0 \quad \forall i$$

$\alpha_i^* + \beta_i^* = C$
$\forall i$

$\hookrightarrow$ (A)

Various cases possible

Case 1: $\alpha_i^* = 0$ $\overset{④}{\Rightarrow}$ $\beta_i^* = C$ $\overset{\boxed{CS}}{\Rightarrow}$ $\underline{\xi_i^*} = 0$



$1 - (w^{*T} x_i) y_i - \underbrace{\xi_i^*}_{=0} \leq 0$ [Primal feasibility]

$\Rightarrow 1 - (w^{*T} x_i) y_i \leq 0$

$\Rightarrow w^{*T} x_i y_i \geq 1$

$\Rightarrow w^*$ classifies (x_i, y_i) correctly.

Case 2: $0 < \alpha_i^* < C \overset{(A)}{\Longrightarrow} 0 < \beta_i^* < C \overset{CS}{\Longrightarrow} \underline{\xi_i^*} = 0$

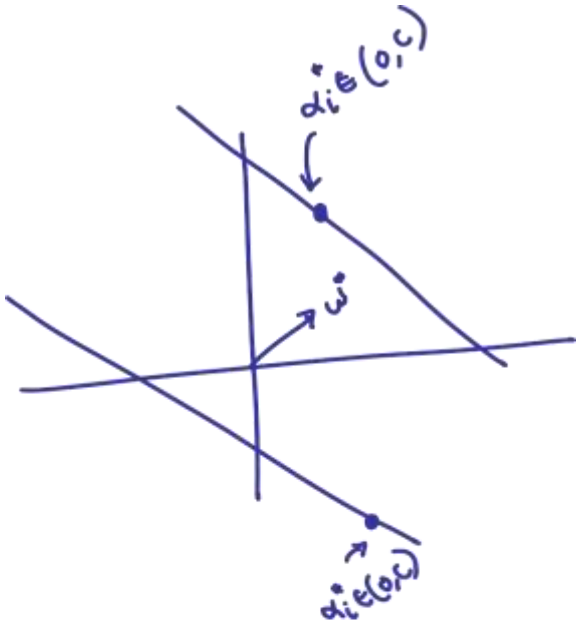
$$\Big\Downarrow \boxed{CS}$$

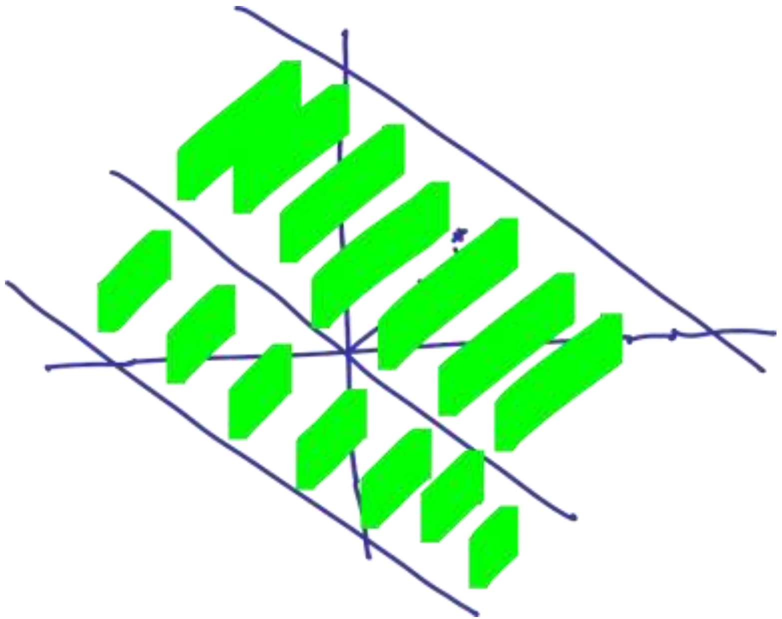
$$1 - (w^{*T} x_i) y_i - \xi_i^* = 0$$

$$\Big\Downarrow$$

$$(w^{*T} x_i) y_i = 1$$

$\Rightarrow$ (x_i, y_i) lies on the supporting hyperplane.





$\underline{\text{Case 3}}$: $\quad \underline{\alpha_i^* = C} \Rightarrow \beta_i^* = 0 \Rightarrow \xi_i^* \geq 0$

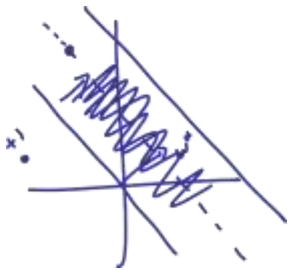
$\Downarrow \boxed{CS}$

$$1 - w^{*T} x_i y_i - \xi_i^* = 0$$

$$\xi_i^* = 1 - w^{*T} x_i y_i \geq 0$$

$$\Rightarrow \boxed{w^{*T} x_i y_i \leq 1}$$

Let's see this from P.O.V of data



CASE 1

$$\boxed{w^{*T} x_i y_i < 1}$$

$$1 - w^T x_i y_i - \xi_i^* \leq 0$$

$$w^{*T} x_i y_i \geq 1 - \xi_i^*$$

$$\boxed{\xi_i^* \geq 1 - w^{*T} x_i y_i}$$

$$\Rightarrow \xi_i^* > 0 \Rightarrow \beta_i^* = 0 \Rightarrow \alpha_i^* = C$$

$$\alpha_i^* (1 - w^{*T} x_i y_i - \xi_i^*) = 0$$

$$\beta_i^* \xi_i^* = 0.$$

CASE 2: $w^{*T} x_i y_i = 1$

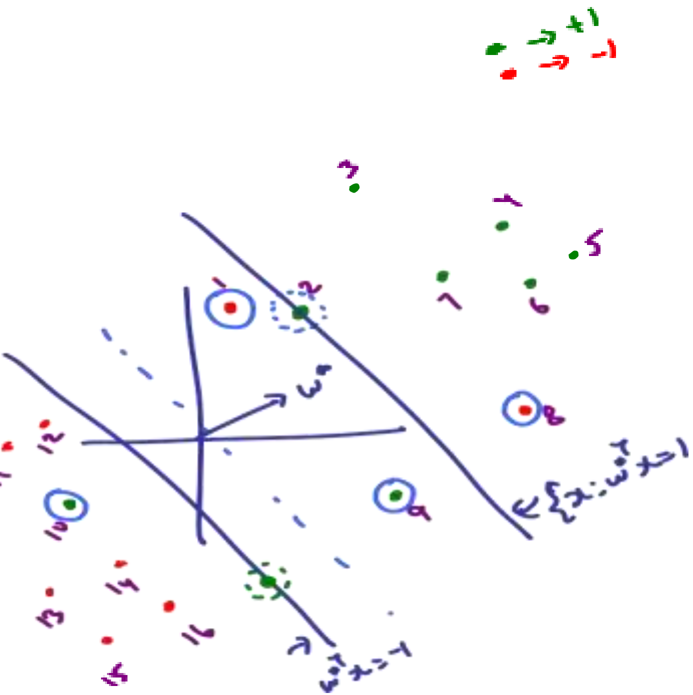
$$\xi_i^* \geq 1 - w^{*T} x_i y_i$$

$$\Rightarrow \quad \xi_i^* \geq 0 \quad \Rightarrow \quad \alpha_i^* \in [0, C]$$

CASE 3 $w^{*T} x_i y_i > 1$

$$1 - w^{*T} x_i y_i - \xi_i^* \leq 0 \quad \text{[Primal feasibility]}$$

$$\Rightarrow \quad 1 - w^{*T} x_i y_i - \xi_i^* < 0 \quad \overset{\boxed{\text{CS}}}{\Longrightarrow} \quad \alpha_i^* = 0$$



