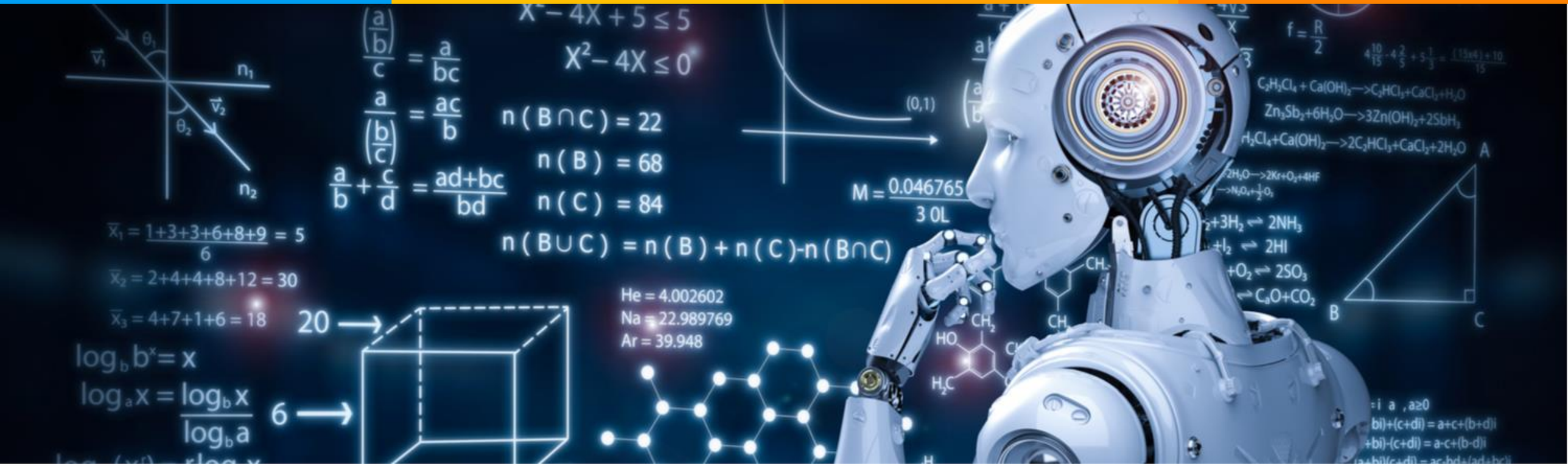
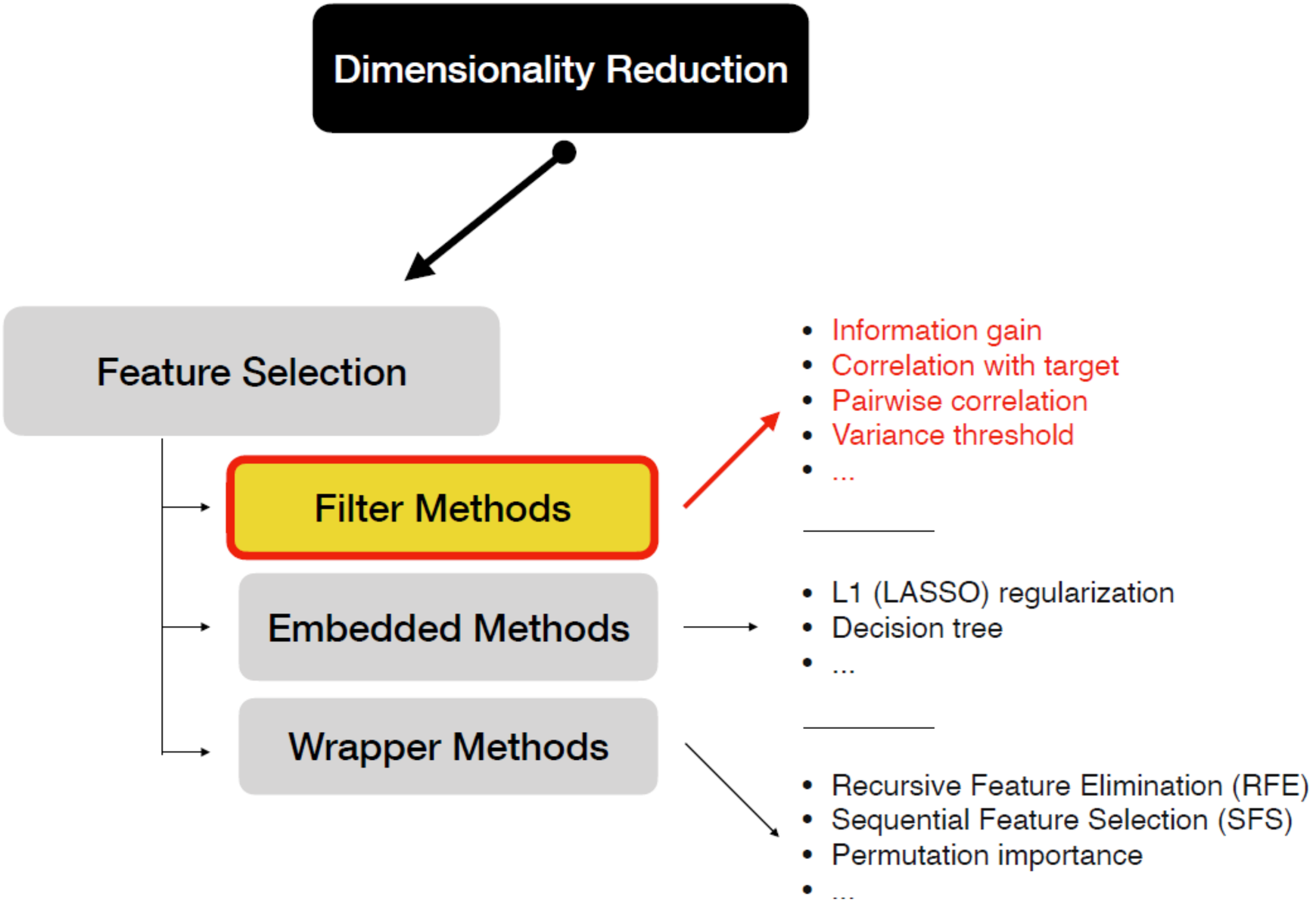
Duration: 60 min	Srikanth Shenoy	AML 5102	M.E. AI & ML

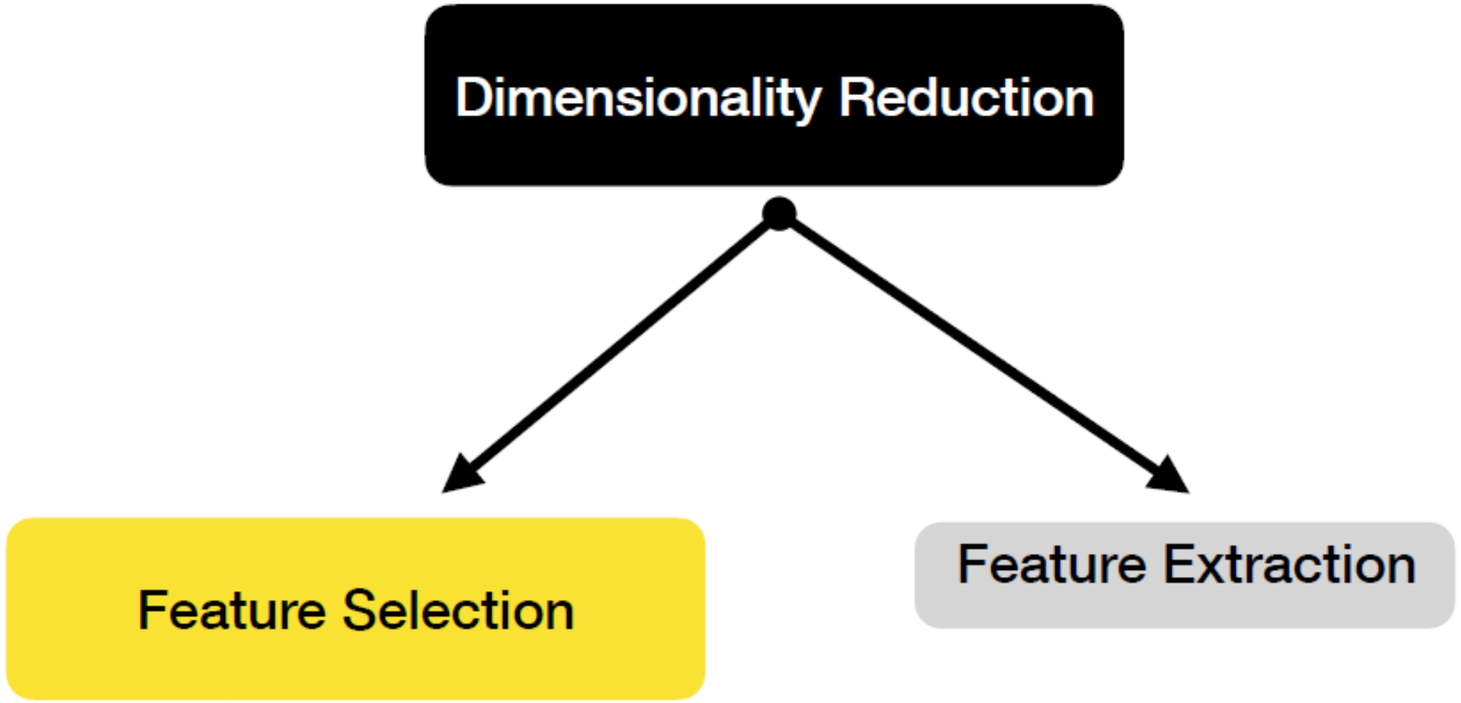


Lecture 30 & 31: Perceptron & SVM

Recap


Dimensionality Reduction
Feature Selection
Filter Methods
Embedded Methods
Wrapper Methods
Information gain
Correlation with target
Pairwise correlation
Variance threshold
...
L1 (LASSO) regularization
Decision tree
...
Recursive Feature Elimination (RFE)
Sequential Feature Selection (SFS)
Permutation importance
...

Feature Extraction



- Feature Extraction is combination of features
- Automatic elimination of unwanted features
- PCA, Kernel PCA, t-SNE, UMAP, Autoencoder


$X^2-4X+5\le 5$
$X^2-4X\le 0$
$n(B\cap C)=22$
$n(B)=68$
$n(C)=84$
$n(B\cup C)=n(B)+n(C)-n(B\cap C)$
$\frac{a}{b}+\frac{c}{d}=\frac{ad+bc}{bd}$
$\log_b b^x = x$
$\log_a x = \frac{\log_b x}{\log_b a}$
$\bar{x}_1 = \frac{1+3+3+6+8+9}{6} = 5$
$\bar{x}_2 = 2+4+4+8+12 = 30$
$\bar{x}_3 = 4+7+1+6 = 18$
20
6
(0,1)
$M = \frac{0.046765}{3\,0L}$
He = 4.002602
Na = 22.989769
Ar = 39.948
$f = \frac{R}{2}$
$C_2H_3Cl_4 + Ca(OH)_2 \rightarrow C_2HCl_3 + CaCl_2 + H_2O$
$Zn_3Sb_2 + 6H_2O \rightarrow 3Zn(OH)_2 + 2SbH_3$
$+3H_2 \rightleftharpoons 2NH_3$
$+I_2 \rightleftharpoons 2HI$
$+O_2 \rightleftharpoons 2SO_3$
A
B
C

Perceptron

Stage 2 – Early enthusiasm

Early enthusiasm (1957-70)

1. Chatbot, humanoid robot
2. Perceptron by Rosenblatt
3. AI was oversold

AI Rebirth (1980-90)

1. As expert systems.
2. As Machine Learning
3. Decision Trees

Rise of ML (1995-2011)

1. Backgammon success
2. SVM, Ensembles
3. Deep Blue - Gary Kasparov

Stage 1 | Stage 2 | Stage 3 | Stage 4 | Stage 5 | Stage 6 | Stage 7

Birth of AI (1952-56)

1. Samuel's Checker Program
2. Logic Theorists - First AI Program
3. Term "AI" was coined

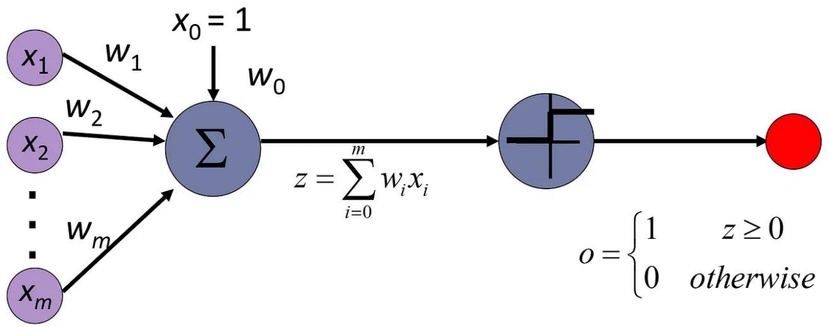
First AI Winter (1970-80)

1. Minsky Report - XOR
2. Bubble burst. Funding collapsed.

2nd AI Winter (1990-95)

1. Expert systems failed
2. Not enough computational power for Neural Networks
3. Funding collapsed

Rise of Deep Learning (2011-Now)



Perceptron is the idea behind neural networks

Stage 3 – First AI winter

Stage 1 — Birth of AI (1952-56)

1. Samuel's Checker Program
2. Logic Theorists - First AI Program
3. Term "AI" was coined

Stage 2 — Early enthusiasm (1957-70)

