SVMs

Two different approaches to regression/classification

Generative

- Assume something about P(x,y) $P(y|x)$
- Find f which maximizes likelihood of training data assumption
  - Often reformulated as minimizing loss

Versus

Discriminative

- Pick a loss function
- Pick a set of hypotheses H linear & NN ..
- Pick f from H which minimizes loss on training data

Our description of logistic regression was the former

- **Learn**: f:**X —>**Y
  - **X** – features
  - **Y – target classes**

$$Y \in \{-1, 1\}$$

- **Expected loss of f:**

$$\mathbb{E}_{XY}[\mathbf{1}\{f(X) \neq Y\}] = \mathbb{E}_X[\mathbb{E}_{Y|X}[\mathbf{1}\{f(x) \neq Y\}|X = x]]$$

$$\mathbb{E}_{Y|X}[\mathbf{1}\{f(x) \neq Y\}|X = x] = 1 - P(Y = f(x)|X = x)$$

- **Bayes optimal classifier:**

$$f(x) = \arg\max_y \mathbb{P}(Y = y|X = x)$$

- **Model of logistic regression:**

$$P(Y = y|x, w) = \frac{1}{1 + \exp(-y\, w^T x)}$$

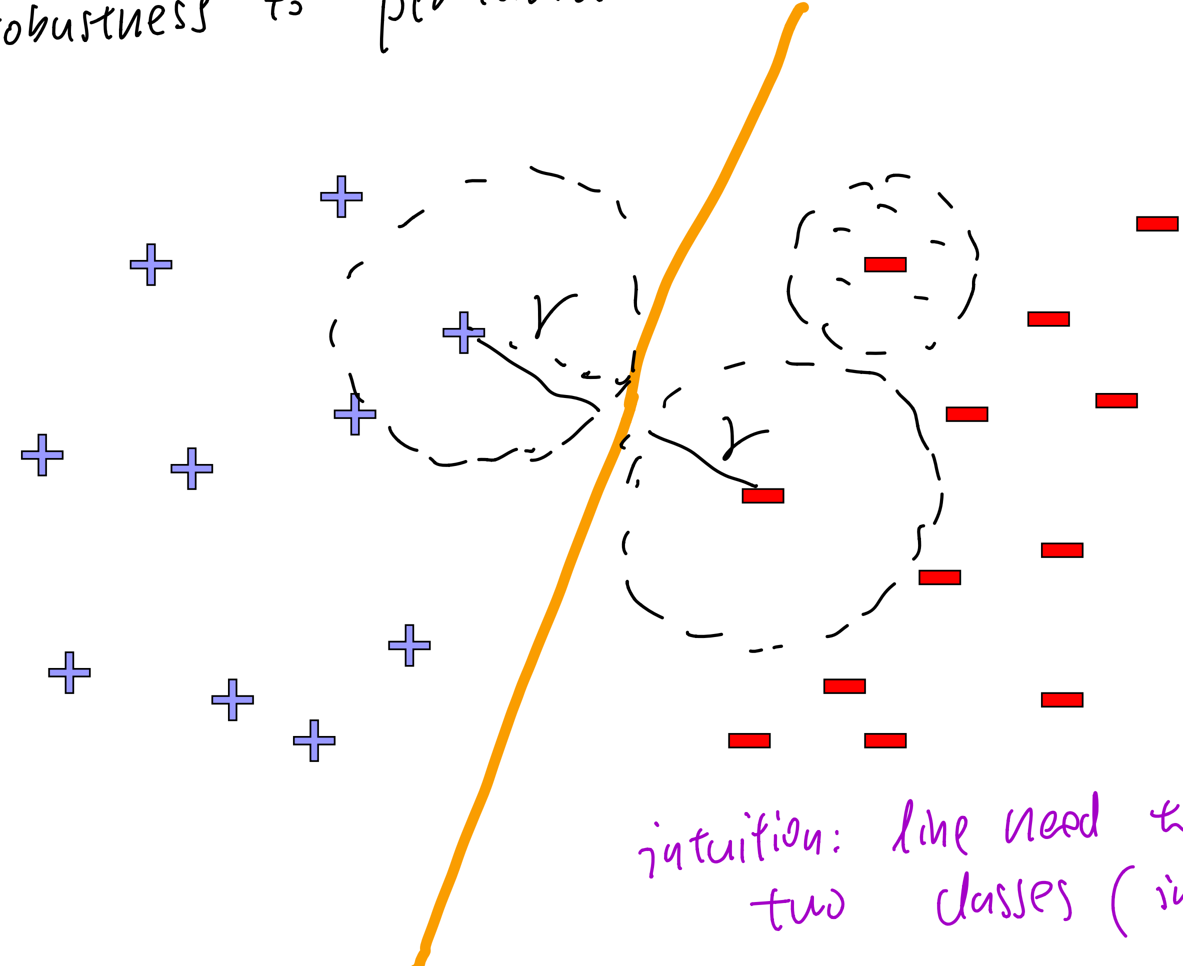
- **Loss function:**

$$\ell(f(x), y) = \mathbf{1}\{f(x) \neq y\}$$

What if the model is wrong? What other ways can we pick linear decision rules?