SUMMARY

$$\alpha_i^* = 0 \Rightarrow w^{*T} x_i y_i \geq 1$$

$$0 < \alpha_i^* < C \Rightarrow w^{*T} x_i y_i = 1$$

$$\alpha_i^* = C \Rightarrow w^{*T} x_i y_i \leq 1$$

✓ $w^{*T} x_i y_i < 1 \Rightarrow \alpha_i^* = C$

$\rightarrow$ $w^{*T} x_i y_i = 1 \Rightarrow \alpha_i^* \in [0, C]$

$\rightarrow$ $w^{*T} x_i y_i > 1 \Rightarrow \alpha_i^* = 0.$

Binary classification

GENERATIVE

Naive Bayes

G.D.A

DISCRIMINATIVE

K-NN

Decision trees

Perceptron

Support-vector-machines

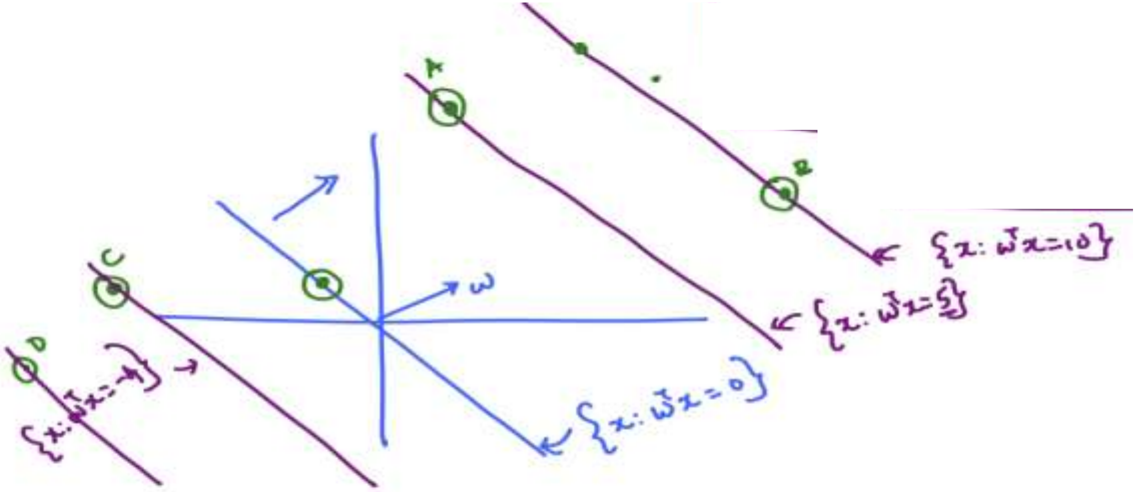
Dosen't "really" model $\boxed{P(y/x)}$

→ Just finds $f: \mathbb{R}^d \to \{\pm 1\}$

- Can we model $P(y=+1/x)$ differently?

Start with a simple model

Given $x \in \mathbb{R}^d$ $\quad z = w^T x \quad w \in \mathbb{R}^d$.



$$\boxed{P(y=+1/x)} = g(w^T x)$$

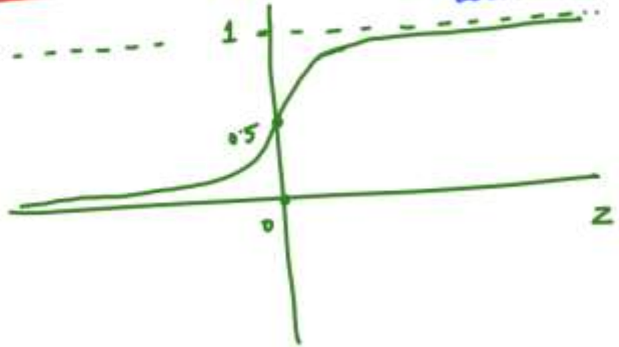
LINK FUNCTION

- $\underline{g(z)} \in \underline{[0,1]}$
- $g(z) \rightarrow \boxed{1}$ as $z \rightarrow \boxed{\infty}$
- $g(z) \rightarrow 0$ as $z \rightarrow -\infty$
- $g(z) = 0.5$ if $z = 0$.

ONE POPOLAR CHOICE

$$g(z) = \frac{1}{1+e^{-z}}$$

SIGMOID FUNCTION
LOGISTIC FUNCTION.



MODEL: LOGISTIC REGRESSION

Data: $\{ (x_1, y_1) \cdots (x_n, y_n) \}$ $\quad x_i \in \mathbb{R}^d$

$y_i \in \{0,1\}$

Max. Likelihood

$$\mathcal{L}(w, \text{Data}) = \prod_{i=1}^{n} \left(g(w^T x_i)\right)^{y_i} \left(1 - g(w^T x_i)\right)^{(1-y_i)}$$

$$\log \mathcal{L}(w, \text{Data}) = \sum_{i=1}^{n} y_i \log\left(g(\underline{w^T x_i})\right) + (1-y_i) \log\left(1 - g(w^T x_i)\right)$$

$$= \sum_{i=1}^{n} \left[y_i \log\left(\frac{1}{1+e^{-w^T x_i}} \right) + (1-y_i) \log\left(\frac{e^{-w^T x_i} \times 1}{1+e^{-w^T x_i}} \right) \right]$$

$$= \sum_{i=1}^{n} \left[\log\left(\frac{e^{-w^T x_i}}{1+e^{-w^T x_i}} \right) - y_i(-w^T x_i) \right]$$

$$= \sum_{i=1}^{n} \left[(1-y_i)(-w^T x_i) - \log\left(1+e^{-w^T x_i}\right) \right]$$

- No closed form solution

• Gradient ascent

$$\nabla \log L(w) = \sum_{i=1}^{n} (1-y_i)(-x_i) - \frac{e^{-w^T x_i}}{1+e^{-w^T x_i}}(-x_i)$$

$$= \sum_{i=1}^{n} x_i \left(y_i - \left(1 - \frac{e^{-w^T x_i}}{1+e^{-w^T x_i}} \right) \right)$$

$i=1$

$$= \boxed{\sum_{i=1}^{n} \overset{\mathbb{R}^d}{x_i} \left(\overset{\{0,1\}}{y_i} - \overset{g(w^T x_i)}{\frac{1}{1+e^{-w^T x_i}}} \right)}$$

$$\boxed{w_{t+1} = w_t + \eta_t \nabla \log L(w_t)}$$

REGULARIZED VERSION

$$\min_{w} \sum_{i=1}^{n} (1-y_i) w^T x_i + \log\left(1 + e^{-w^T x_i}\right) + \frac{\lambda}{2} \|w\|^2$$

KERNEL VERSION

- Can argue $w = \sum_{i=1}^{n} \alpha_i x_i$ [Formal Theorem: Representer Theorem]

Exercise: Derive the kernel version of logistic regression

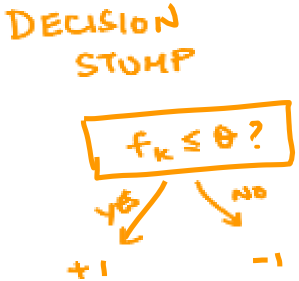
META CLASSIFIERS (or)

ENSEMBLE CLASSIFIERS.

WEAK CLASSIFIERS → STRONG CLASSIFIERS

[better than random]

Weak classifiers



high bias, low variance

Overfit decision tree

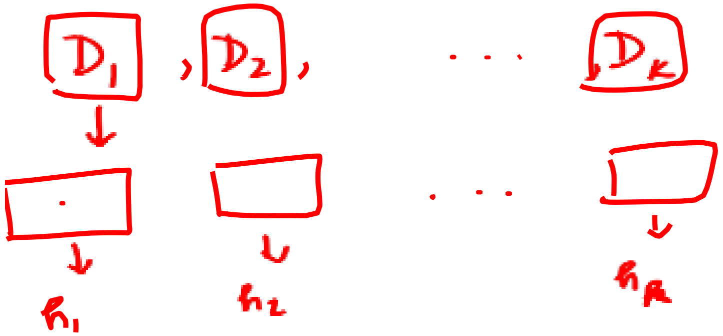
...

low bias, high variance

$$x_1, x_2, \dots x_n \sim \mathcal{N}(\mu, 1)$$

$$\hat{\mu}_1 = x_1 \qquad \hat{\mu}_2 = x_2, \dots \hat{\mu}_N = x_n \qquad \hat{\mu}_{ML} = \frac{1}{n}\sum x_i$$

Overfit decision trees



$$h_i : \mathbb{R}^d \to \{\pm 1\}$$

$$h^*(x) = \text{majority}(h_1(x), \dots, h_k(x))$$

BAGGING - Bootstrap Aggregation.

$D = \{(x_1,y_1) \dots (x_n,y_n)\}$

- Create datasets $D_1, \dots, D_k$ from D by "sampling with replacement".

- Run weak classifier on $D_1, \dots, D_k$ to get $h_1, \dots, h_k$

- Aggregate $h_1, \dots, h_k$ using majority.