1. Chatbot, humanoid robot
2. Perceptron by Rosenblatt
3. AI was oversold

Stage 3 — First AI Winter (1970-80)

1. Minsky Report - XOR
2. Bubble burst. Funding collapsed.

Stage 4 — AI Rebirth (1980-90)

1. As expert systems.
2. As Machine Learning
3. Decision Trees

Stage 5 — 2nd AI Winter (1990-95)

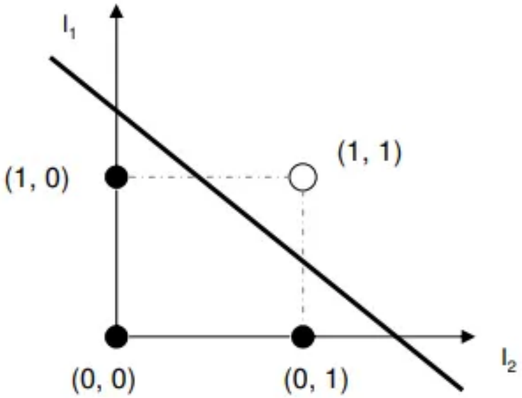
1. Expert systems failed
2. Not enough computational power for Neural Networks
3. Funding collapsed

Stage 6 — Rise of ML (1995-2011)

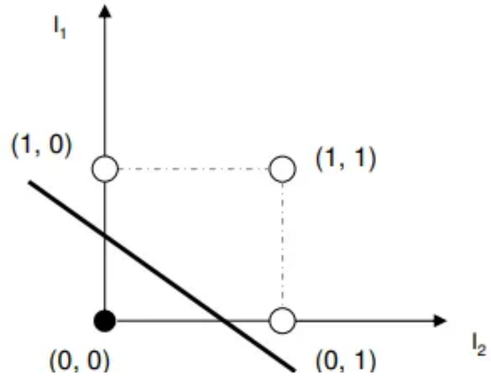
1. Backgammon success
2. SVM, Ensembles
3. Deep Blue - Gary Kasparov

Stage 7 — Rise of Deep Learning (2011-Now)

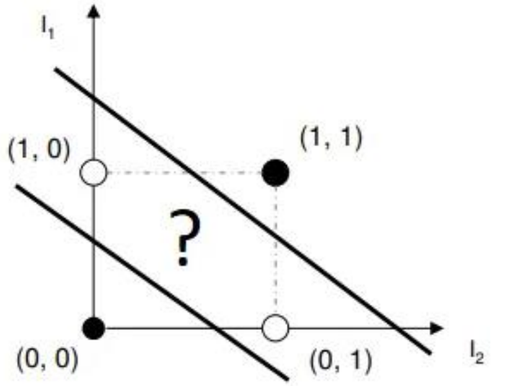
AND		
I_1	I_2	out
0	0	0
0	1	0
1	0	0
1	1	1



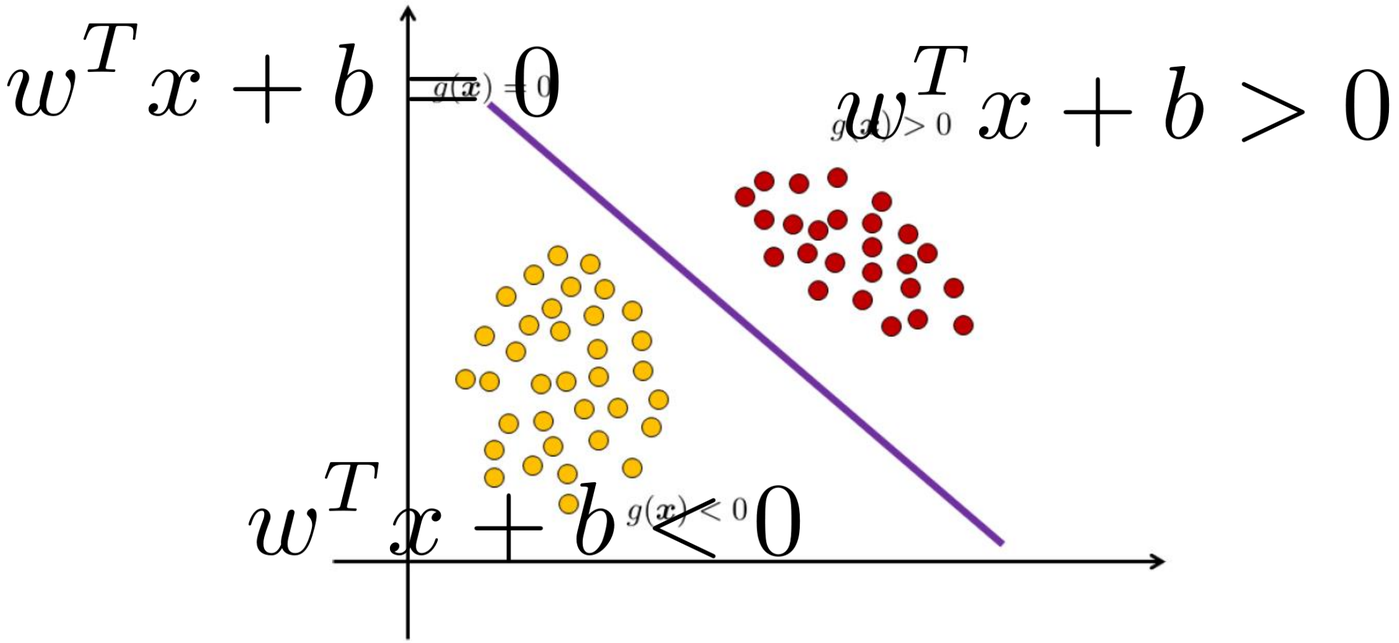
OR		
I_1	I_2	out
0	0	0
0	1	1
1	0	1
1	1	1

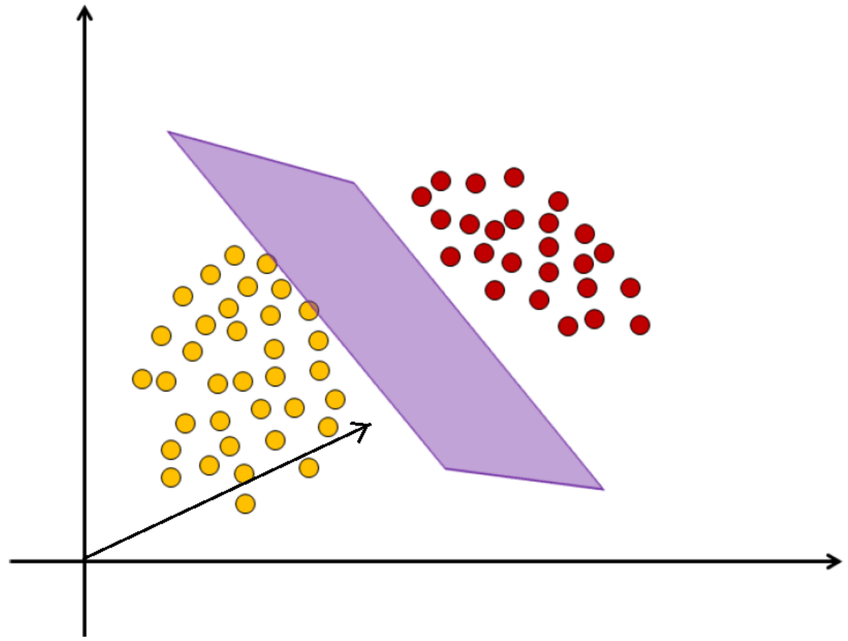


XOR		
I_1	I_2	out
0	0	0
0	1	1
1	0	1
1	1	0



Decision boundary in binary classification



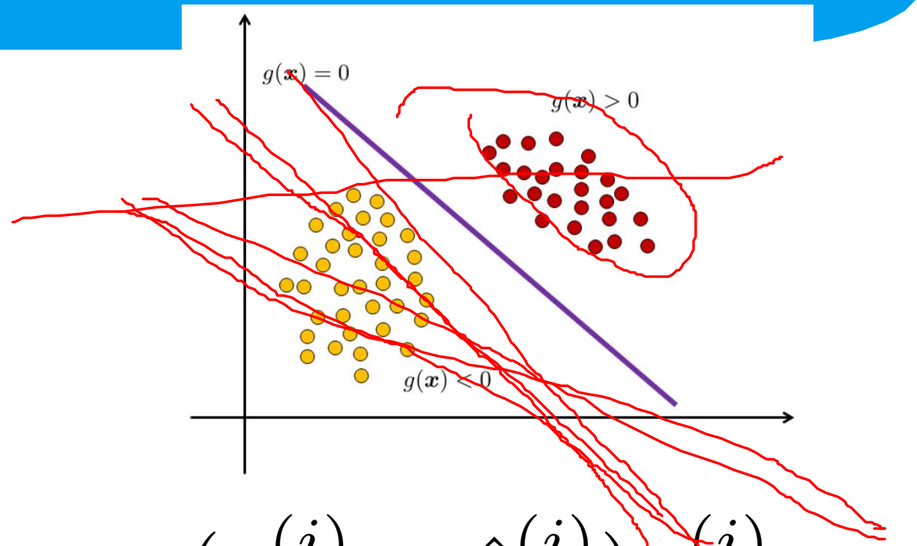


- y is +1, -1
- Calculate wTx+b for a given x
- Product of y and wTx+b
- Sign() function

y	y-hat	Prediction
1	1	Correct
-1	-1	Correct
1	-1	Incorrect
-1	1	Incorrect

Perceptron learning algorithm

- Select random w and b
- while num_iter < K
  - for each record in dataset
    - yhat = wTx + b
    - If y * yhat < 0
      - adjust w and b



$$w = w + \alpha(y^{(i)} - \hat{y}^{(i)})x^{(i)}$$

$$b = b + \alpha(y^{(i)} - \hat{y}^{(i)})$$

$$\hat{y}^{(i)}_{updated} = y^{(i)} \cdot (w^{T}_{updated}x^{(i)} + b_{updated})$$

Simplify

$$\hat{y}^{(i)}_{updated} = (positive\ \ number)$$

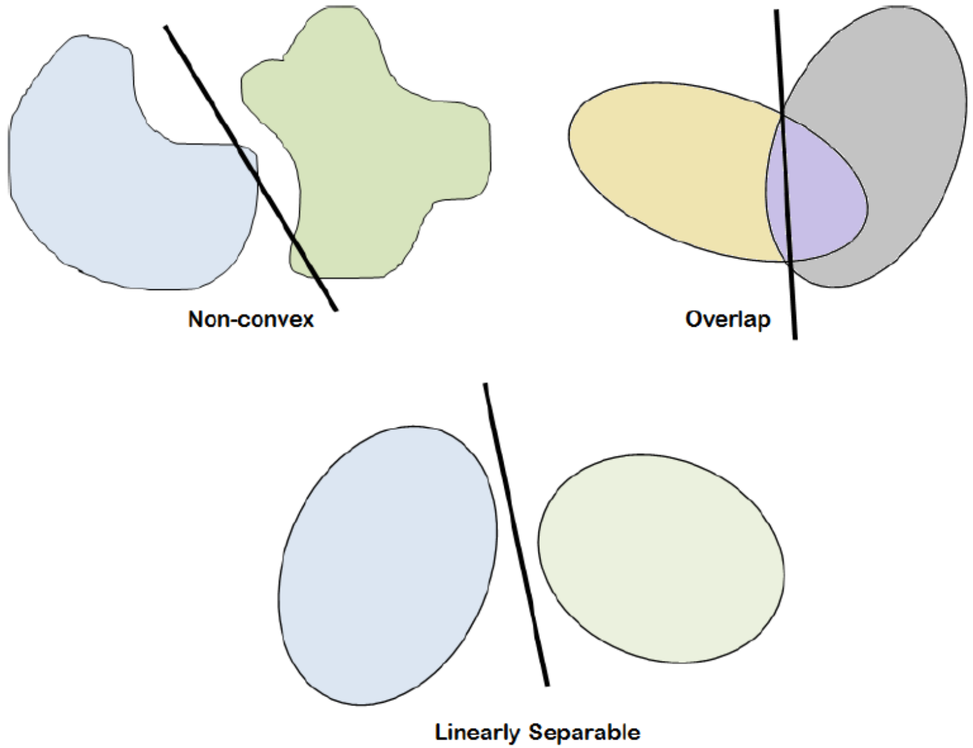
Problems with Perceptron learning algorithm