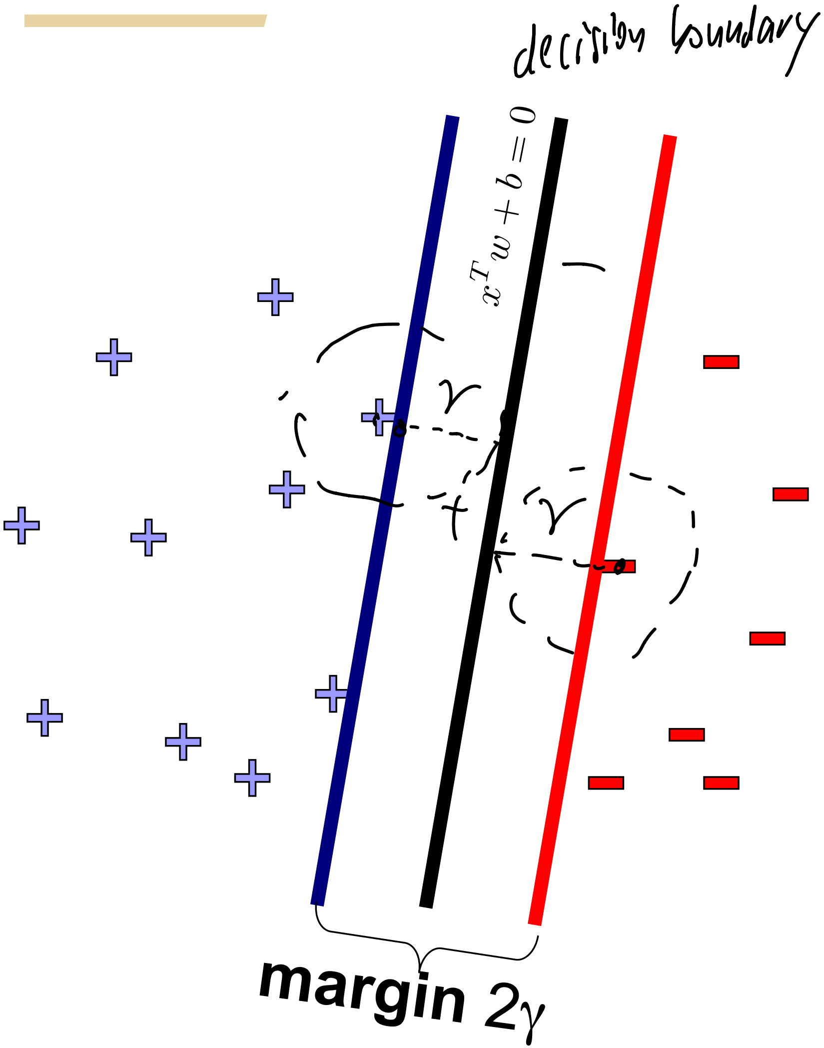
Linear classifiers – Which line is better?

robustness to perturbation



intuition: line need to balance two classes (in the middle)

Pick the one with the largest margin!