Chance that a point appears in a dataset

$$1 - \underbrace{\left(1-\frac{1}{n}\right)\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{1}{n}\right)}$$

$$1 - \underbrace{\left(1-\frac{1}{n}\right)^n}$$

$$1 - \frac{1}{e} \quad (\text{as } n \to \infty)$$

$$\simeq 66\%$$

FEATURE BAGGING → Bag the features in addition to data points

Feature bagged decision trees -> RANDOM FOREST

BOOTSTRAP - Sampling with Replacement ?

AGGREGATION - Majority. ?

BOOSTING

[Freund & Schappire. 1995 Godel Prize]

ADA-BOOST

Distribution D over $(\mathcal{X} \times \mathcal{Y})$ — $\mathcal{X}$: $\mathbb{R}^d$, $\mathcal{Y}$: $\{+1,-1\}$

unknown but fixed.

$x_1, \ldots, x_n$ are iid from D.

$h: \underset{\mathcal{X}}{\mathbb{R}^d} \to \underset{\mathcal{Y}}{\{\pm 1\}}$.

Measure performance using

$$P_{(x,y)\sim D} \left(h(x) \neq y \right)$$

Misclassification probability.

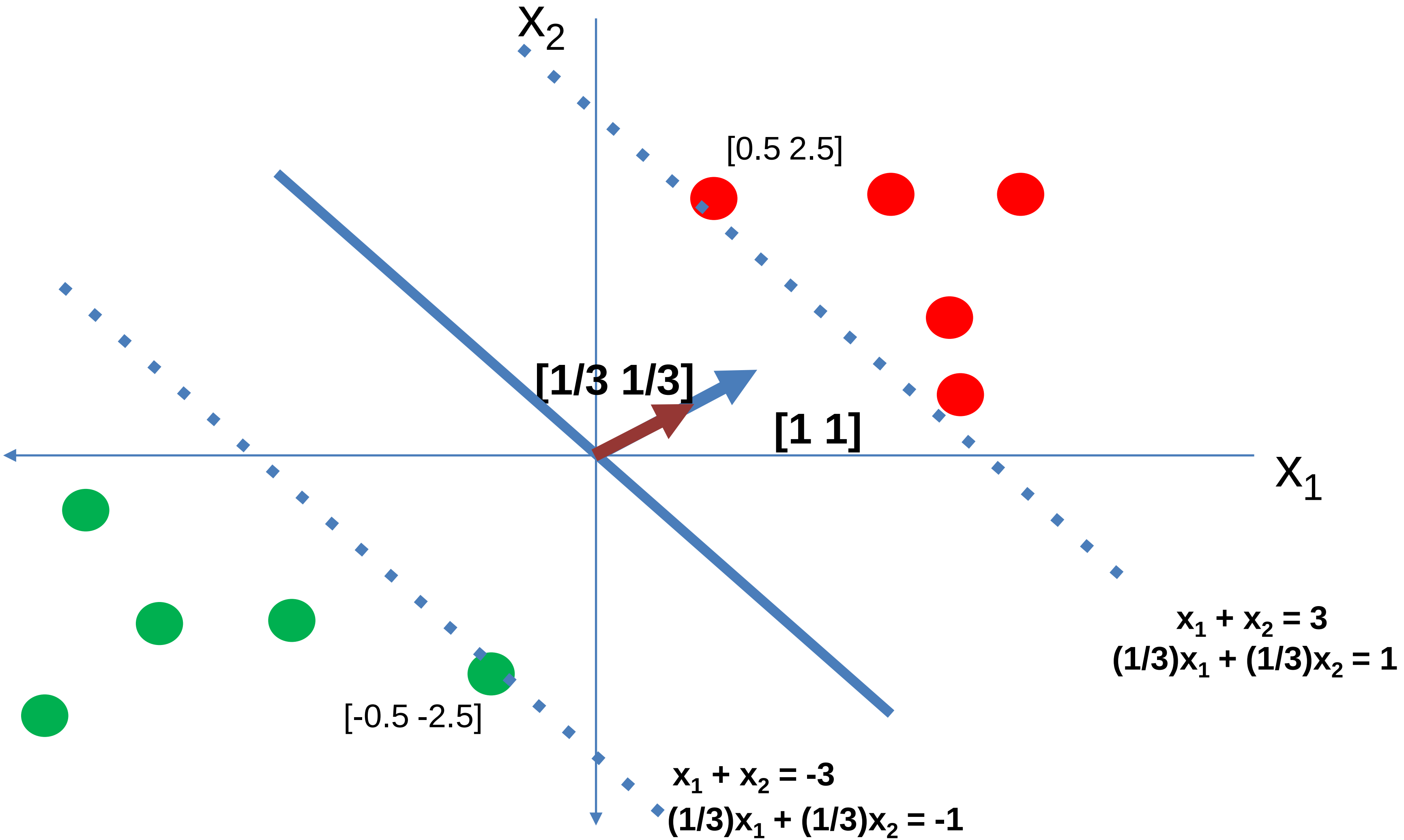
A weak learner is one which outputs a classifier
Strong

h for which

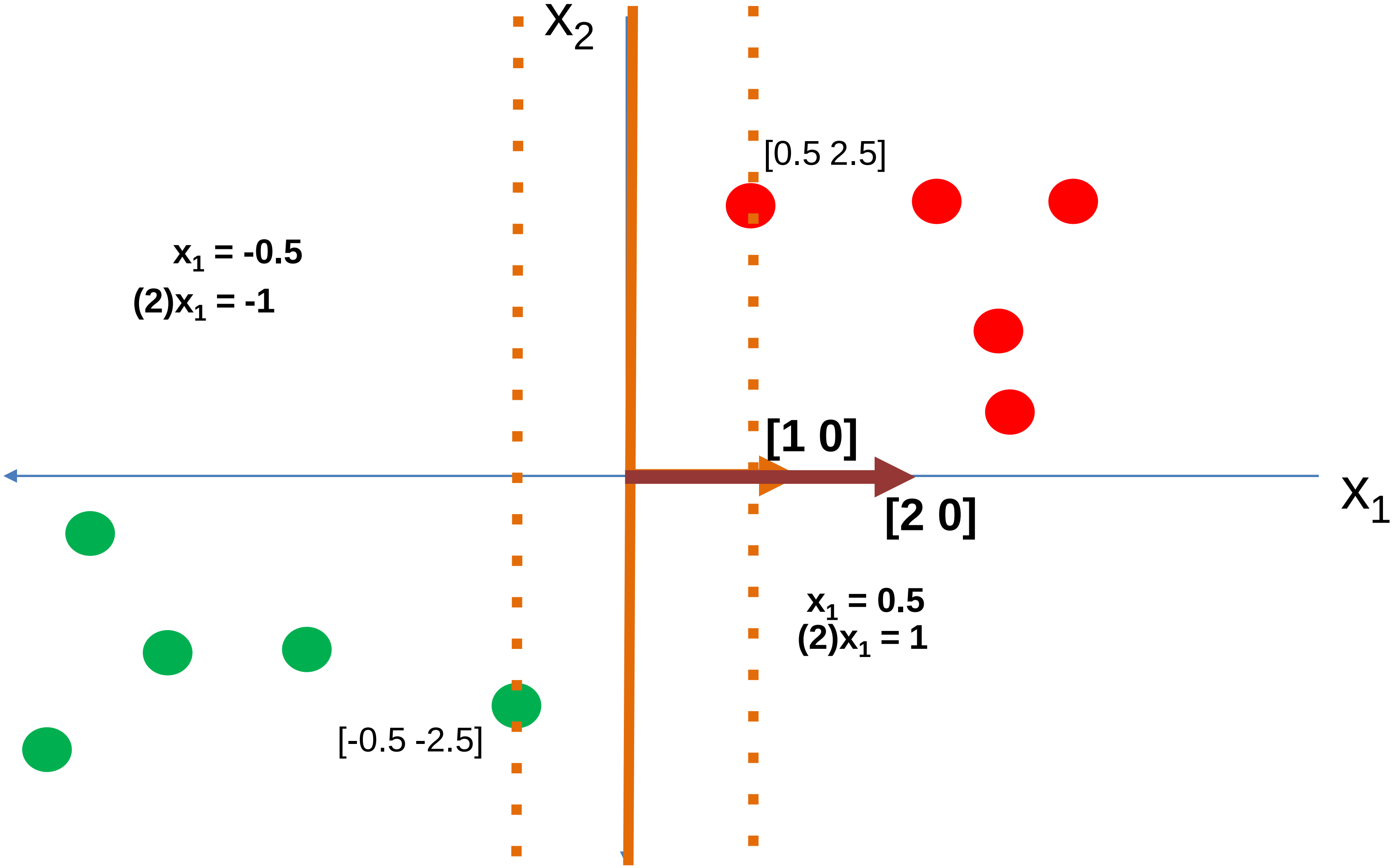
$$P_{x,y\sim D} \left(h(x) = y \right) \geq \frac{1}{2} + \gamma \quad 1 - \epsilon$$