- Adjustments to w & b not systematic

$$w = w + \alpha(y^{(i)} - \hat{y}^{(i)})x^{(i)}$$

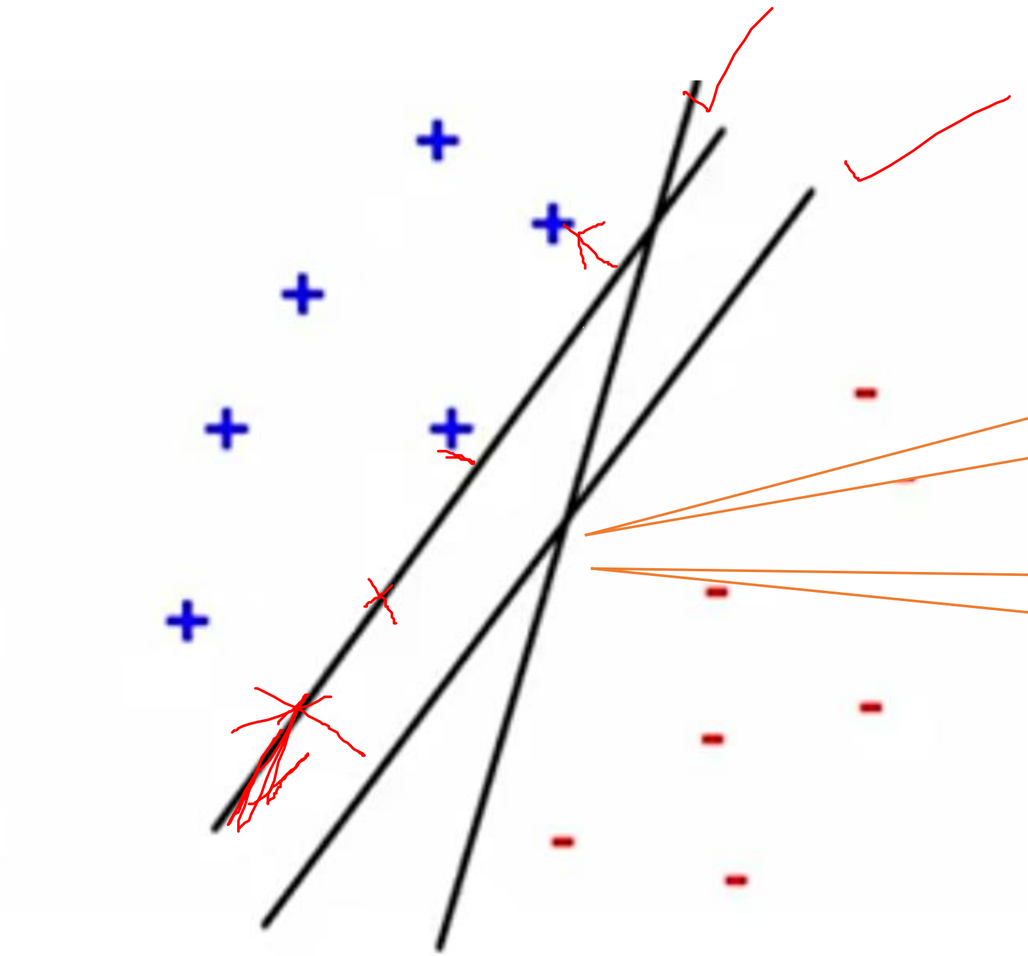
$$b = b + \alpha(y^{(i)} - \hat{y}^{(i)})$$

- Does not converge for non linear decision boundary



Problems with perceptron learning algorithm

- May get any of these boundaries

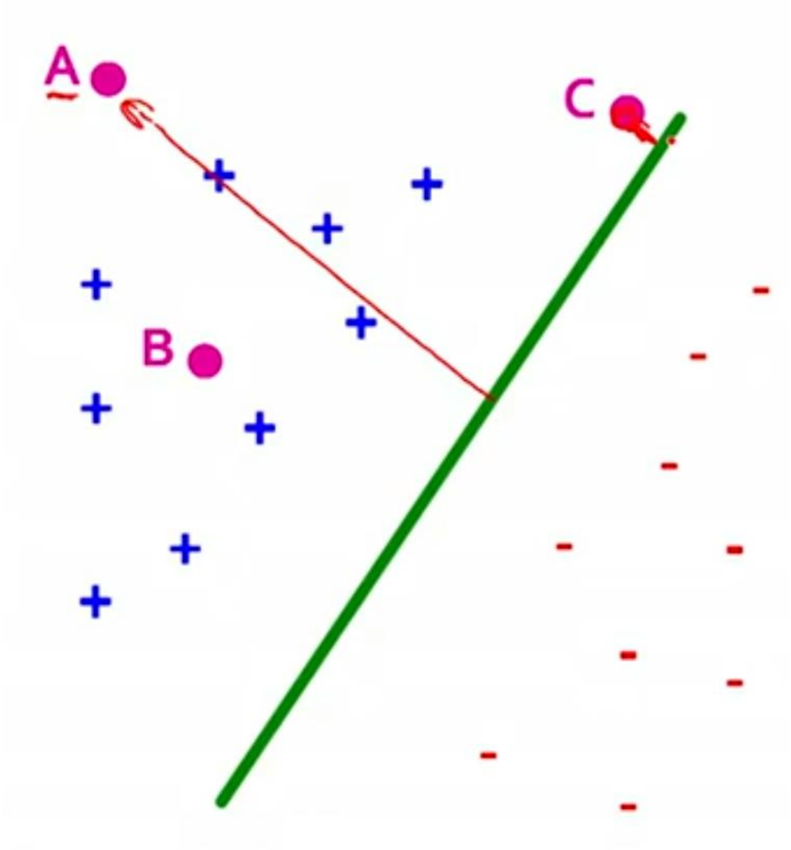


How to maximize margin?

Ans: SVM

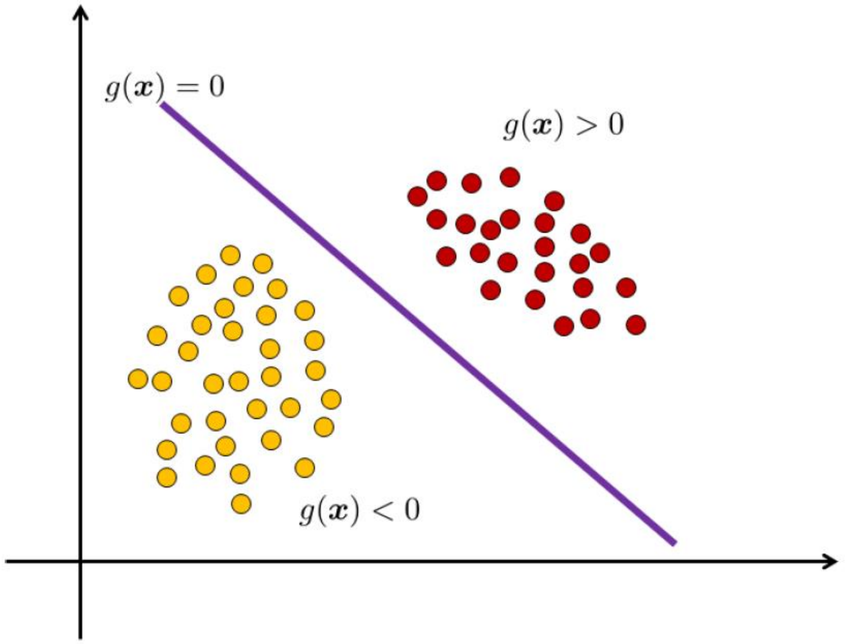
SVM intuition: Distance and confidence level

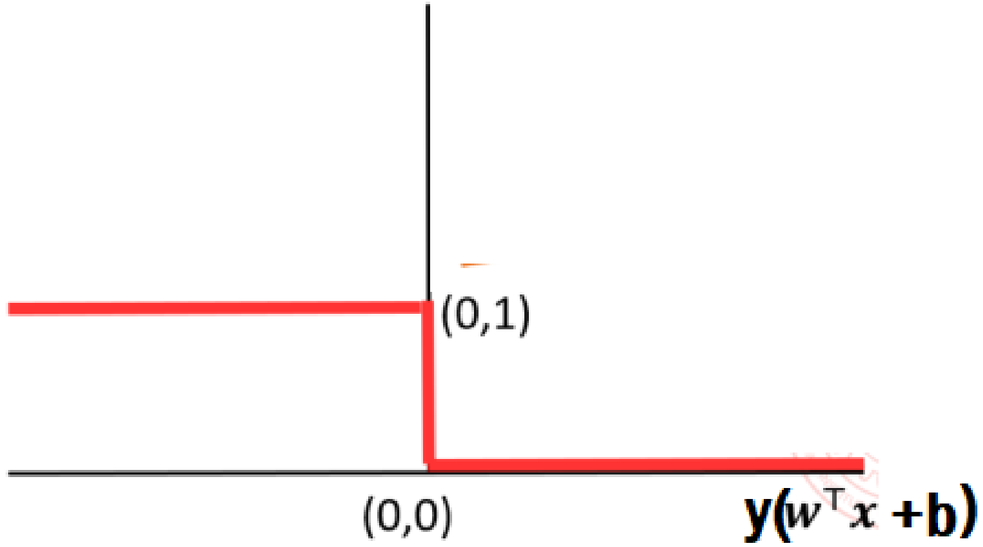
- Distance from the hyperplane is a measure of confidence of prediction



Perceptron & SVM loss functions

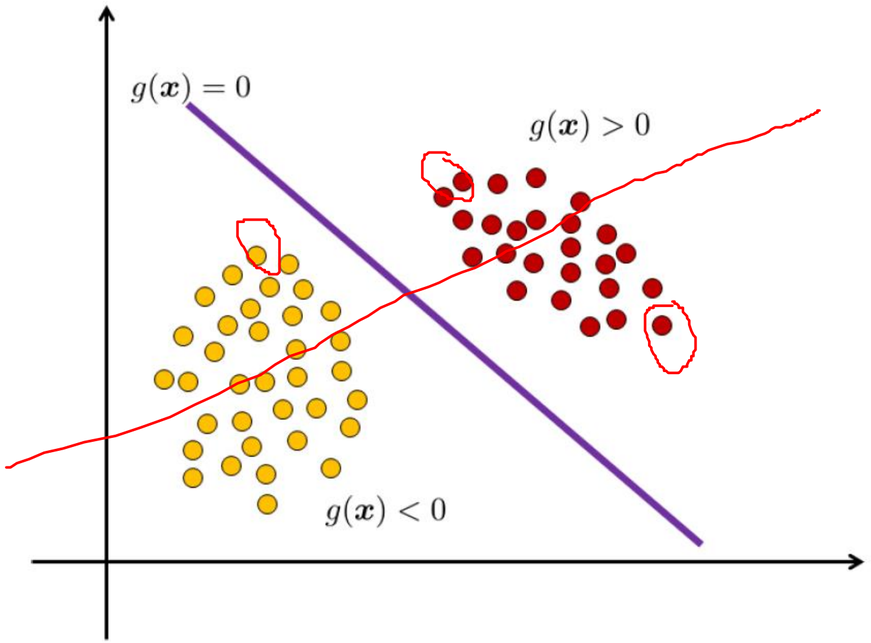
A rudimentary loss function in perceptron

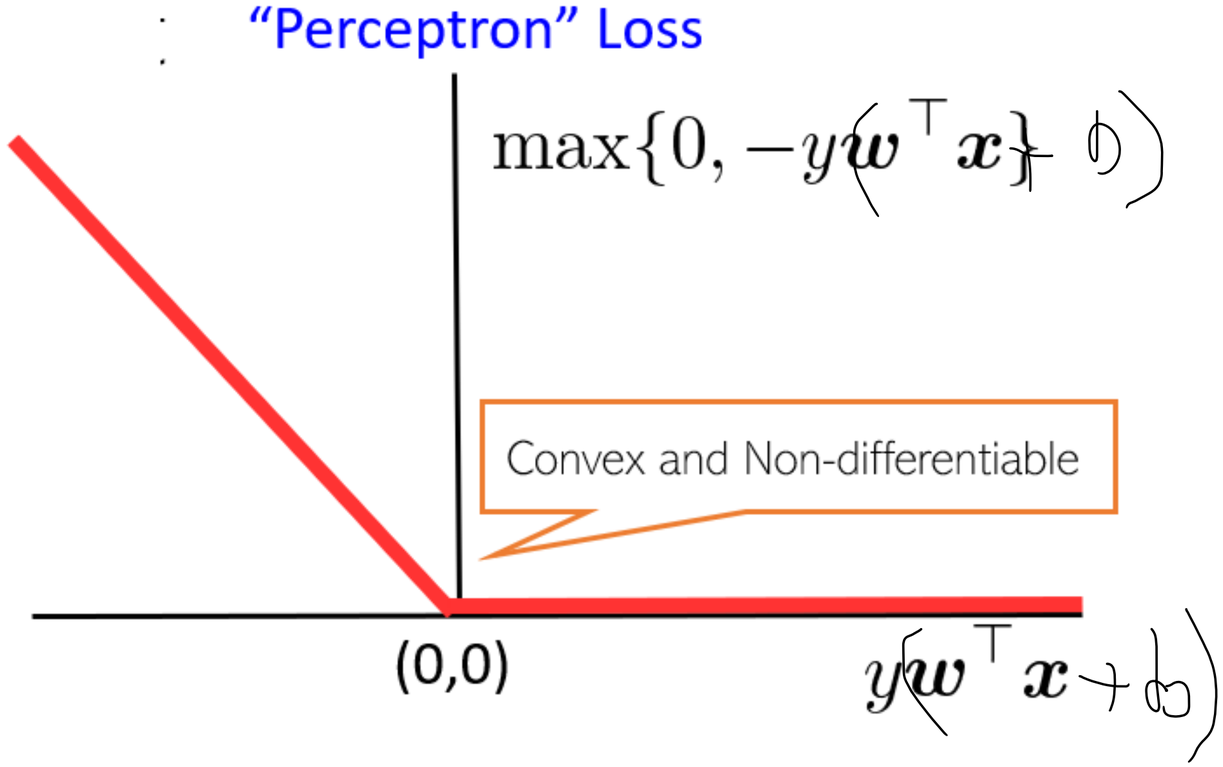




- Problems
  - Non convex, non differentiable
  - Loss does not take distance into account