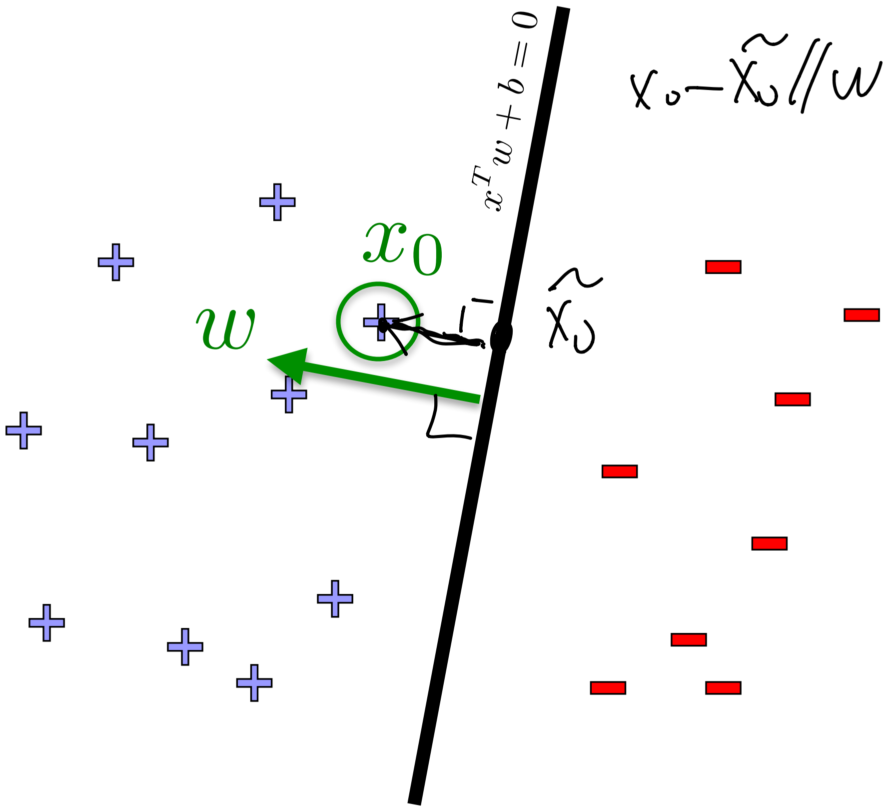
γ: margin

the minimum distance from training point to decision boundary

Goal: (w, b)

find decision boundary with maximum γ

Pick the one with the largest margin!



$\tilde{x}_0^T w + b = 0$ ($\tilde{x}_0$ is on decision boundary)

Distance from x_0 to hyperplane defined by $x^T w + b = 0$?

distance:

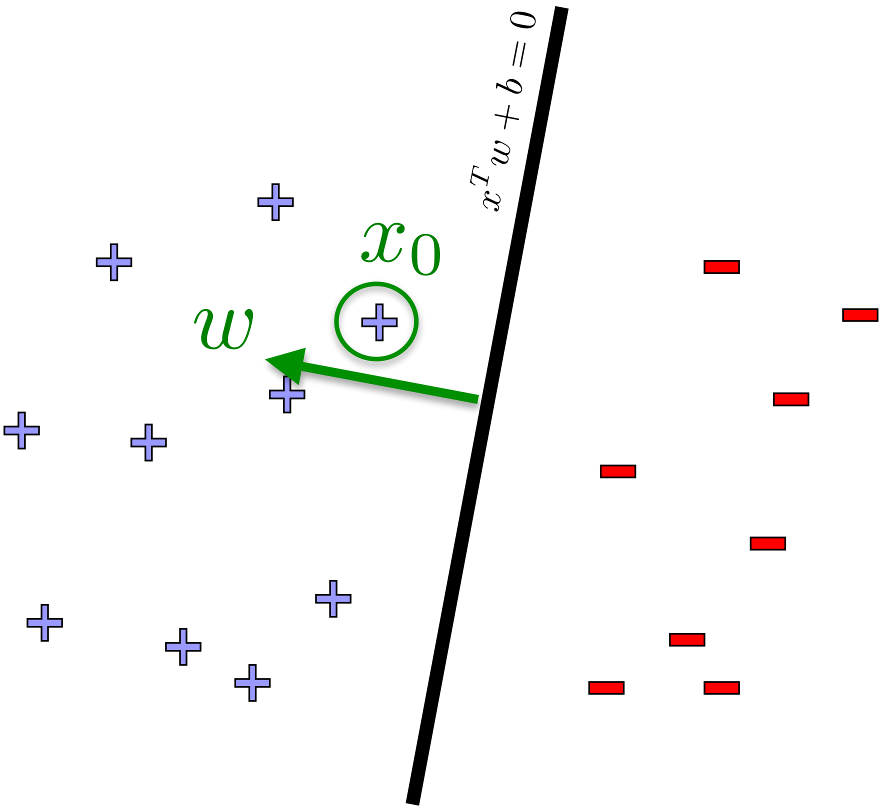
$\|x_0 - \tilde{x}_0\|_2$

$= \left| (x_0 - \tilde{x}_0)^T \frac{w}{\|w\|_2} \right|$ (projection on its direction)

$= \frac{1}{\|w\|_2} \left| x_0^T w - \tilde{x}_0^T w \right|$

$= \frac{1}{\|w\|_2} \left| x_0^T w + b \right|$

Pick the one with the largest margin!



Distance from x_0 to hyperplane defined by $x^T w + b = 0$?