$\boxed{\gamma > 0}$

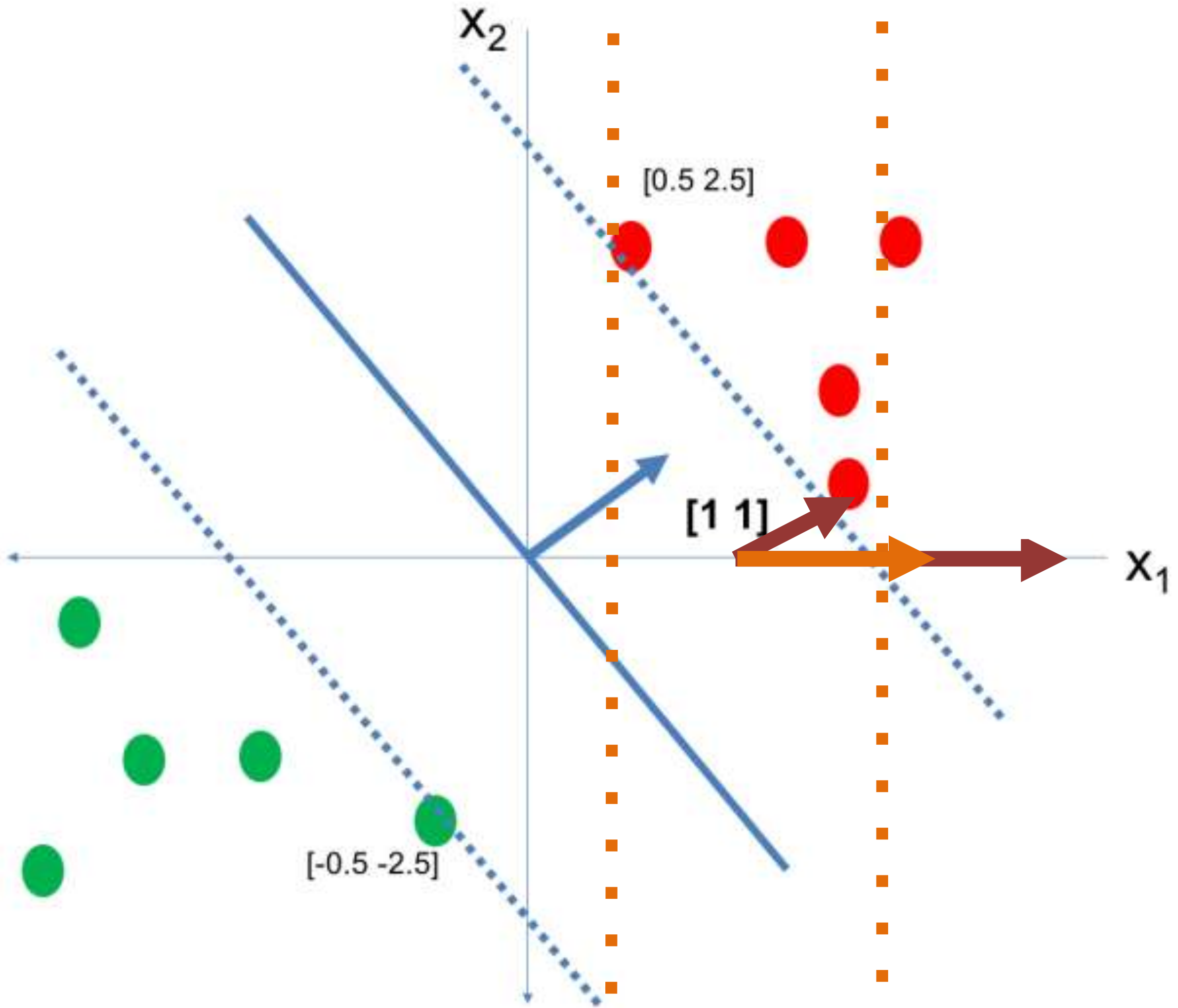
for any unknown but fixed distribution D.



X2
[0.5 2.5]
[1/3 1/3]
[1 1]
X1
x1 + x2 = 3
(1/3)x1 + (1/3)x2 = 1
[-0.5 -2.5]
x1 + x2 = -3
(1/3)x1 + (1/3)x2 = -1



X_2
[0.5 2.5]
x_1 = -0.5
(2)x_1 = -1
[1 0]
[2 0]
X_1
x_1 = 0.5
(2)x_1 = 1
[-0.5 -2.5]



Notice what happens to the lengths of the w as we adjust it to have margin 1

OBSERVATIONS

-> Once a direction is fixed,
the width between the margin lines
is fixed

-> If the width is large, then the w that
achieves margin 1 in that direction
has smaller length

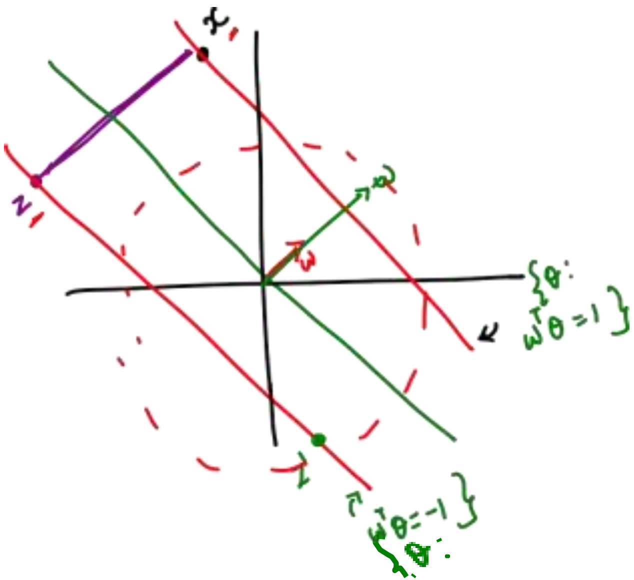
-> If the width is small, then the w that
achieves margin 1 in that direction
has larger length

-> In general, width(w) seems to be
inversely proportional to length(w)

$$\max_{w} \quad \boxed{\text{width}(w)}$$

$$\text{s.t.} \quad (w^T x_i) y_i \geq 1 \quad \forall i$$

What is width(w) ?



$$\min_{z} : \underbrace{\frac{1}{2}\|x - z\|^2}$$

$$\text{s.t.} \quad \left.\begin{array}{l} w^T x = 1 \\ w^T z = -1 \end{array}\right\}$$

[Exercise]

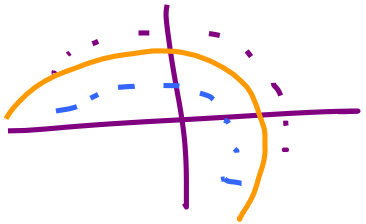
$$\text{width}(w) = \frac{2}{\|w\|^2}$$

$$\max_{w} \frac{2}{\|w\|^2}$$

$$\text{s.t } \forall i \quad (w^T x_i) y_i \geq 1$$

$$\min_{w} \frac{1}{2}\|w\|^2$$

$$\text{s.t } \forall i \quad (w^T x_i) y_i \geq 1$$



Issues

- L.S is a strong assumption
- Non-linear structure ?

DETOUR

$$\boxed{\begin{array}{l}\min_{w} \; f(w) \\ \quad g(w) \leq 0\end{array}}$$

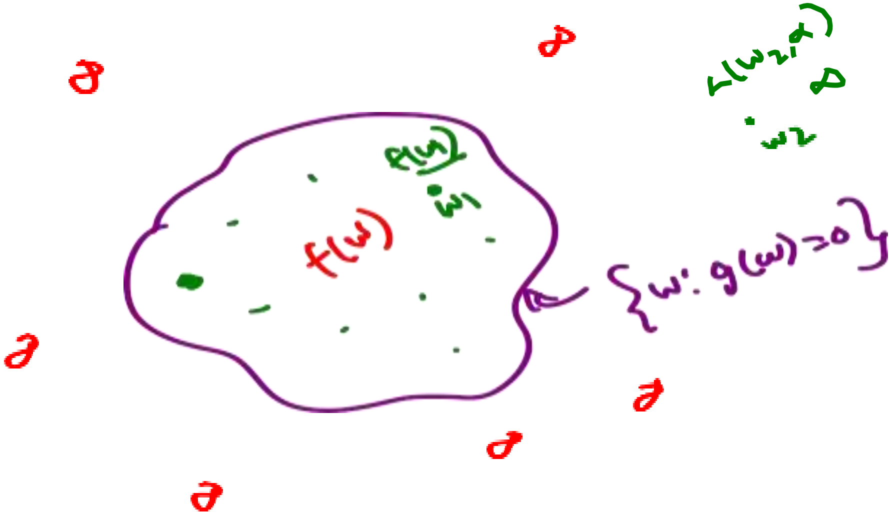
$$\mathcal{L}(w, \alpha) = f(w) + \alpha \cdot g(w)$$

Fix some w.