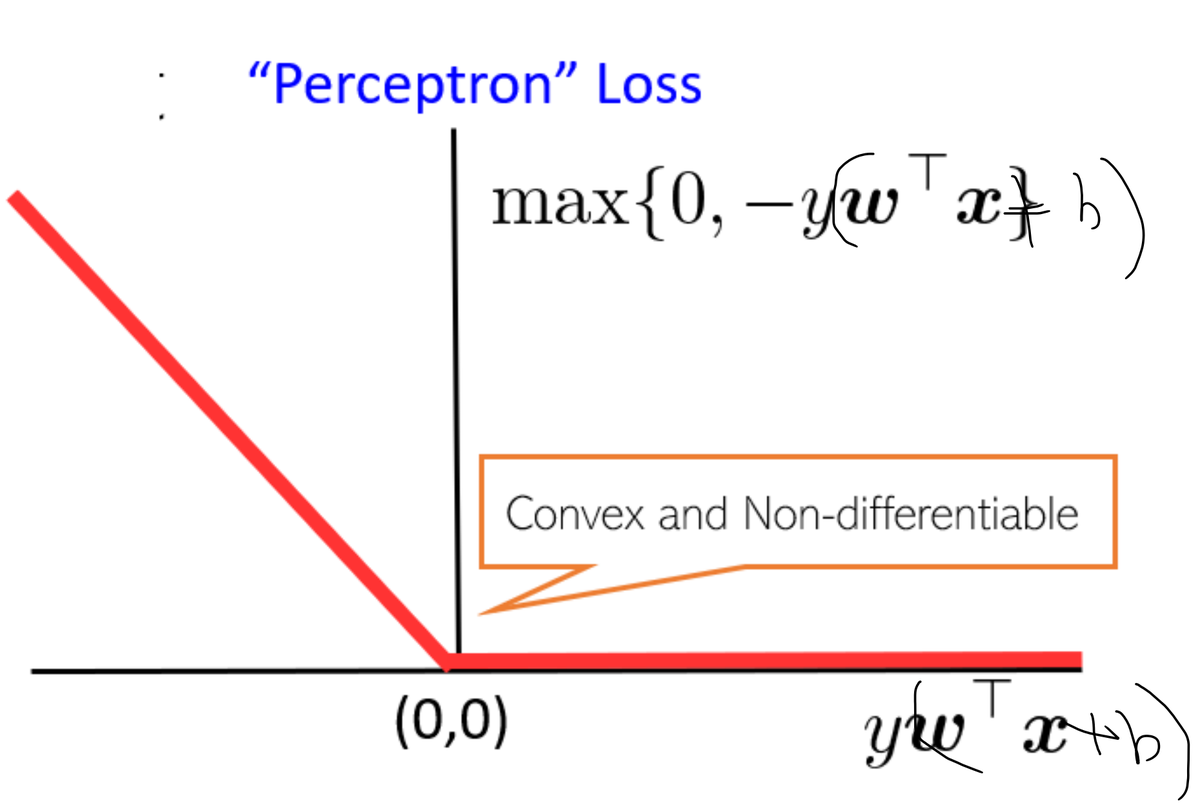
A better loss function

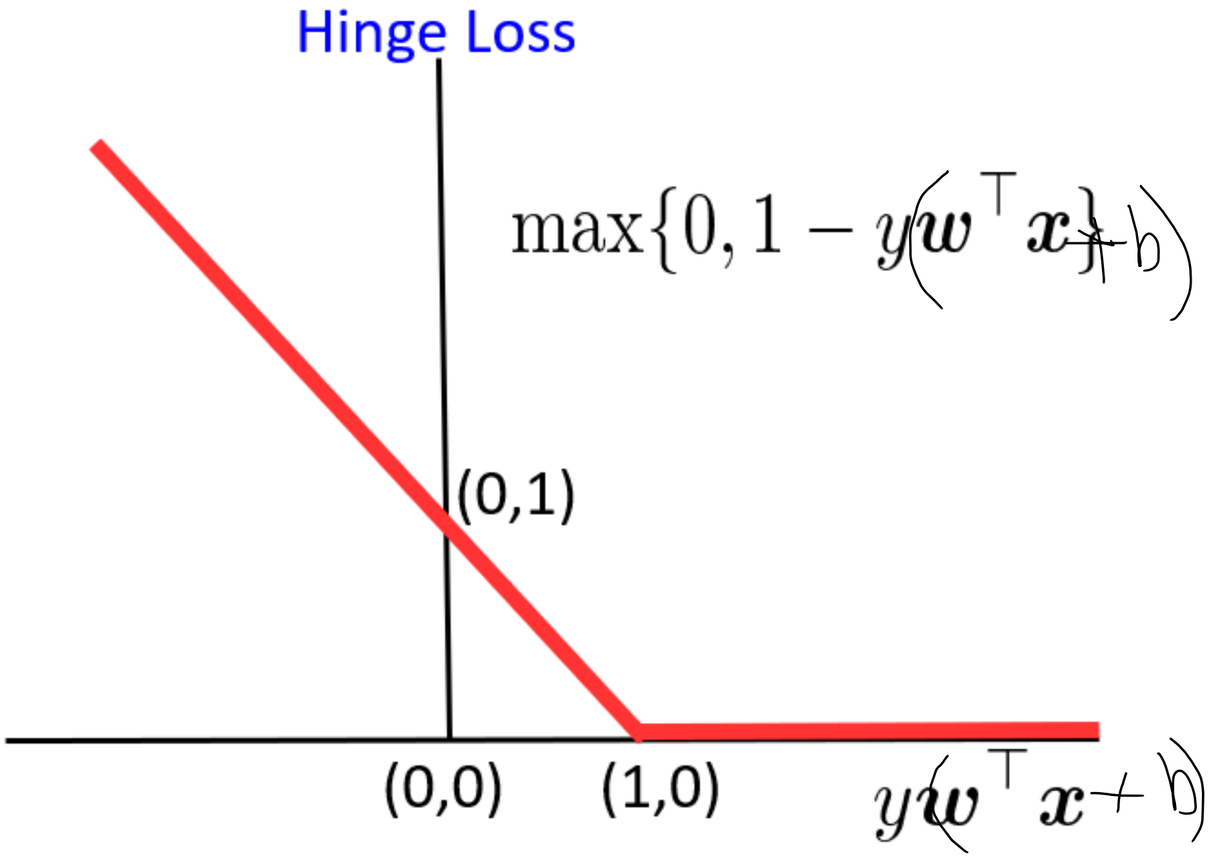




- But How to calculate distance?

An even better loss function



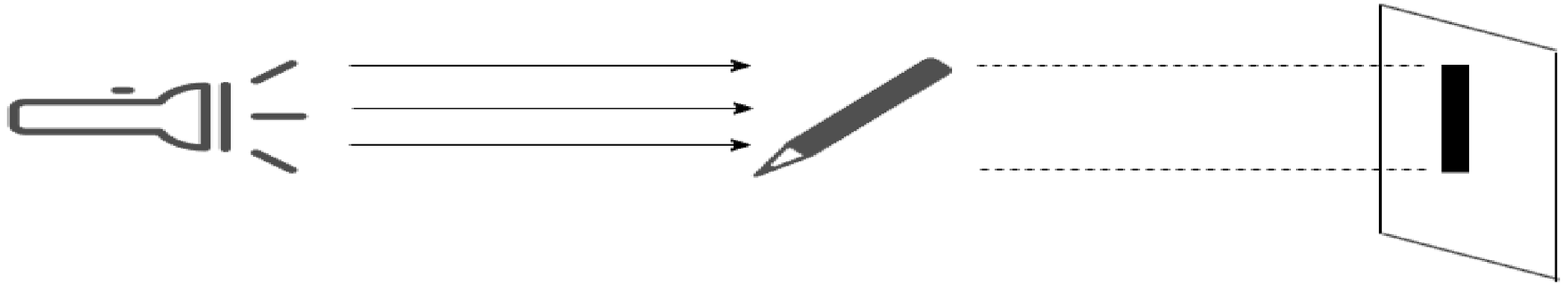


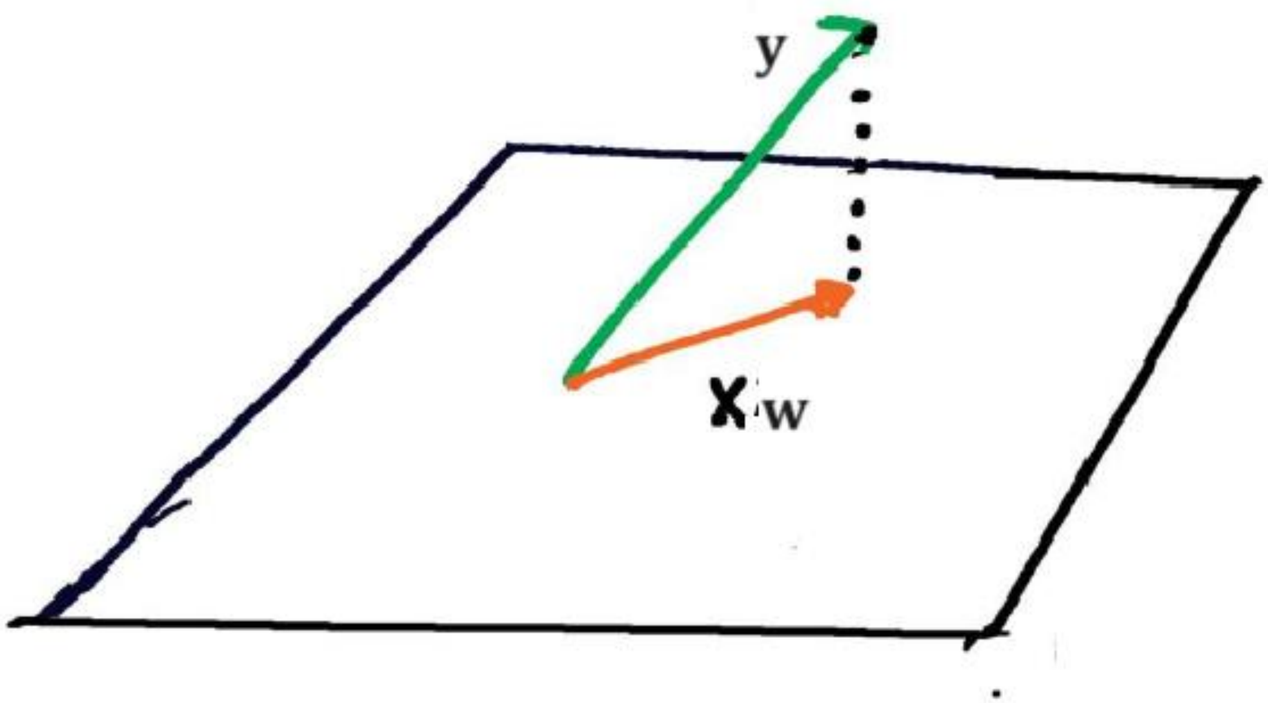
Linear Algebra of separating hyperplane

Length of the projection

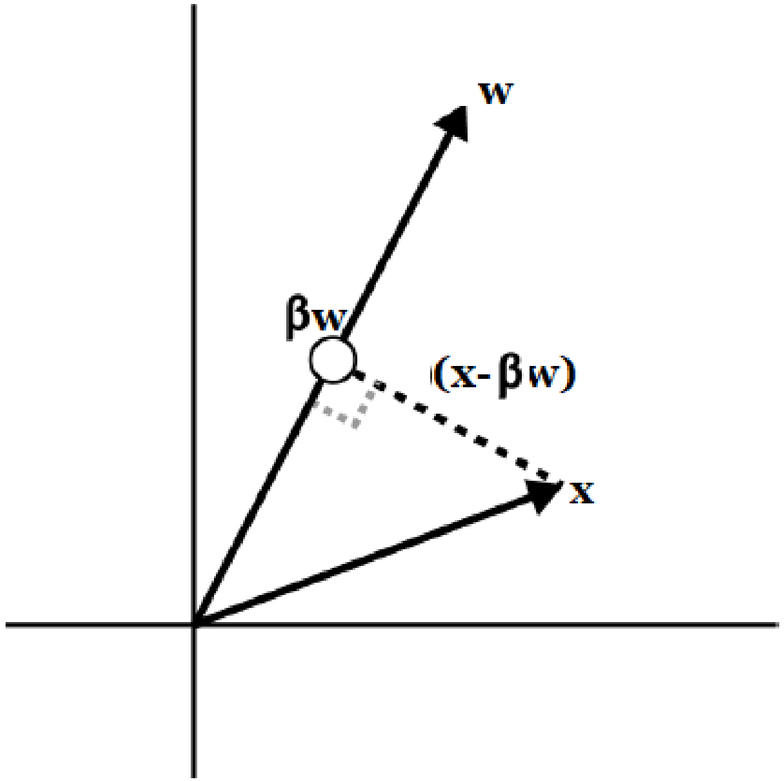
- Dot product
- Orthogonality
- Dot product of orthogonal vectors = 0