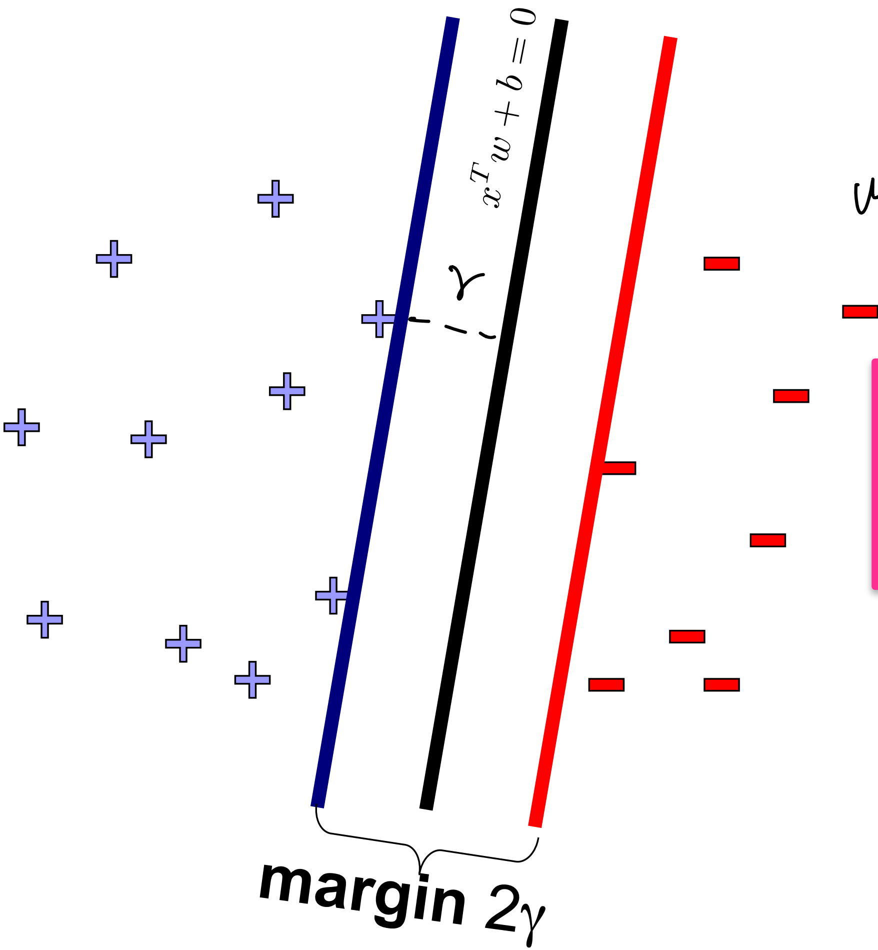
If $\widetilde{x}_0$ is the projection of x_0 onto the hyperplane then

$$||x_0 - \widetilde{x}_0||_2 = |(x_0^T - \widetilde{x}_0)^T \frac{w}{||w||_2}|$$

$$= \frac{1}{||w||_2} |x_0^T w - \widetilde{x}_0^T w|$$

$$= \underbrace{\frac{1}{||w||_2}}_{\text{normalization}} \underbrace{|x_0^T w + b|}_{\text{decision rule}}$$

Pick the one with the largest margin!



Distance of x_0 from hyperplane $x^T w + b$:

want $$\frac{1}{||w||_2}(x_0^T w + b) \geq \gamma$$

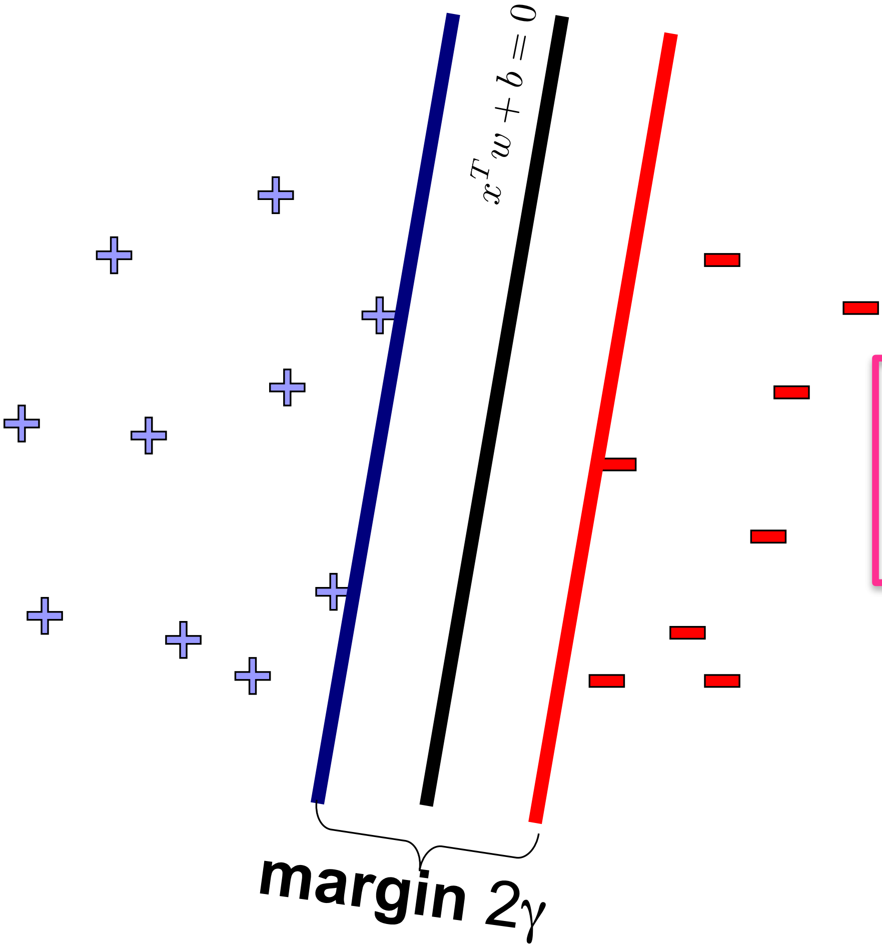
Optimal Hyperplane

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

subject to: $\frac{1}{||w||_2} y_i(x_i^T w + b) \geq \gamma, \forall i$

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

Pick the one with the largest margin!



Distance of x_0 from hyperplane $x^T w + b$:

$$\frac{1}{||w||_2}(x_0^T w + b)$$

Optimal Hyperplane

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

$$\text{subject to } \frac{1}{||w||_2} y_i(x_i^T w + b) \geq \gamma \quad \forall i$$

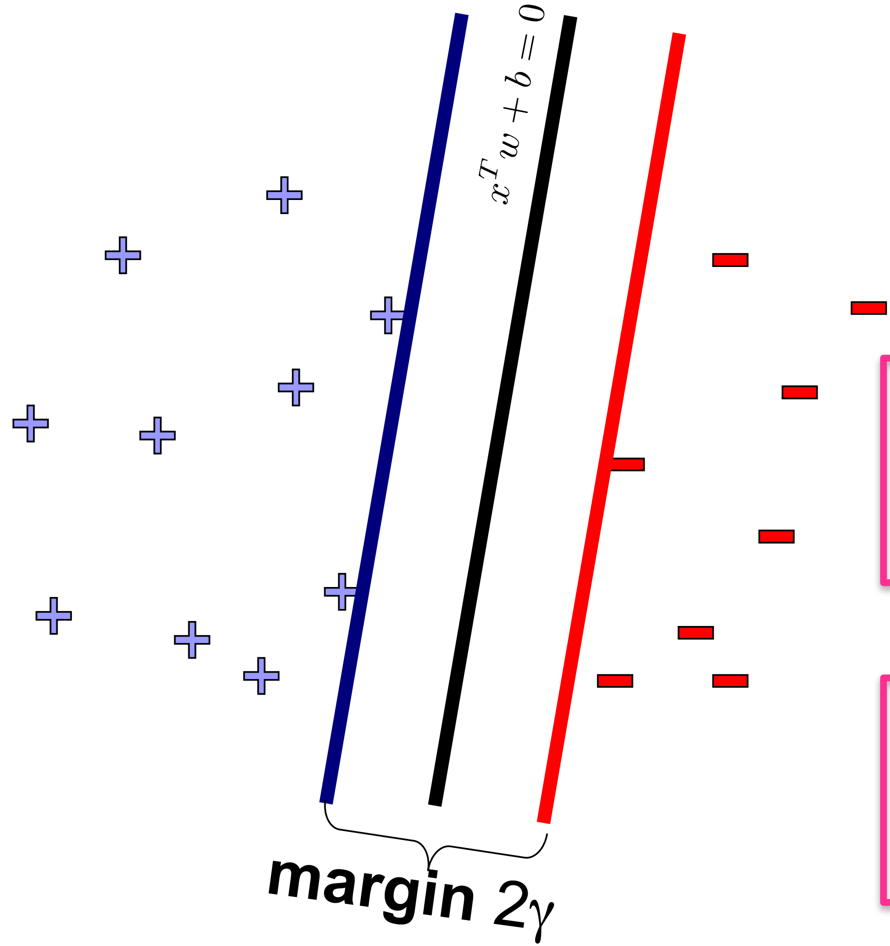
change of variable:

let $\tilde{w} = \frac{w}{\|w\|_2 \gamma}$ (1)

$\tilde{b} = \frac{b}{\|w\|_2 \cdot \gamma}$

(1) $\Rightarrow \|\tilde{w}\|_2 = \left\|\frac{w}{\|w\|_2}\right\| \cdot \frac{1}{\gamma} \Rightarrow \gamma = \frac{1}{\|\tilde{w}\|_2}$

Pick the one with the largest margin!