Consider

$$\boxed{\max_{\alpha \geq 0} \; L(w, \alpha)}$$

$$= \max_{\alpha \geq 0} \; f(w) + \alpha \, g(w)$$



$$\begin{cases} \infty & \text{if } \; g(w) > 0 \\ f(w) & \text{if } \; g(w) \leq 0 \end{cases}$$

$$\min_{w} \underbrace{\left[\max_{\alpha \geq 0} \overset{\theta(w)}{\overline{\mathcal{L}(w, \alpha)}} \right]}$$

equivalent

$$\equiv \quad \min_{w} F(w)$$

$$\text{s.t } g(w) \leq 0.$$

- Can we swap min and max?

mi
w

Multiple constraints

→ Same idea

$$\min_{w} f(w) \quad \text{s.t.} \quad g_i(w) \le 0 \quad \forall i = 1 \ldots k \quad \equiv \quad \min_{w} \left[\max_{\{\alpha_1, \ldots, \alpha_k \ge 0\}} \left[f(w) + \underbrace{\sum_{i=1}^{k} \alpha_i \underline{g_i(w)}} \right] \right]$$

$$\parallel \quad \text{Strong duality for convex } f, g_i$$

$$\max_{\alpha_1, \ldots, \alpha_k \ge 0} \min_{w} f(w) + \sum_{i=1}^{k} \alpha_i g_i(w)$$

$$\boxed{\begin{aligned} &\min_{w} \quad \underbrace{\tfrac{1}{2}\|w\|^2} \leftarrow f(w) \\ &\text{s.t} \quad \underbrace{(w^T x_i) y_i \geq 1} \quad \forall i \\ &\qquad \underbrace{1 - (w^T x_i) y_i} \leq 0 \end{aligned}}$$

$$g_i(w) = 1 - (w^T x_i) y_i$$

$$L(w, \alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right)$$

$$\alpha \in \mathbb{R}^n$$

$$\min_{w} \left[\max_{\alpha \geq 0} \frac{1}{2} \|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right) \right]$$

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} \geq 0$$

$$|||$$

$$\max_{\alpha \geq 0} \left[\min_{w} \frac{1}{2} \|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right) \right]$$

Fix $\alpha \geq 0$

$$\min_{w} \left[\underbrace{\frac{1}{2}\|w\|^2}_{} + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right) \right]$$

Grad w.r.t w

$$w^* + \sum_{i=1}^{n} -\alpha_i x_i y_i = 0$$

$\in \mathbb{R}^d$

$$\boxed{w^* = \sum_{i=1}^{n} \alpha_i x_i y_i}$$

$\{+1, -1\}$

Fixed Choice

In matrix notation

$$\boxed{w^* = X Y \alpha}$$

$$X = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix}_{d \times n} \quad Y = \begin{bmatrix} y_1 & & & \\ & \ddots & & \\ & & & y_n \end{bmatrix}_{n \times n} \quad \alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}_{n \times 1}$$

Substituting $\underset{\sim}{\text{soln}}$ back in the objective.

$$\frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right)$$

$$= \frac{1}{2} \underline{w^T w} + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i (w^T x_i) y_i$$

$$\underbrace{\frac{1}{2}(XY\alpha)^T(XY\alpha)} + \alpha^T \mathbf{1} - \underbrace{\sum_{i=1}^{n}(XY\alpha)^T x_i\, y_i\, \alpha_i}$$

$$\begin{bmatrix}\alpha_1\\ \vdots \\ \alpha_n\end{bmatrix}\begin{bmatrix}\vdots\\ \vdots\end{bmatrix}$$

on simplification [please do this]

$$\alpha^T \mathbf{1} - \frac{1}{2}(XY\alpha)^T(XY\alpha)$$

DUAL PROBLEM

$$\max_{\boxed{\alpha \geq 0}} \; \alpha^T \mathbf{1} - \frac{1}{2}\alpha^T Y^T X^T X Y \alpha$$

Solving in n instead of d.

easy constraints

can be KERNELIZED!

$n \times n$.

Revisiting the Lagrangian

$$\underbrace{\min_{w} \left[\overbrace{\max_{\alpha \geq 0} f(w) + \alpha g(w)} \right]}_{\text{PRIMAL}} = \underbrace{\max_{\alpha \geq 0} \left[\min_{w} f(w) + \alpha g(w) \right]}_{\text{DUAL}}$$

w^* $\qquad\qquad$ α^*

$$\boxed{\underbrace{\max_{\alpha \geq 0} f(w^*) + \alpha\, g(w^*)}} = \min_{w} f(w) + \alpha^* g(w)$$

$$f(w^*) = f(w') + \alpha^* g(w')$$

$$f(w^*) \leq f(w^*) + \alpha^* g(w^*)$$

$$\Rightarrow \boxed{\alpha^* g(w^*) \geq 0}$$

But we know $\alpha^* g(w^*) \leq 0$

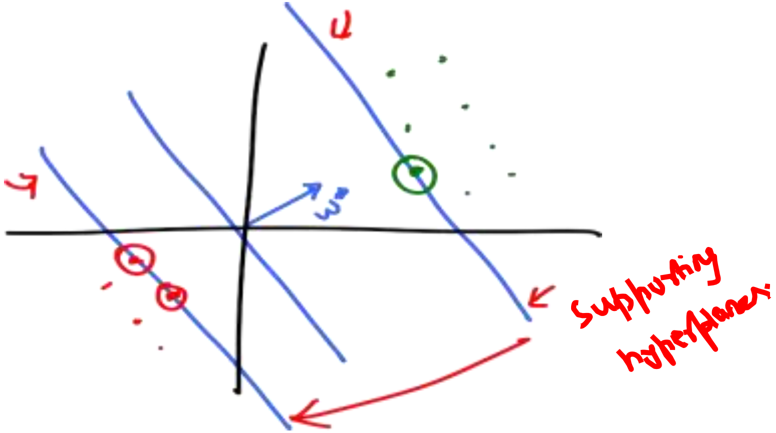
$$\Rightarrow \boxed{\alpha^* g(w^*) = 0} \rightarrow \text{COMPLEMENTARY SLACKNESS}$$

For multiple constraints

$$\alpha_i^* \, g_i(w^*) = 0 \quad \forall i$$

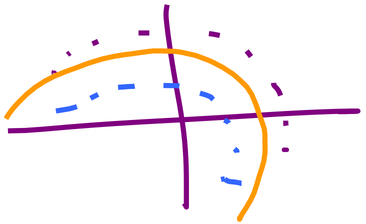
For our problem

$$\boxed{\alpha_i^* \left(1 - (w^{*T} x_i) y_i\right) = 0 \quad \forall i}$$



$$\min_{w} \quad \frac{1}{2}\|w\|^2$$

$$\text{s.t.} \ \forall i \quad (w^T x_i) y_i \geq 1$$



Issues

- L.S is a strong assumption

- Non-linear structure?

So far

Support Vector Machines

Primal Problem – Margin Maximization

Dual Problem

- Kernel Version

Now

- What if there are **outliers** in the problem?

Idea (to deal with outliers):

Fix any w. w classifies some points correct and some incorrectly. Let the incorrect points pay "bribe" to get to the correct side.