Dot Product, Shortest distance





Orthogonal component w.r.t. another vector

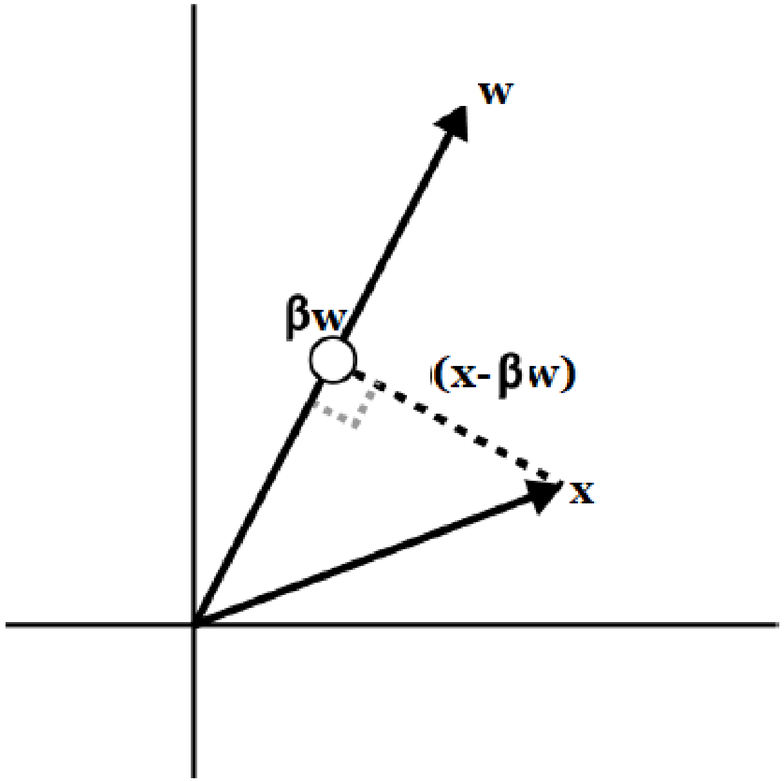


- Projection of x onto w βw
- Difference of projection vector βw and x is $x - \beta w$
- Projection vector βw is such as to minimize distance $x - \beta w$
- Then w and $x - \beta w$ are orthogonal

$$w^T(x - \beta w) = 0 \implies w^T x = \beta w^T w \implies \beta = \frac{w^T x}{w^T w}$$

$$\implies \beta w = \frac{w^T x}{\|w\|^2} w$$

Orthogonal component w.r.t. another vector



$$\beta = \frac{w^T x}{w^T w} \implies \beta w = \frac{w^T x}{\|w\|^2} w$$

$$\beta w = \frac{w^T x}{\|w\|} \frac{w}{\|w\|}$$

Unit vector in the direction of w

Length of projection $= \dfrac{w^T x}{\|w\|}$

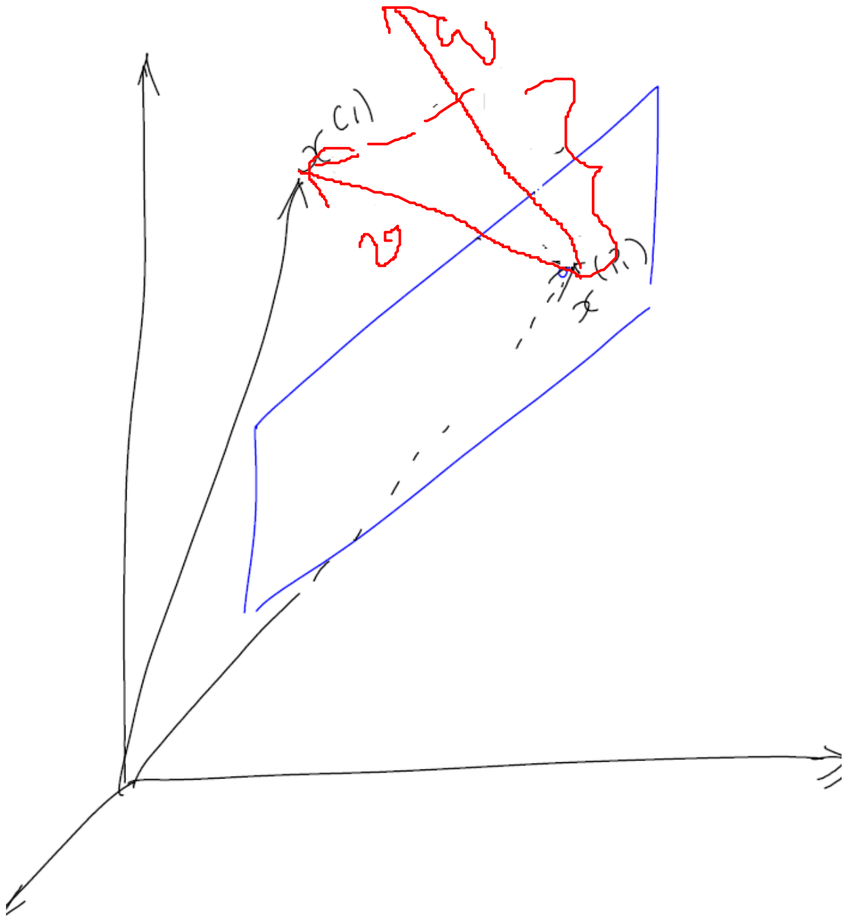
w and separating hyperplane

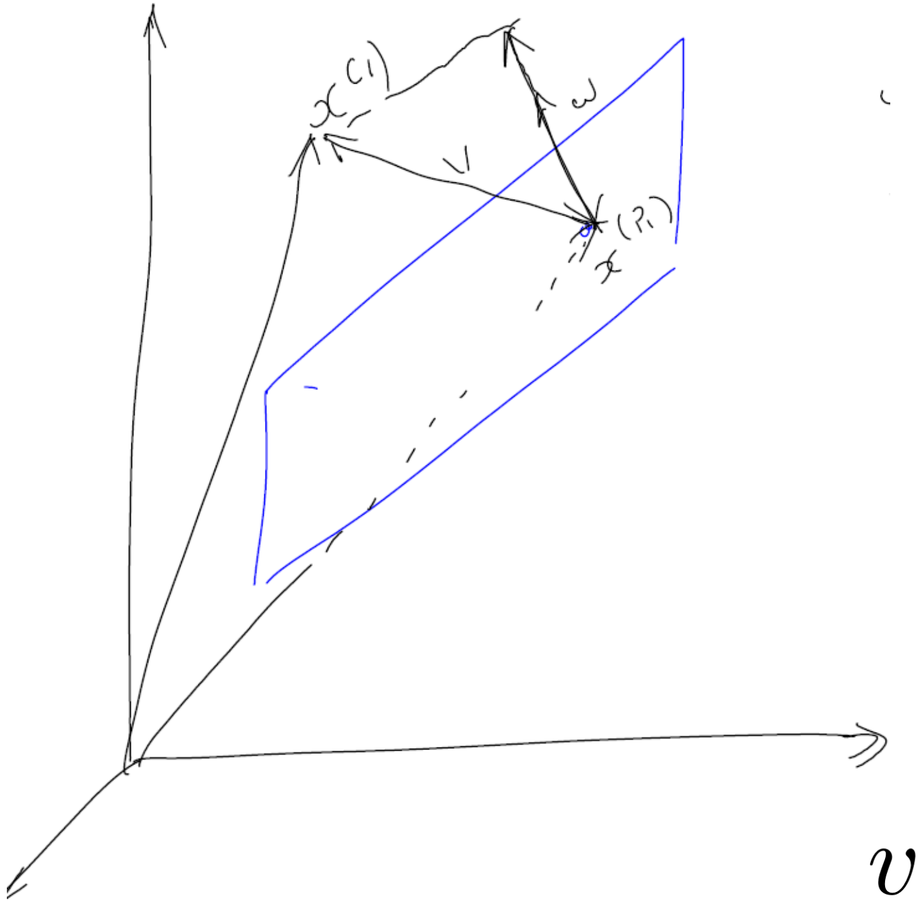
- Demo $w^T x = 0$
- Demo $w^T x = k; \quad w^T x = -k$

$$w^T x - k = 0; \quad w^T x + k = 0$$

- Generic equation of hyperplane $w^T x + b = 0$
- w is orthogonal to hyperplane
- NOTE: w does not include the intercept

Distance of a point from hyperplane

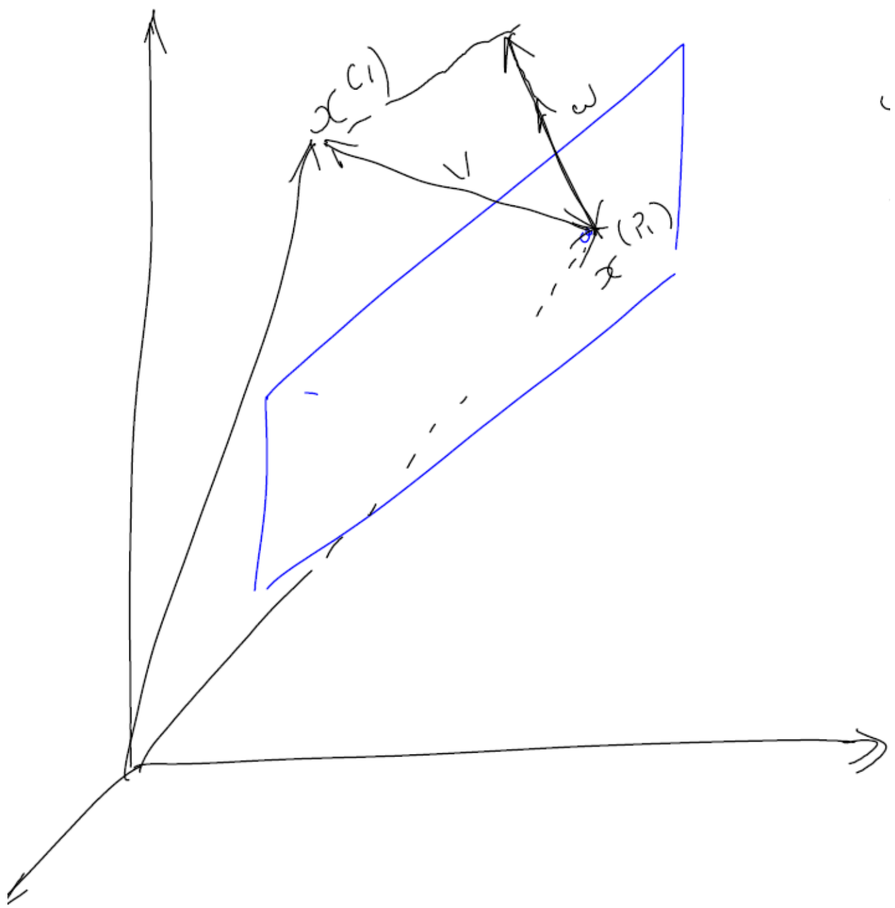




$$v = x^{(1)} - x^{(p)}$$

Dist = projection of v on w = $\dfrac{v^T w}{\|w\|}$

Distance of a point from hyperplane



$$\frac{v^T w}{\|w\|} = \frac{w^T x^{(1)} - w^T x^{(p)}}{\|w\|}$$

$$w^T x^{(p)} + b = 0$$

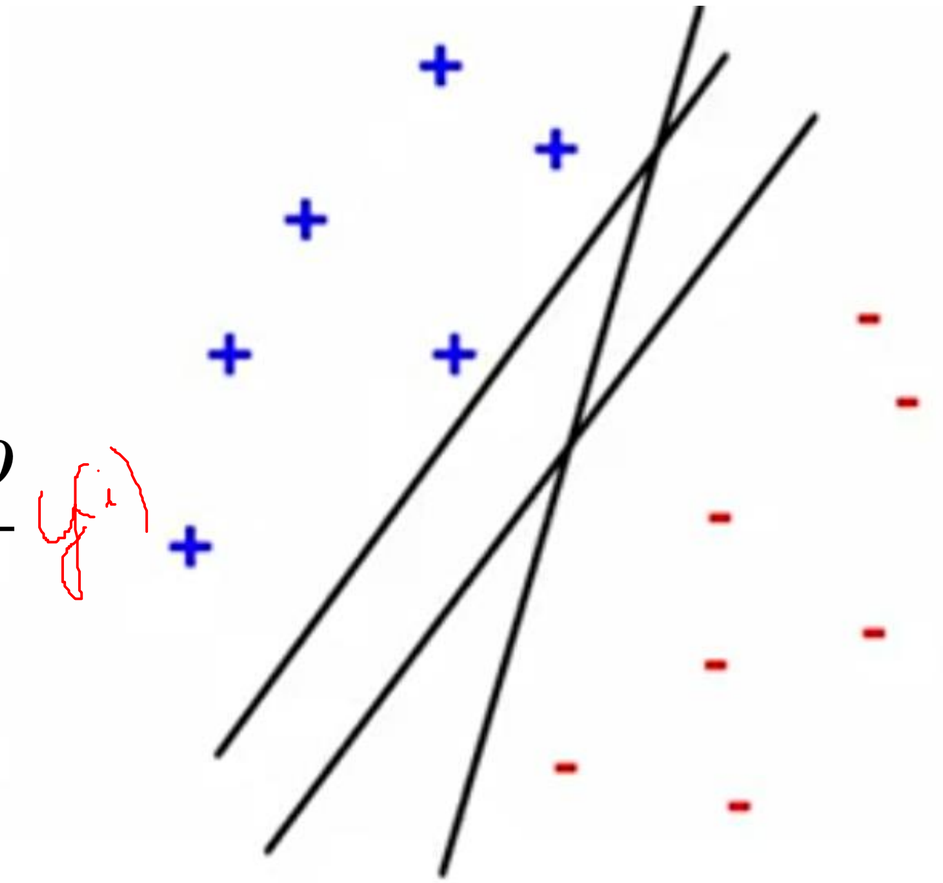
Dist = $$\frac{w^T x^{(1)} + b}{\|w\|}$$