Distance of x_0 from hyperplane $x^T w + b$:

$$\frac{1}{\|w\|_2}(x_0^T w + b)$$

Optimal Hyperplane

$$\max_{w,b} \gamma \qquad \text{non-convex}$$

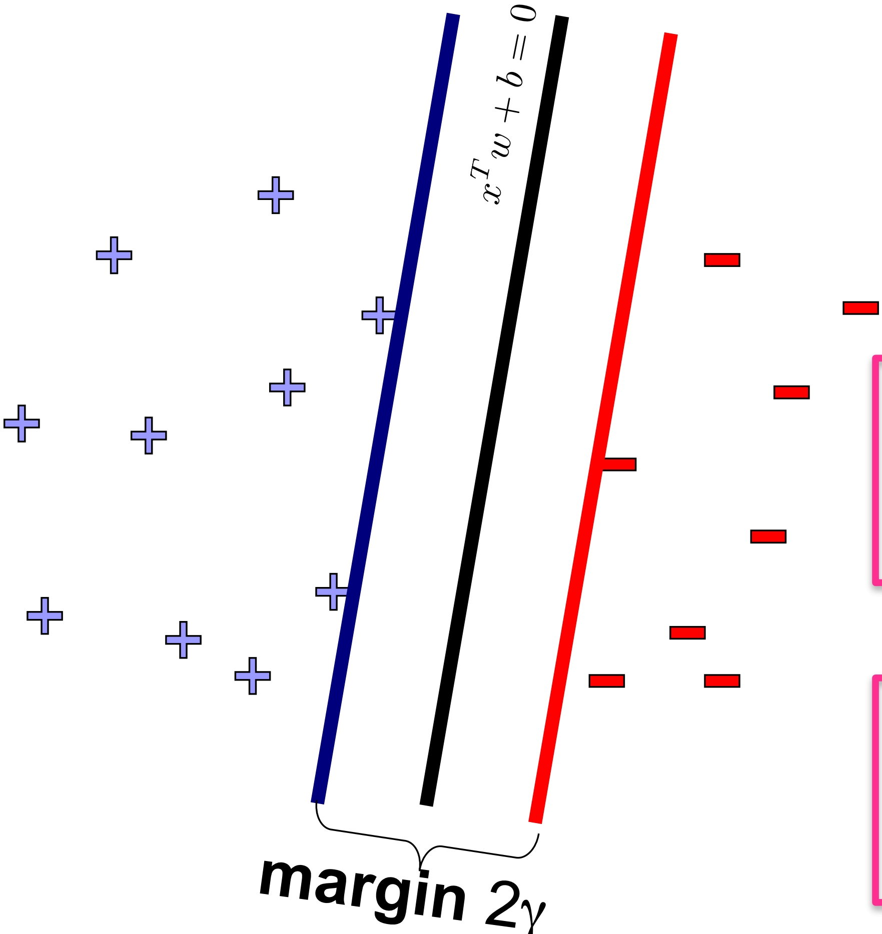
$$\text{subject to } \frac{1}{\gamma \|w\|_2} y_i (x_i^T w + b) \geq \gamma \; 1 \; \forall i$$

Optimal Hyperplane (reparameterized)

$$\max_{\tilde{w}, \tilde{b}} \frac{1}{\|\tilde{w}\|_2} \iff \min_{\tilde{w}, \tilde{b}} \|\tilde{w}\|_2^2$$

$$\text{subject } y_i (x_i^T \tilde{w} + \tilde{b}) \geq 1 \quad \forall i$$

Pick the one with the largest margin!



Distance of x_0 from hyperplane $x^T w + b$:

$$\frac{1}{||w||_2}(x_0^T w + b)$$

Optimal Hyperplane

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

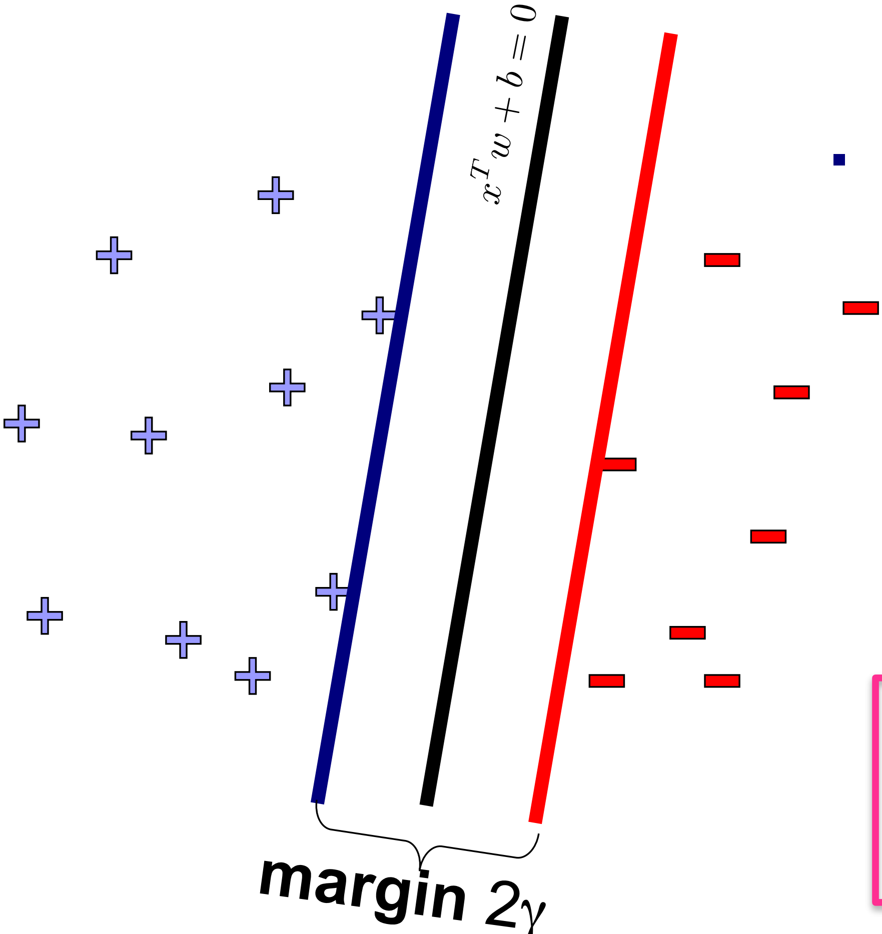
$$\text{subject to } \frac{1}{||w||_2} y_i(x_i^T w + b) \geq \gamma \quad \forall i$$

Optimal Hyperplane (reparameterized)

$$\min_{w,b} ||w||_2^2$$ convex

$$\text{subject to } y_i(x_i^T w + b) \geq 1 \quad \forall i$$

Pick the one with the largest margin!

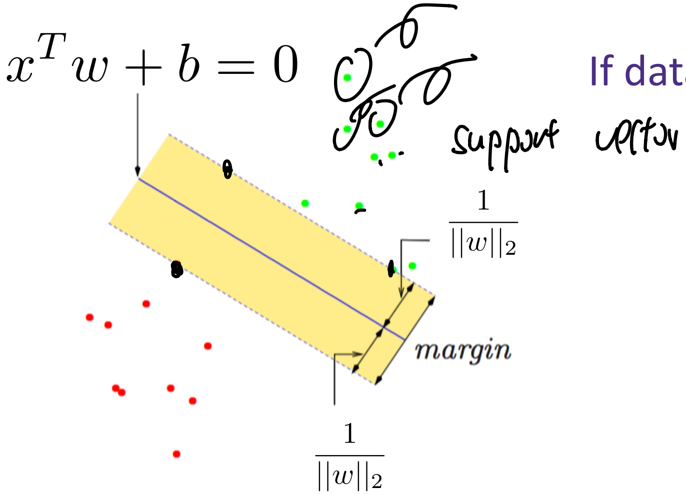


- Solve efficiently by many methods, e.g.,
  - quadratic programming (QP)
    - Well-studied solution algorithms
  - Stochastic gradient descent
  - Coordinate descent (in the dual)

Optimal Hyperplane (reparameterized)

$$\min_{w,b} ||w||_2^2$$
$$\text{subject to } y_i(x_i^T w + b) \geq 1 \quad \forall i$$

What are support vectors



If data is linearly separable

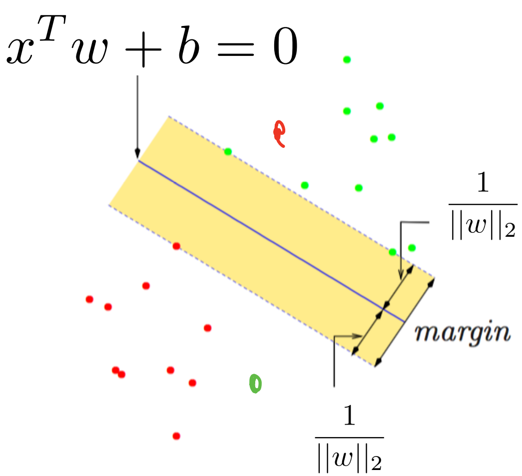
$$\min_{w,b} ||w||_2^2$$

$$\underbrace{y_i(x_i^T w + b) \geq 1 \quad \forall i}$$

x_i, $y_i(x_i^T w + b) = 1$

Note: the solution of this can be written in terms of very few of the training points. These points are known as support vectors.