Modified formulation

$C > 0$ [hyper parameter]

$$\min_{\boxed{w, \xi}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

$\rightarrow (w^T x_i) y_i + \xi_i \geq 1 \quad \leftarrow \forall i$

$\rightarrow \xi_i \geq 0 \quad \leftarrow \forall i$

If $C = 0 \Rightarrow$ Bribes don't cost $\Rightarrow$ $w = 0$ is solution

$C \rightarrow \infty \Rightarrow$ Bribes are too costly $\Rightarrow$ Linear separable case.

$$\mathcal{L}(\omega, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + \underbrace{c\left(\sum_{i=1}^{n} \xi_i\right)}_{\uparrow} + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i - \xi_i\right)$$

$$+ \sum_{i=1}^{n} \beta_i (-\xi_i)$$

Dual:

$$\max_{\substack{\alpha \geq 0 \\ \beta \geq 0}} \min_{\omega} \mathcal{L}(w, \xi, \alpha, \beta)$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w^* = \sum_{i=1}^{n} \alpha_i x_i y_i$$

$$\boxed{w^* = X Y \alpha}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \quad \Rightarrow \quad \boxed{C - \alpha_i - \beta_i = 0}$$

$$\boxed{\alpha_i + \beta_i = C} \quad \forall i$$

Substitute $w^* = XY\alpha$ in the original objective

$$\frac{1}{2}(XY\alpha)^T(XY\alpha) + \sum_{i=1}^{n} (C - \underline{\alpha_i - \beta_i})\xi_i + \alpha^T 1 - (XY\alpha)^T(XY\alpha)$$

SOFT - MARGIN

SUPPORT

VECTOR

MACHINE

$$\max_{\substack{\alpha \geq 0 \\ \beta \geq 0 \\ \alpha_i + \beta_i = C}} \quad \alpha^T 1 - \frac{1}{2} (x y \alpha)^T (x y \alpha)$$

$$\equiv \quad \boxed{\max_{\underbrace{0 \leq \alpha \leq C}_{\text{BOX CONSTRAINT.}}} \quad \underbrace{\alpha^T 1 - \frac{1}{2} \alpha^T y^T (x^T x) y \alpha}}$$

HARD-MARGIN SVM

PRIMAL

$$\min_{w} \frac{1}{2}\|w\|^2$$

$$\text{s.t.} \quad \underbrace{(w^T x_i) y_i \geq 1}_{1 - w^T x_i y_i \leq 0} \quad \forall i$$

DUAL

$$\max_{\alpha \geq 0} \; \alpha^T 1 - \alpha^T y^T x^T x \, y \alpha$$

$$\alpha_i^*(1 - w^{*T} x_i y_i) = 0 \quad \forall i$$

$$\boxed{w^* = \sum_{i=1}^{n} \alpha_i^* x_i y_i}$$

SOFT-MARGIN SVM

PRIMAL ✓

$$\min_{w, \xi} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

$$\text{s.t.} \quad (w^T x_i) y_i + \xi_i \geq 1 \quad \forall i = 1, \ldots n \quad (\alpha)$$

$$\xi_i \geq 0 \quad \forall i = 1, \ldots n \quad (\beta)$$

DUAL ✓

$$\max \; \alpha^T 1 - \alpha^T y^T x^T x \, y \alpha$$

$$\boxed{\alpha + \beta = C}$$

$$\underbrace{\alpha \geq 0, \quad \beta \geq 0}_{0 \leq \alpha \leq C}$$

- Let $(\underline{w}^*, \underline{\xi}^*)$ be the primal optimal solution
- Let $(\underline{\alpha}^*, \underline{\beta}^*)$ be the dual optimal solution

COMPLEMENTARY SLACKNESS

$$\underline{\alpha_i^*}\left(\underline{1 - ({w^*}^T x_i) y_i - \xi_i^*}\right) = 0 \quad \forall i$$

$$\beta_i^* \xi_i^* = 0 \quad \forall i$$

$\alpha_i^* + \beta_i^* = C$
$\forall i$
$\hookrightarrow$ (A)

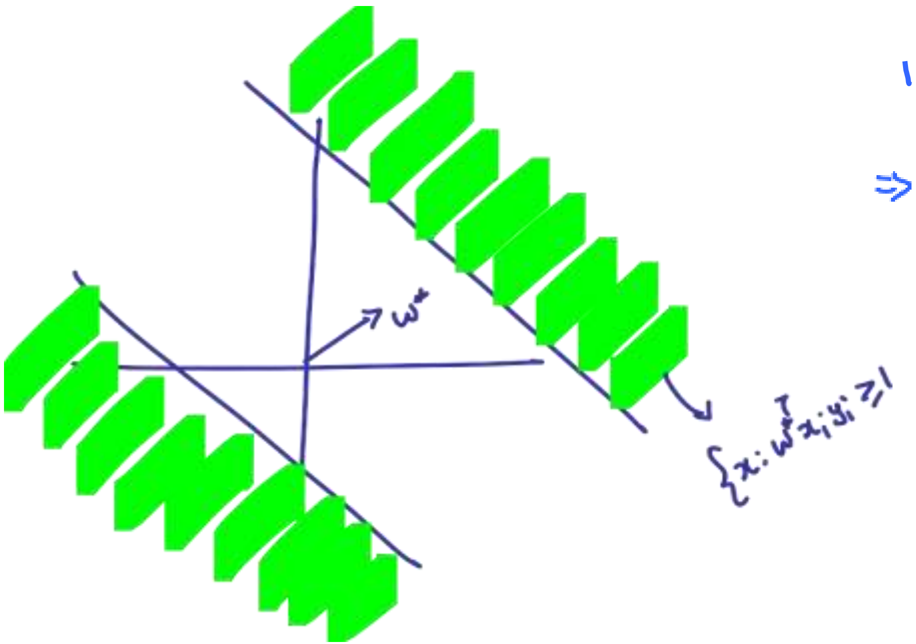
Various cases possible

Case 1: $\alpha_i^* = 0 \overset{4}{\Rightarrow} \beta_i^* = C \overset{\boxed{CS}}{\Rightarrow} \underline{\xi_i^*} = 0$

$1 - (w^{*T} x_i) y_i - \underbrace{\xi_i^*}_{=0} \leq 0$ [Primal feasibility]

$\Rightarrow \quad 1 - (w^{*T} x_i) y_i \leq 0$

$\Rightarrow \quad w^{*T} x_i y_i \geq 1$



$\Rightarrow$ w^* classifies (x_i, y_i) correctly.

Case 2: $0 < \alpha_i^* < C \overset{(A)}{\Longrightarrow} 0 < \beta_i^* < C \overset{CS}{\Longrightarrow} \underline{\xi_i^*} = 0$

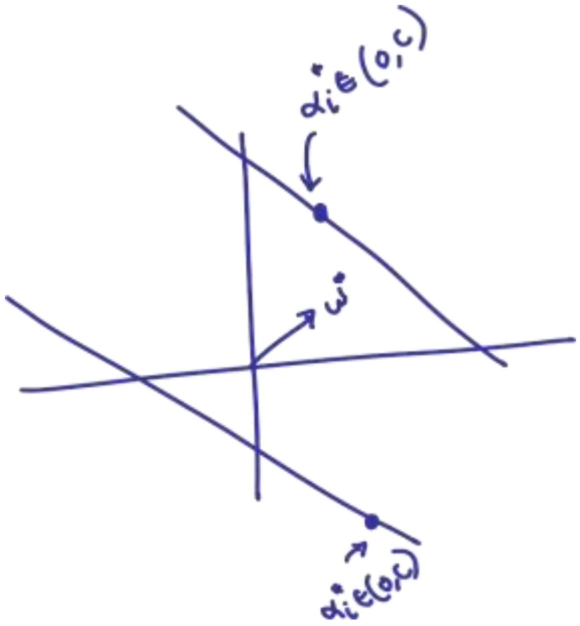
$$\Big\Downarrow \boxed{CS}$$

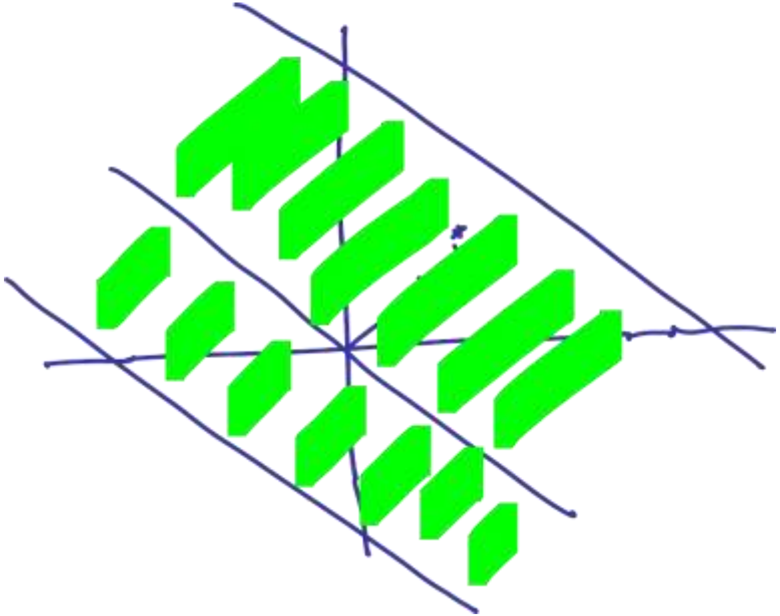
$$1 - (w^{*T} x_i) y_i - \xi_i^* = 0$$

$$\Big\Downarrow$$

$$(w^{*T} x_i) y_i = 1$$

$\Rightarrow$ (x_i, y_i) lies on the supporting hyperplane.