Support Vector Machine

First attempt at SVM objective function

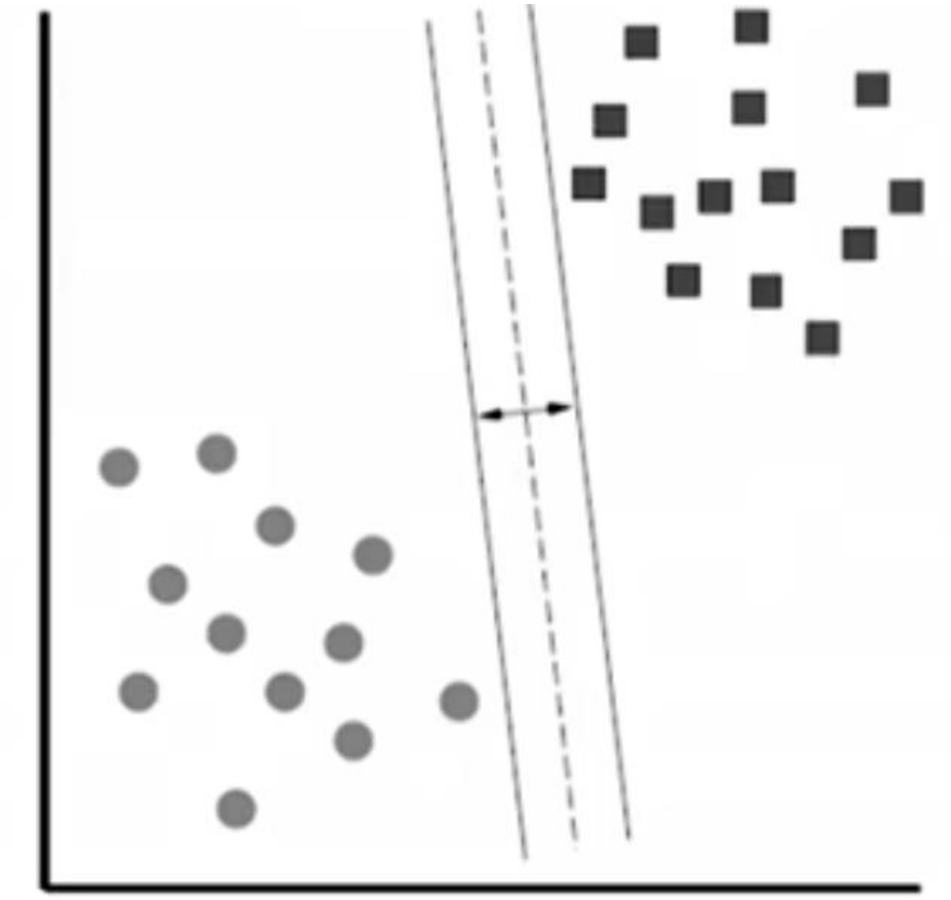
- What is the best linear separator?
- How about maximizing mean distance?

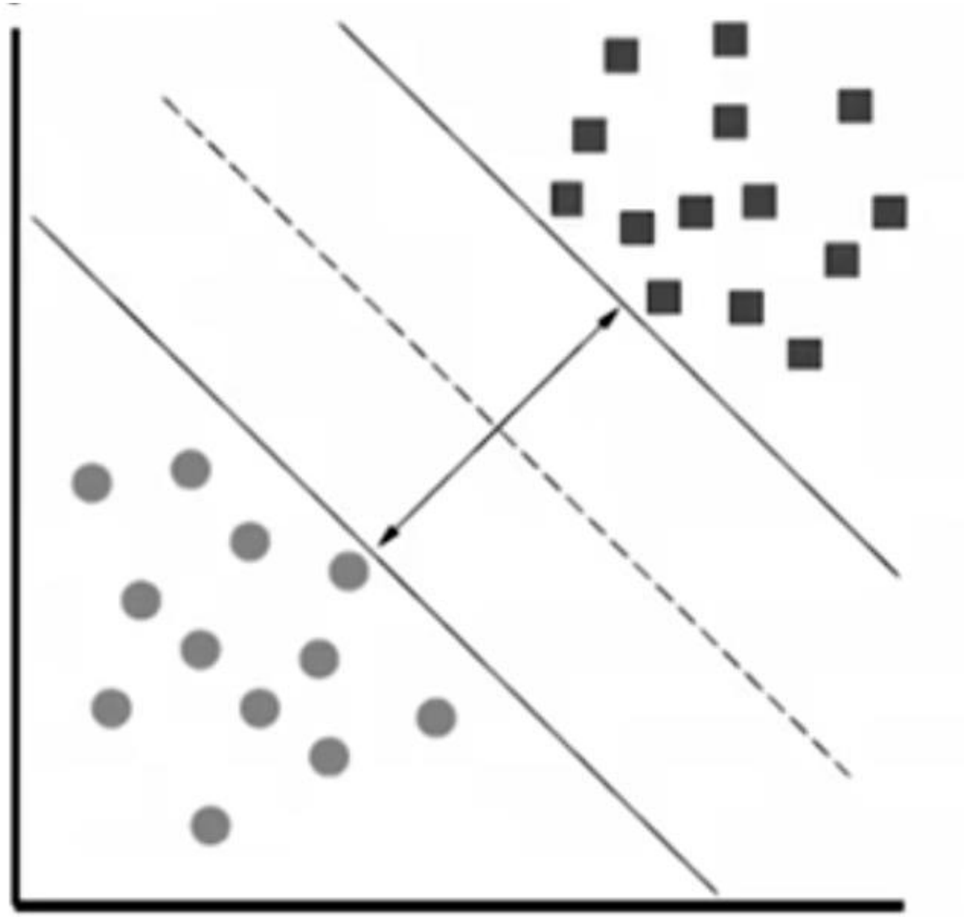
$$\mathcal{J} = \arg\max_{w} \frac{1}{m} \sum_{i=1}^{m} \frac{w^T x^{(i)} + b}{\|w\|}$$



- Two problems
  - Closest points don't add much
  - Increasing w increases avg

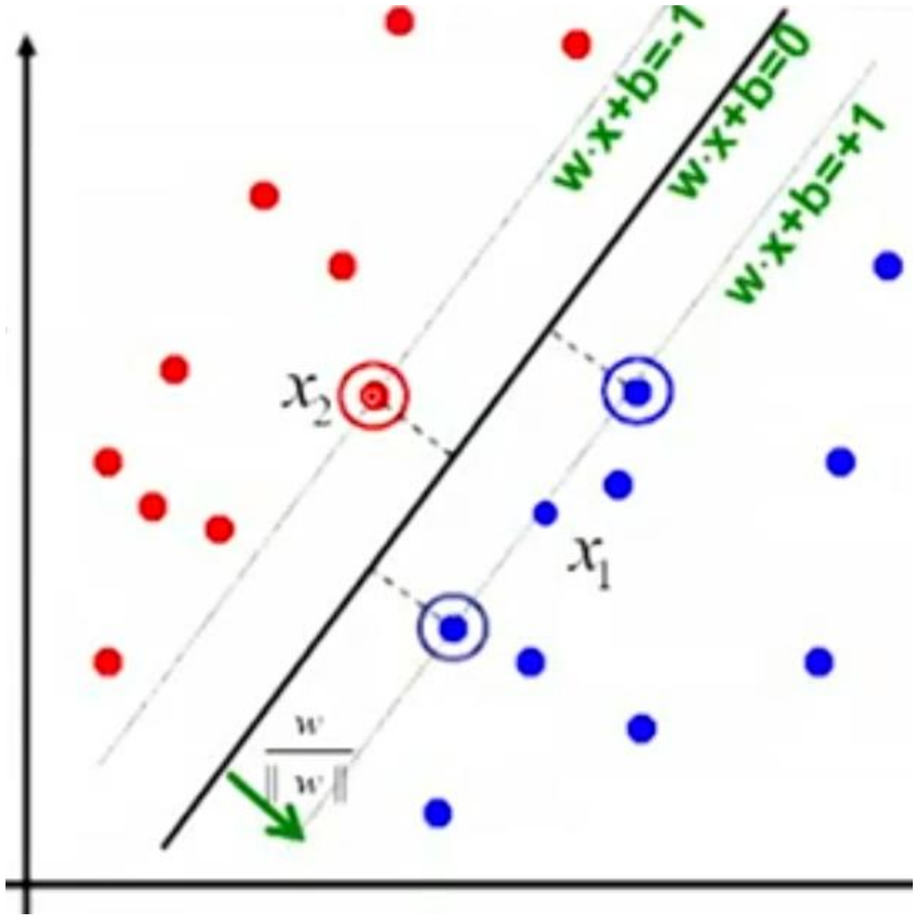
SVM Intuition – Margin





- Margin γ - distance of closest example from hyperplane

SVM Intuition – Support Vectors

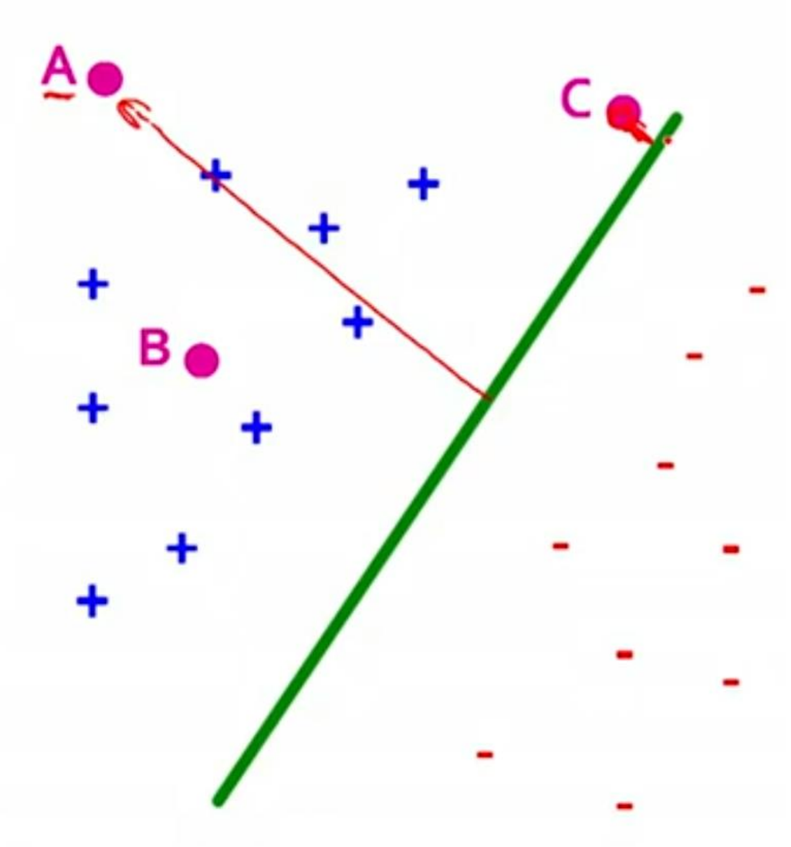


- Focus on the two outer lines instead of hyperplane
- Three data points (vectors) support the two lines
- Hence the name Support Vector Machine
- In general d dimensional data requires d+1 support vectors at minimum)

Formulate objective function using margin

- Maximum Margin
  - Pick w such that the closest data point has largest distance among all possible hyperplanes for different w

$$\gamma = \min dist$$



- For ith data point
  - dist $= \left(\frac{w^T x^{(i)} + b}{\|w\|} \right) y^{(i)}$

$$\mathcal{J} = \arg\max_{w,b} [min\ dist] = \arg\max_{w} \gamma$$

$$s.t. \forall i, y^{(i)}(w^T x^{(i)} + b) \geq \gamma$$

How to calculate margin?