What if the data is not linearly separable?



If data is linearly separable

$$\min_{w,b} ||w||_2^2$$

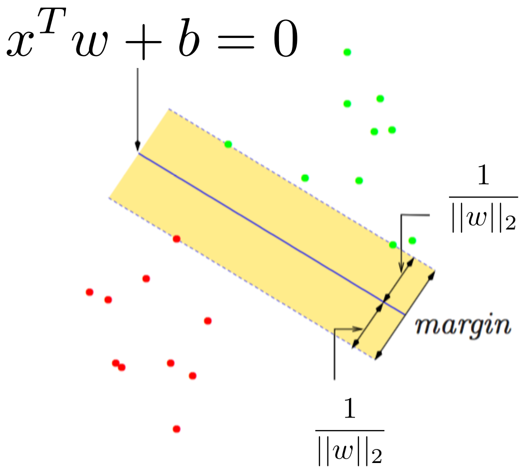
$$y_i(x_i^T w + b) \geq 1 \quad \forall i$$

If data is not linearly separable, some points don't satisfy margin constraint:

Two options:

1. Introduce slack to this optimization problem
2. Lift to higher dimensional space $\leftarrow$ Kernel

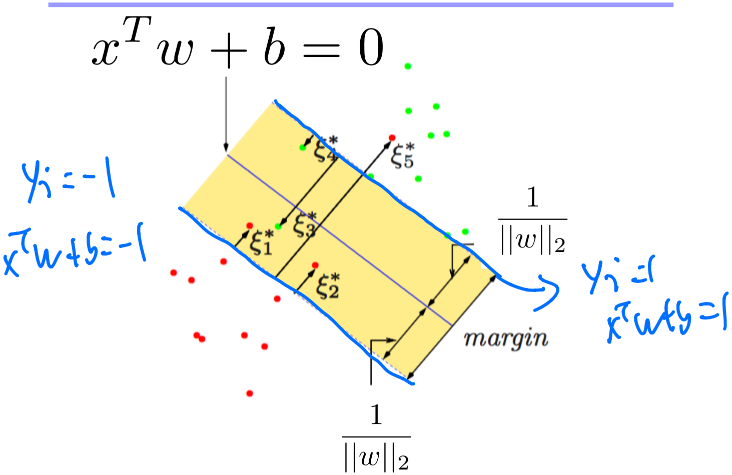
What if the data is not linearly separable?



If data is linearly separable:

$$\min_{w,b} ||w||_2^2$$

$$y_i(x_i^T w + b) \geq 1 \quad \forall i$$



If data is not linearly separable,

some points don't satisfy margin constraint:

$$\min_{w,b} ||w||_2^2$$

slack variable

$$y_i(x_i^T w + b) \geq 1 - \underline{\xi_i} \quad \forall i$$

$$\xi_i \geq 0, \sum_{j=1}^{n} \xi_j \leq \nu$$

convex

SVM as penalization method

- Original quadratic program with linear constraints:

$$\min_{w,b} ||w||_2^2$$

make margin big

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

$$\xi_i \geq 0, \sum_{j=1}^{n} \xi_j \leq \nu$$

violates constraints as little as possible

SVM as penalization method

ν_1 training data → (w_1, b_1) → validation set ↓ pick ν
ν_2
ν_3
ν_4 (w_4, b_4)

training val test

- Original quadratic program with linear constraints:

$$\min_{w,b} ||w||_2^2$$

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

$$\xi_i \geq 0, \sum_{j=1}^{n} \xi_j \leq \nu$$

$\nu = 1$

KKT condition

- Using same constrained convex optimization trick as for lasso:

For any $\nu \geq 0$ there exists a $\lambda \geq 0$ such that the solution the following solution is equivalent:

constraints margin

$$\min \boxed{\sum_{i=1}^{n} \max\{0, 1 - y_i(b + x_i^T w)\} + \lambda ||w||_2^2}$$

SVMs: optimizing what?

SVM objective:

data set: $\{(x_i, y_i)\}_{i=1}^{n}$

$x_i \in \mathbb{R}^d$

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

$$\boxed{\frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i(b + x_i^T w)\} + \lambda ||w||_2^2 \quad = \sum_{i=1}^{n} \ell_i(w, b)}$$

$$\nabla_w \ell_i(w, b) = \begin{cases} -x_i y_i + \frac{2\lambda}{n} w & \text{if } y_i(b + x_i^T w) < 1 \\ \frac{2\lambda}{n} & \text{otherwise} \end{cases}$$

$$\nabla_b \ell_i(w, b) = \begin{cases} -y_i & \text{if } y_i(b + x_i^T w) < 1 \\ 0 & \text{otherwise} \end{cases}$$