Case 3:

$$\alpha_i^* = C \Rightarrow \beta_i^* = 0 \Rightarrow \xi_i^* \geq 0$$

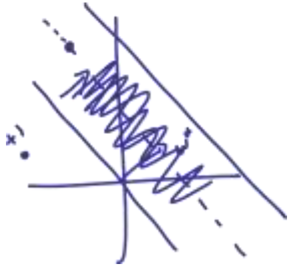
$$\Downarrow \boxed{CS}$$

$$1 - w^{*T} x_i y_i - \xi_i^* = 0$$

$$\xi_i^* = 1 - w^{*T} x_i y_i \geq 0$$

$$\Rightarrow \boxed{w^{*T} x_i y_i \leq 1}$$

Let's see this from P.o.v of data



CASE 1

$$\boxed{\underline{w^{*}x_i y_i} \ \underline{< 1}}$$

$$1 - w^{T}x_i y_i - \xi_i^{*} \leq 0$$

$$w^{*}x_i y_i \geq 1 - \xi_i^{*}$$

$$\boxed{\xi_i^{*} \geq 1 - \underline{w^{*}x_i y_i}}$$

$$\Rightarrow \xi_i^{*} > 0 \Rightarrow \beta_i^{*} = 0 \Rightarrow \underline{\alpha_i^{*} = C}$$

$$\underline{\alpha_i^{*}}\left(\underline{1 - w^{T}x_i y_i - \xi_i^{*}}\right) = 0$$

$$\beta_i^{*}\xi_i^{*} = 0.$$

CASE 2 : $w^{*T} x_i y_i = 1$

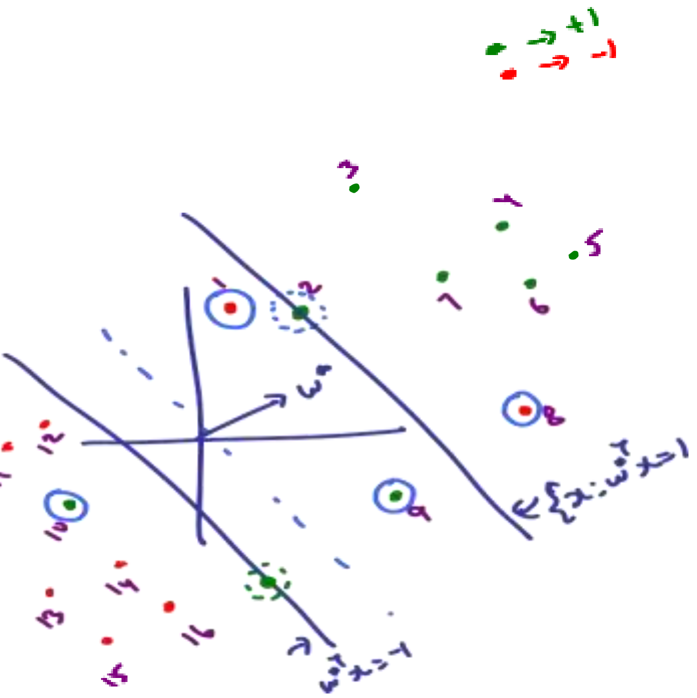
$$\xi_i^* \geq 1 - \underline{w^{*T} x_i y_i}$$

$$\Rightarrow \quad \xi_i^* \geq 0 \quad \Rightarrow \quad \alpha_i^* \in [0, C]$$

CASE 3 $w^{*T} x_i y_i > 1$

$$1 - w^{*T} x_i y_i - \xi_i^* \leq 0 \quad \text{[Primal feasibility]}$$

$$\Rightarrow \quad 1 - w^{*T} x_i y_i - \xi_i^* < 0 \quad \overset{\boxed{\text{C.S}}}{\Longrightarrow} \quad \alpha_i^* = 0$$



SUMMARY

$$\alpha_i^* = 0 \Rightarrow w^{*T} x_i y_i \geq 1$$

$$0 < \alpha_i^* < C \Rightarrow w^{*T} x_i y_i = 1$$

$$\alpha_i^* = C \Rightarrow w^{*T} x_i y_i \leq 1$$

✓ $w^{*T} x_i y_i < 1 \Rightarrow \alpha_i^* = C$

→ $w^{*T} x_i y_i = 1 \Rightarrow \alpha_i^* \in [0, C]$

→ $w^{*T} x_i y_i > 1 \Rightarrow \alpha_i^* = 0.$

Binary classification

GENERATIVE

- Naive Bayes
- G.D.A

DISCRIMINATIVE

- K-NN
- Decision trees
- Perceptron
- Support-vector-machines

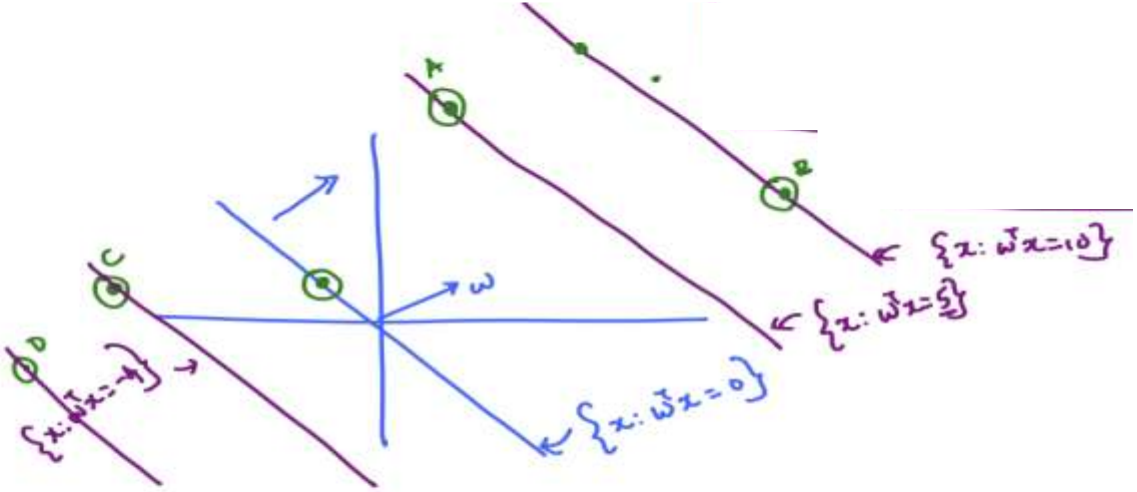
Dosen't "really" model $\boxed{P(y/x)}$

→ Just finds $f: \mathbb{R}^d \to \{\pm 1\}$

- Can we model $P(y=+1/x)$ differently?

Start with a simple model

Given $x \in \mathbb{R}^d$ $\quad z = w^T x \quad w \in \mathbb{R}^d$.



$$\boxed{P\left(y=+1/x\right)} = g(w^T x)$$

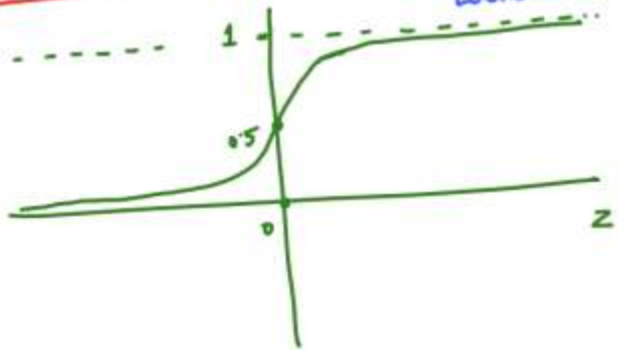
- $g(z) \in [0,1]$
- $g(z) \to \boxed{1}$ as $z \to \boxed{\infty}$
- $g(z) \to 0$ as $z \to -\infty$
- $g(z) = 0.5$ if $z = 0$.