Maximizing Margin

- Equations for supporting hyperplanes

$$w^T x^{(1)} + b = 1$$

$$w^T x^{(2)} + b = -1$$

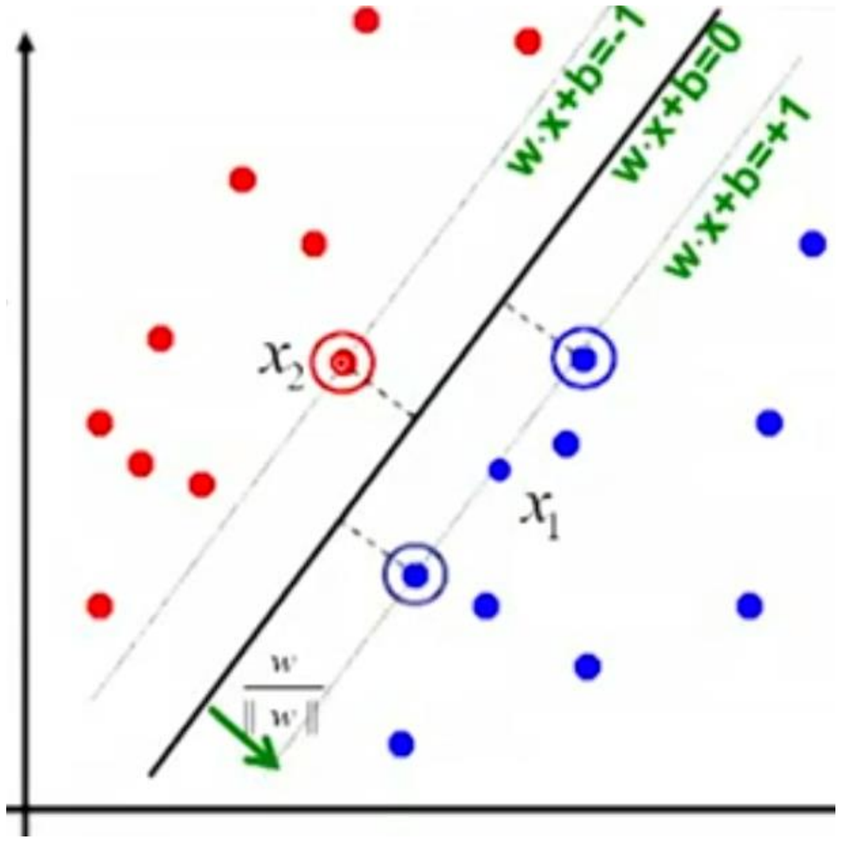
Let us assume wx+b= +/-1 for closest data point

- Relation between x1 and x2

$$x^{(1)} - x^{(2)} = 2\gamma \frac{w}{\|w\|}$$

- Solving, we get

$$\gamma = \frac{1}{\|w\|}$$



Final Objective function for SVM

- Objective function $\mathcal{J} = \arg\max_{w,b} \gamma$

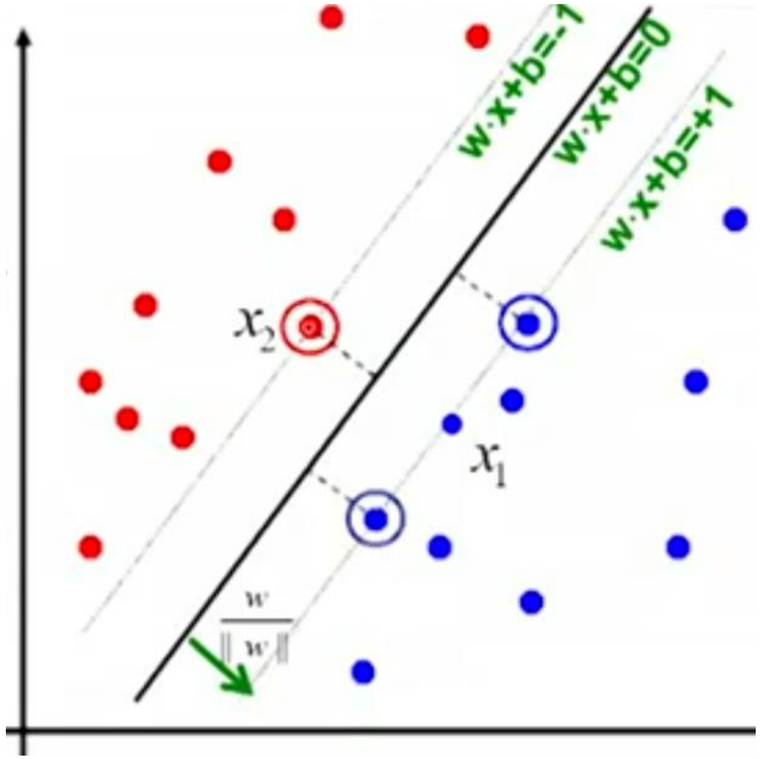
$$s.t. \forall i, y^{(i)}(w^T x^{(i)} + b) \geq \gamma$$

$$\gamma = \frac{1}{\|w\|}$$

SVM with hard margin

$$\mathcal{J} \approx \max_{w,b} \frac{1}{\|w\|} \approx \min \|w\| \approx \min \frac{\|w\|^2}{2}$$

$$s.t. \forall i, \quad y^{(i)}(w^T x^{(i)} + b) \geq 1$$

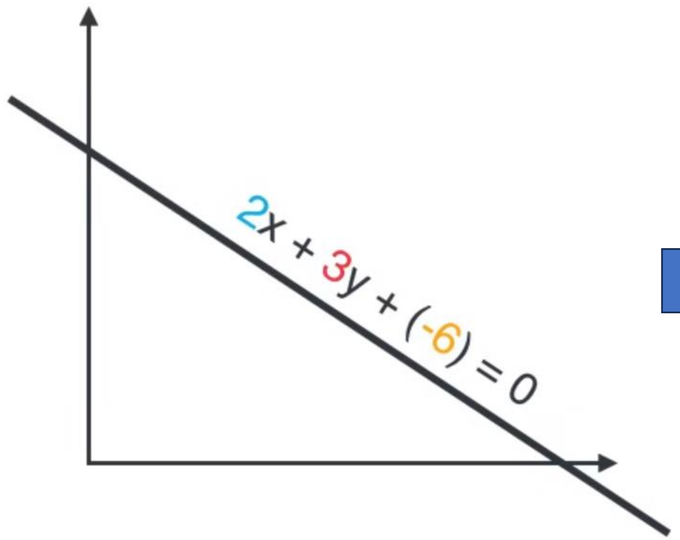


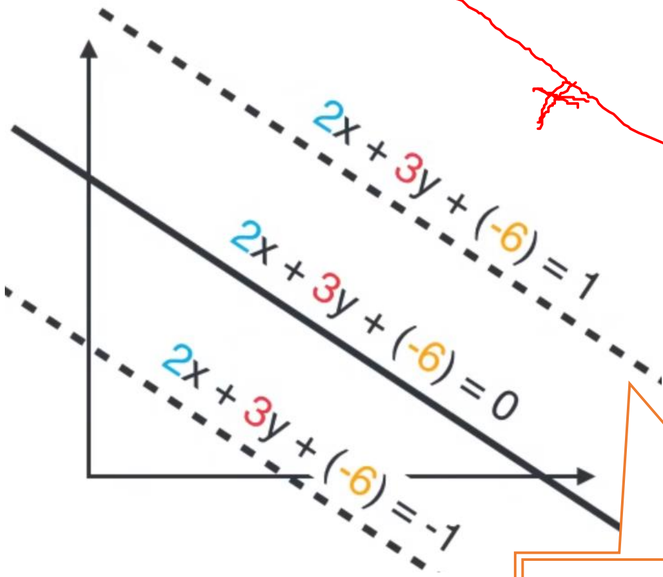
Multiple Margin Condition/Constraints

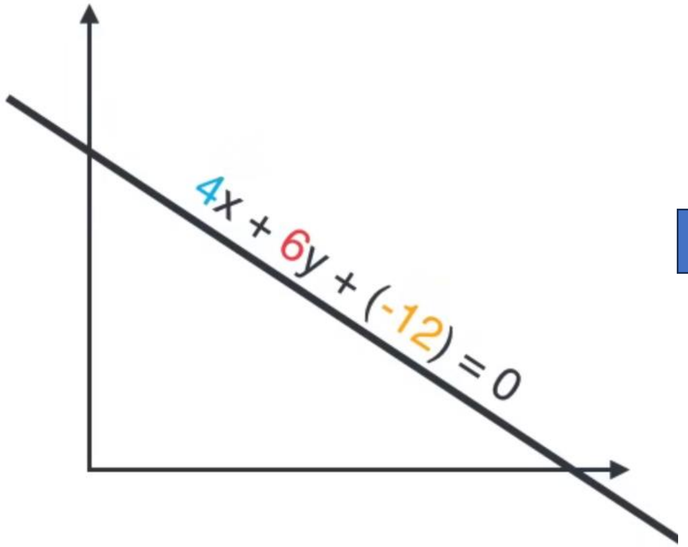
Equality holds for only support vectors

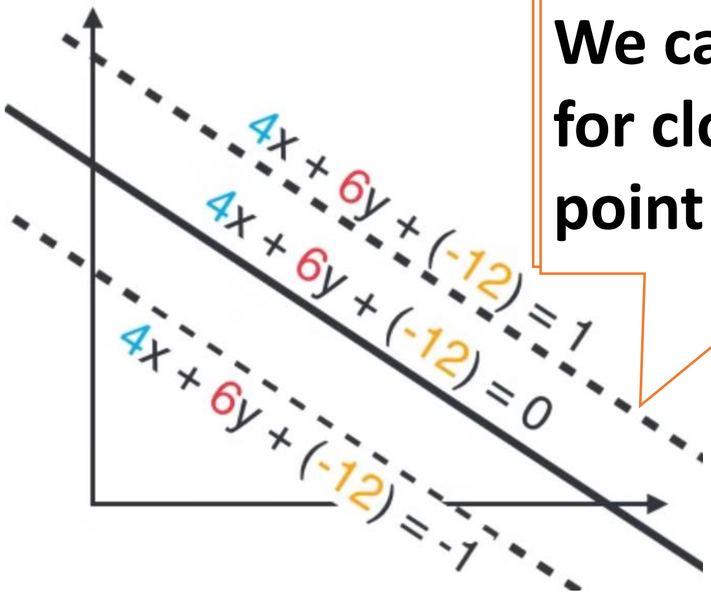
min w = 0 vector without constraints

Why wx+b = +/- 1 for supporting hyperplanes?







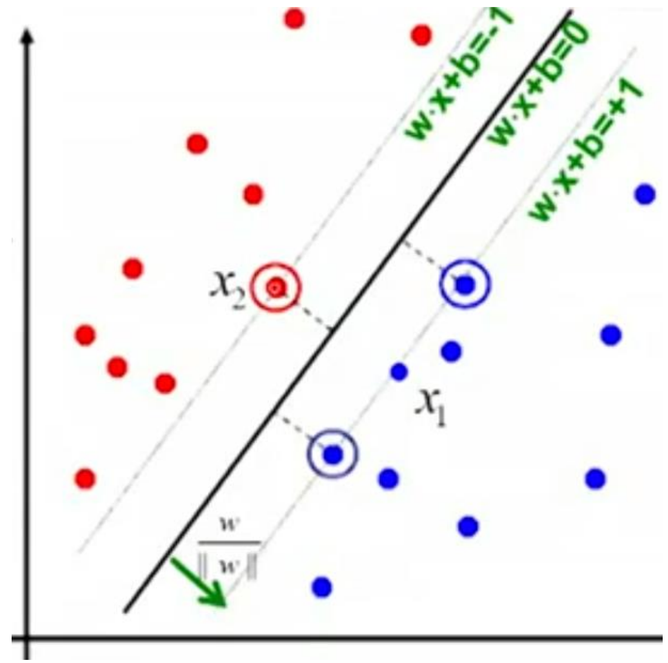


We can fit wx+b=1 for closest data point by adjusting w

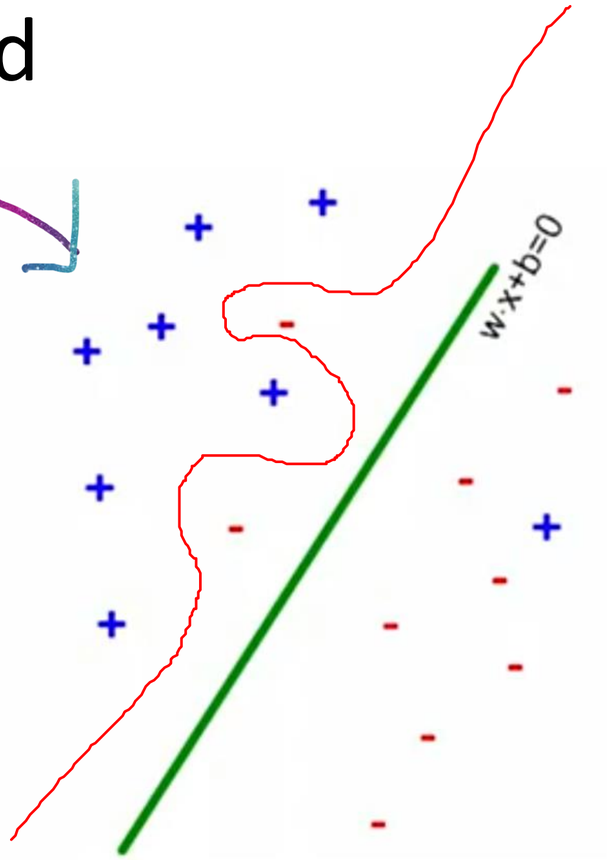
Soft Margin SVM

Recap: Objective function for Hard margin SVM

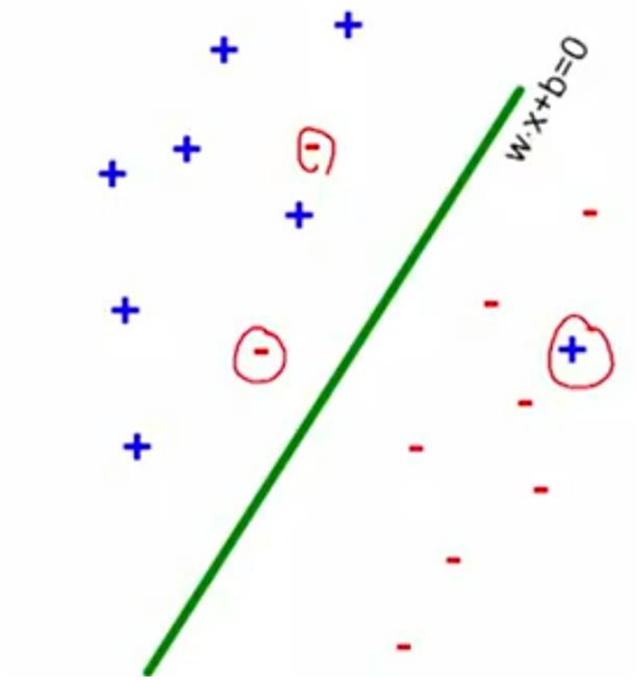
$$\mathcal{J} = \min \frac{\|w\|^2}{2} \quad s.t. \forall i, \;\; y^{(i)}(w^T x^{(i)} + b) \geq 1$$



- Welcome to the real world
  - Datasets are noisy
  - Not linearly separable



Introduce penalty for mistakes

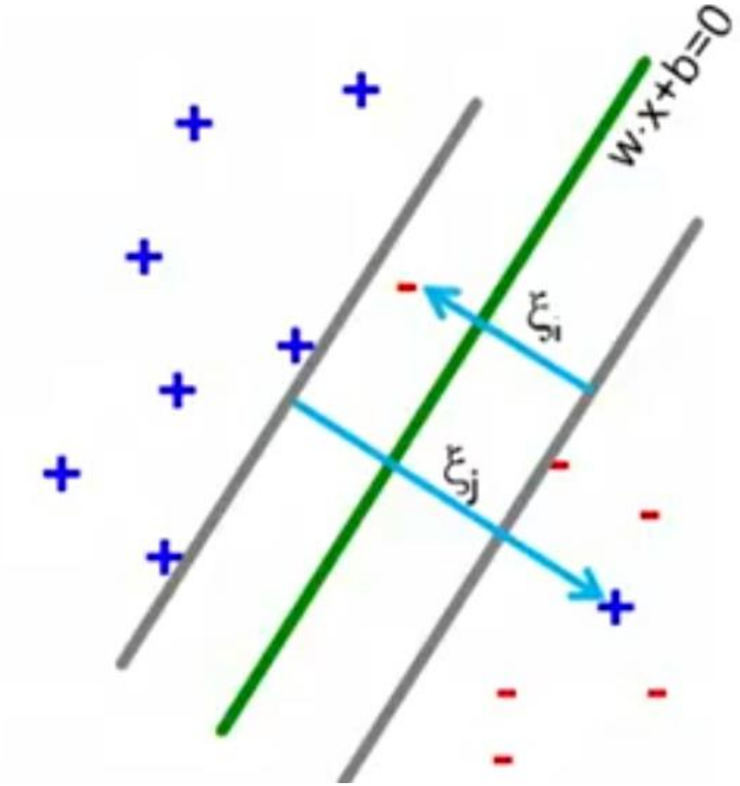


$$\mathcal{J}(w) = \min_{w,b} \frac{\|w\|^2}{2} + C \times number\ of\ mistakes$$

$$s.t. \forall i, \quad y^{(i)}(w^T x^{(i)} + b) \geq 1$$

- Find w such that number of mistakes is small
- C is determined by cross validation
- Penalizing mistakes – Not all mistakes are equally bad

Quantifying penalty for mistakes



- By introducing notion of slack variable ξ_i

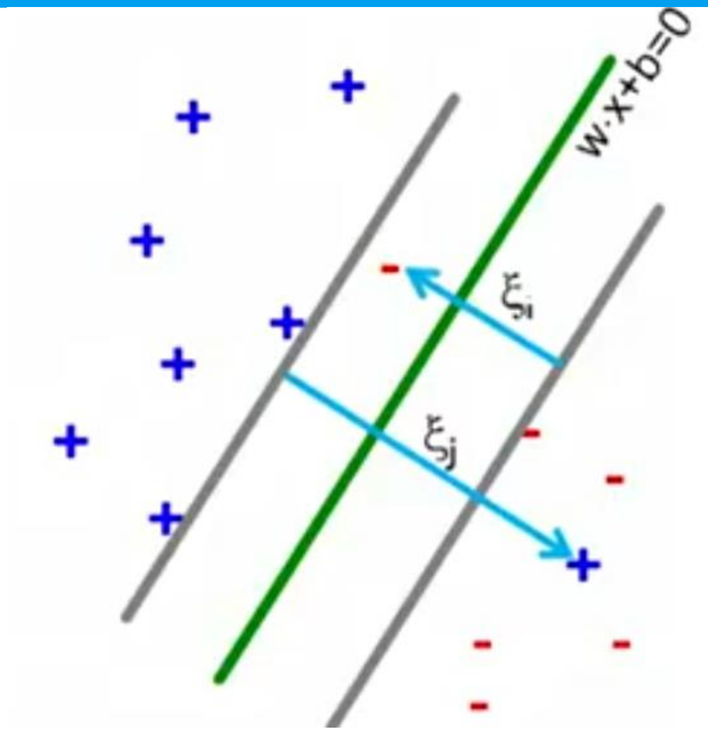
$$\mathcal{J}(w) = \min_{w,b} \frac{\|w\|^2}{2} + C \sum_{i=1}^{m} \xi_i$$

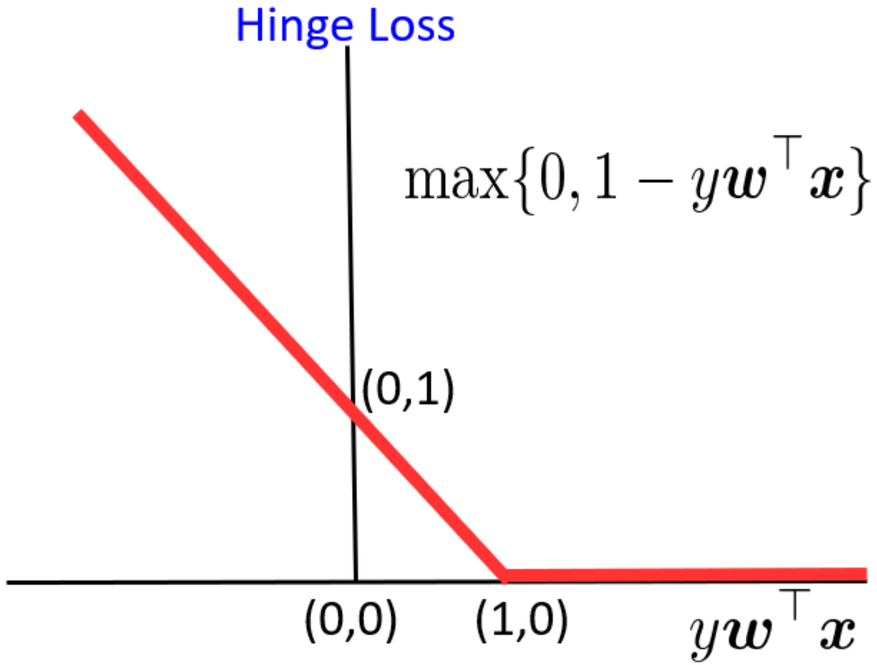
$$s.t. \forall i, \quad y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i$$

- If point is on the wrong side of the margin, penalty is non zero

- C = 0, no regularization
- C = large, high amount of regularization

Generic equation for SVM



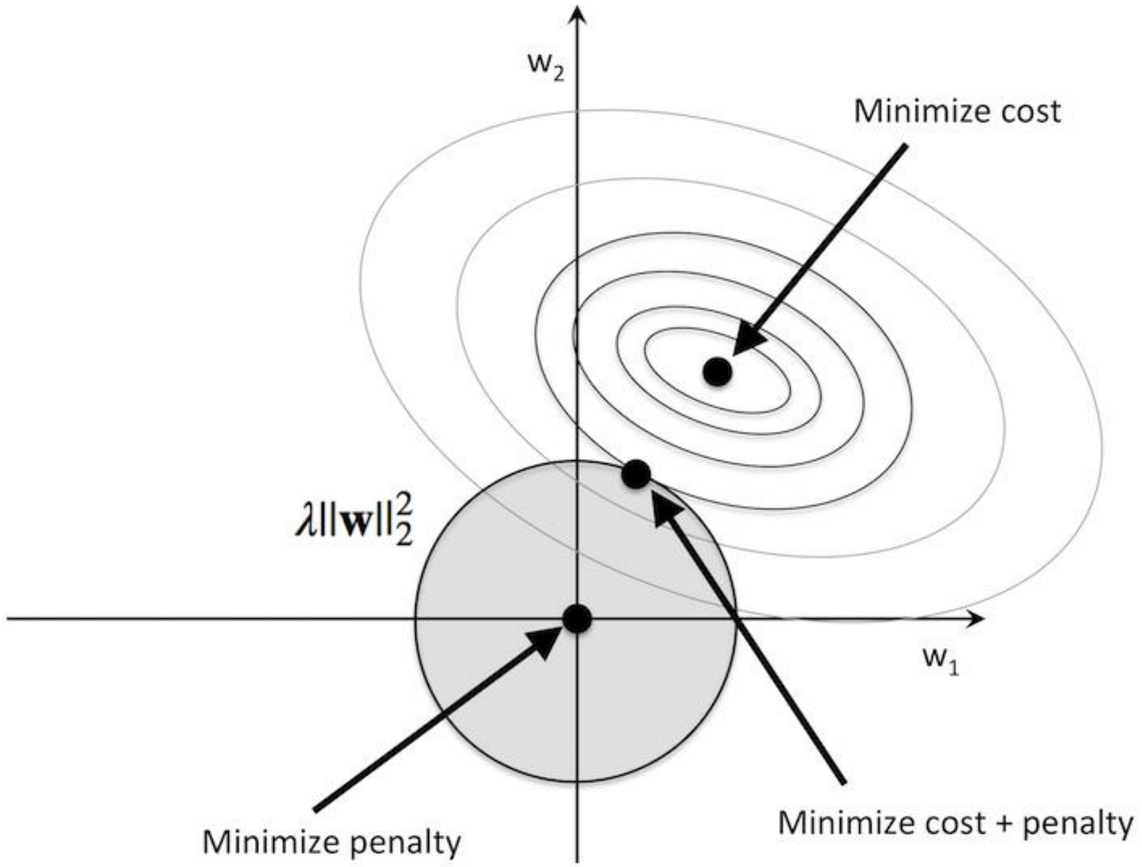


$$\mathcal{J}(w,b) = \min_{w,b} \frac{\|w\|^2}{2} + C\sum_{i=1}^{m} max(0, 1 - y^{(i)}(w^T x^{(i)} + b))$$

- Can perform gradient descent wrt w and b

SVM – Dual form

L2 Regularization in Linear Regression



$$\mathcal{J}(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m}\left(\mathbf{w}^T x^{(i)} - y^{(i)}\right)^2$$

Lagrange Multiplier

$$\nabla_w \mathcal{J} = \lambda \nabla_w \|w\|^2$$

Objective function in Lagrangian notation

$$\mathcal{L}(w, \lambda) = \nabla_w \mathcal{J} + \lambda \nabla_w \|w\|^2 = 0$$

$$\mathcal{J}(w) = \frac{1}{m}(Xw - y)^T(Xw - y)$$

J is a function of w

L is a function of w and lambda

SVM objective function in Lagrangian form

$$\mathcal{J}(w) = \min_{w,b} \frac{\|w\|^2}{2}$$

$$s.t. \forall i, \quad y^{(i)}(w^T x^{(i)} + b) \geq 1$$

Original Problem in Primal form

m separate constraints

$$\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ ..\alpha_m \end{bmatrix}$$

Original problem in Lagrangian notation

No more separate constraints

$$\mathcal{L}(w, b, \alpha) = \min \left[\frac{w^T w}{2} - \sum_{i=1}^{m} \alpha^{(i)} y^{(i)} (w^T x^{(i)} + b) - 1 \right]$$

SVM objective function in dual form

Problem in primal form (Lagrangian notation)

$$\mathcal{L}(w, b, \alpha) = \min_{w,b} \left[\frac{w^T w}{2} - \sum_{i=1}^{m} \alpha^{(i)} y^{(i)} (w^T x^{(i)} + b) - 1 \right]$$

Optimization in column dimension

Equivalent to

$$\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ ..\alpha_m \end{bmatrix}$$

Problem in dual form

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

Optimization in row dimension

Solving SVM objective function in dual form

Problem in dual form

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

$$\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ ..\alpha_m \end{bmatrix}$$

$$\frac{\partial \mathcal{L}}{\partial w} = 0 \implies w = \sum_{i=1}^{m} \alpha^{(i)} y^{(i)} x^{(i)}$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \implies b = \alpha^T y$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ ..y_m \end{bmatrix}$$

Substitute for w and b in this

$$\mathcal{L}(w, b, \alpha) = \min_{w,b} \left[\frac{w^T w}{2} - \sum_{i=1}^{m} \alpha^{(i)} y^{(i)} (w^T x^{(i)} + b) - 1 \right]$$

Solving SVM objective function in dual form

Substitute w and b

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

$$\max_{\alpha^{(i)} \geq 0} \left[\sum_{i=1}^{m} \alpha^{(i)} - \frac{1}{2} \sum_{i,j} \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} {x^{(i)}}^T x^{(j)} \right]$$

$$\min_{\alpha^{(i)} \geq 0} \left[\frac{1}{2} \sum_{i,j} \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} {x^{(i)}}^T x^{(j)} - \sum_{i=1}^{m} \alpha^{(i)} \right]$$

Why solve in dual form?

$$\min_{\alpha^{(i)} \geq 0} \left[\frac{1}{2} \sum_{i,j} \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} {x^{(i)}}^T x^{(j)} - \sum_{i=1}^{m} \alpha^{(i)} \right]$$

- Why dual?
- Solve in single variable vector alpha
- Most alpha are zero
- For wide data sets p >> m
- mp >> m^2
- Kernel friendly
- kernels can be solved only in dual form

$$\mathcal{J}(w) = \min_{w,b} \frac{\|w\|^2}{2}$$

$$s.t. \forall i, \quad y^{(i)}(w^T x^{(i)} + b) \geq 1$$

Further Reading

- SVM Kernels
- SVM polynomial, RBF kernels
  - Statquest by Josh Stammer (youtube)


QUESTIONS