LINK FUNCTION

ONE POPULAR CHOICE

$$g(z) = \frac{1}{1+e^{-z}}$$

SIGMOID FUNCTION
LOGISTIC FUNCTION.



MODEL: LOGISTIC REGRESSION

Data: $\{ (x_1, y_1) \cdots (x_n, y_n) \}$ $\quad x_i \in \mathbb{R}^d$

$y_i \in \{0, 1\}$

Max. Likelihood

$$\mathcal{L}(w, \text{Data}) = \prod_{i=1}^{n} \left(g(w^T x_i)\right)^{y_i} \left(1 - g(w^T x_i)\right)^{(1-y_i)}$$

$$\log \mathcal{L}(w, \text{Data}) = \sum_{i=1}^{n} y_i \log\left(g(\underline{w^T x_i})\right) + (1-y_i) \log\left(1 - g(w^T x_i)\right)$$

$$= \sum_{i=1}^{n} \left[y_i \log\left(\frac{1}{1+e^{-w^T x_i}} \right) + (1-y_i) \log\left(\frac{e^{-w^T x_i} \times 1}{1+e^{-w^T x_i}} \right) \right]$$

$$= \sum_{i=1}^{n} \left[\log\left(\frac{e^{-w^T x_i}}{1+e^{-w^T x_i}} \right) - \underline{y_i(-w^T x_i)} \right]$$

$$= \sum_{i=1}^{n} \left[(1-y_i)(-w^T x_i) - \log\left(1+e^{-w^T x_i}\right) \right]$$

- <u>No closed form solution</u>

• Gradient ascent

$$\nabla \log L(w) = \sum_{i=1}^{n} (1-y_i)(-x_i) - \frac{e^{-w^T x_i}}{1+e^{-w^T x_i}}(-x_i)$$

$$= \sum_{i=1}^{n} x_i \left(y_i - \left(1 - \frac{e^{-w^T x_i}}{1+e^{-w^T x_i}}\right)\right)$$

i = 1

$$= \boxed{\sum_{i=1}^{n} \overset{\mathbb{R}^d}{x_i} \left(\overset{\{0,1\}}{y_i} - \overset{g(w^T x_i)}{\frac{1}{1+e^{-w^T x_i}}} \right)}$$

$$\boxed{w_{t+1} = w_t + \eta_t \nabla \log L(w_t)}$$

REGULARIZED VERSION

$$\min_{w} \sum_{i=1}^{n} (1-y_i) w^T x_i + \log\left(1+e^{-w^T x_i}\right) + \underbrace{\frac{\lambda}{2} \|w\|^2}$$

KERNEL VERSION

- Can argue $w = \underbrace{\sum_{i=1}^{n} \alpha_i x_i}$ [Formal Theorem Representer Theorem]

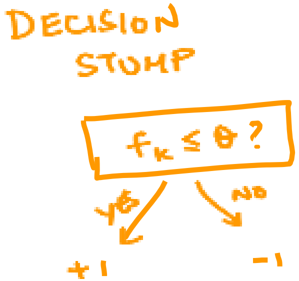
Exercise: Derive the kernel version of logistic regression

META CLASSIFIERS (or)

ENSEMBLE CLASSIFIERS.

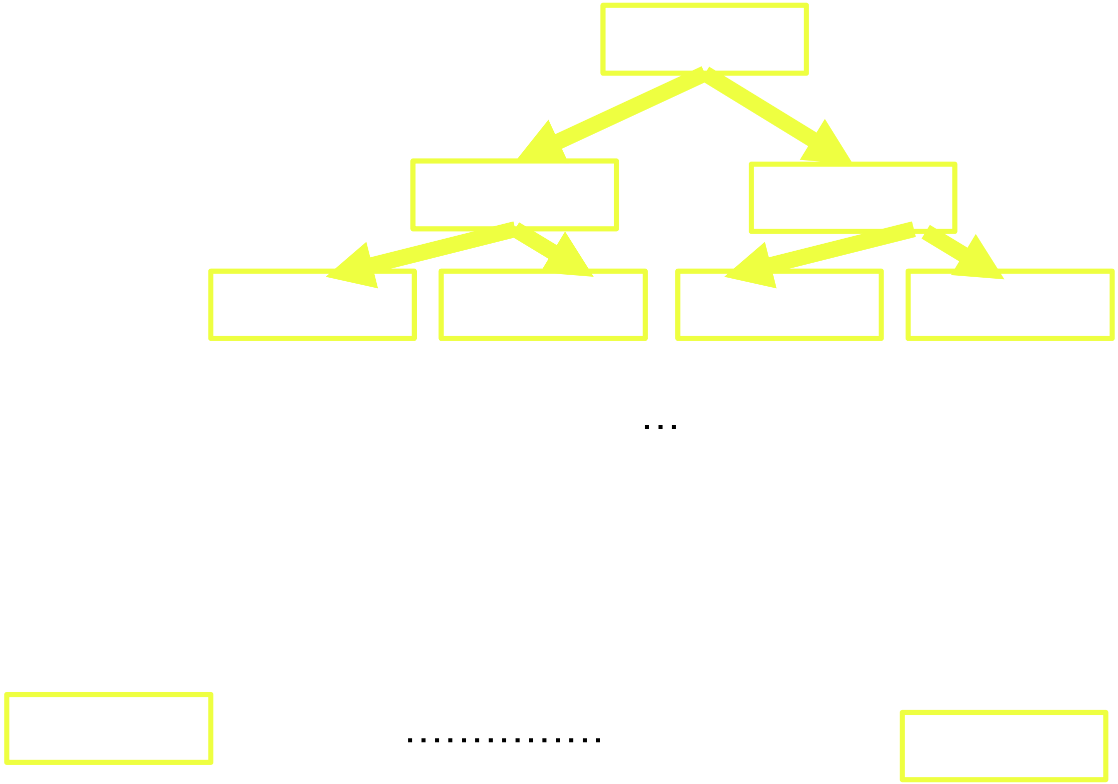
WEAK CLASSIFIERS [better than random] $\longrightarrow$ STRONG CLASSIFIERS

Weak classifiers



high bias, low variance

Overfit decision tree



low bias, high variance

$$x_1, x_2, \ldots x_n \sim N(\mu, 1)$$

$$\hat{\mu}_1 = x_1 \qquad \hat{\mu}_2 = x_2, \ldots \hat{\mu}_N = x_n \qquad \hat{\mu}_{ML} = \frac{1}{n} \sum x_i$$

$$\boxed{D_1}, \boxed{D_2}, \ldots, \boxed{D_k}$$

Overfit decision trees

$$h_1 \quad h_2 \quad \ldots \quad h_k$$

$$h_i : \mathbb{R}^d \to \{\pm 1\}$$

$$h^*(x) = \text{majority}(h_1(x), \ldots, h_k(x))$$

BAGGING - Bootstrap Aggregation.

$D = \{(x_1, y_1) \ldots (x_n, y_n)\}$

- Create datasets $D_1, \ldots, D_k$ from D by "sampling with replacement".
- Run weak classifier on $D_1, \ldots, D_k$ to get $h_1, \ldots, h_k$
- Aggregate $h_1, \ldots, h_k$ using majority.

Chance that a point appears in a dataset

$1 - \underbrace{\left(1-\frac{1}{n}\right)\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{1}{n}\right)}$

$1 - \underbrace{\left(1-\frac{1}{n}\right)^n}$

$1 - \frac{1}{e}$ (as $n \to \infty$)

$\simeq 66\%$

FEATURE BAGGING → Bag the features in addition to data points

Feature bagged decision trees -> RANDOM FOREST

BOOTSTRAP - Sampling with Replacement ?

AGGREGATION - Majority. ?

BOOSTING

ADA-BOOST

[Freund & Schappire.
1995
Godel Prize]

Distribution D over $(\mathcal{X} \times \mathcal{Y})$ $\mathbb{R}^d$ $\{+1,-1\}$

unknown but fixed.

$x_1, \dots, x_n$ are iid from D.

$h : \underset{\mathcal{X}}{\mathbb{R}^d} \to \underset{\mathcal{Y}}{\{\pm 1\}}$.

Measure performance using

$$P_{(x,y)\sim D}\left(h(x) \neq y\right)$$

Misclassification probability.

A weak learner is one which outputs a classifier
Strong

h for which

$$P_{x,y\sim D}\left(h(x) = y\right) \geq \frac{1}{2} + \gamma \qquad \boxed{\gamma > 0}$$

$1 - \epsilon$

for any unknown but fixed distribution D.