SCS 3201

Machine Learning and Neural Computing

Kasun Gunawardana

UCSC

University of Colombo School of Computing

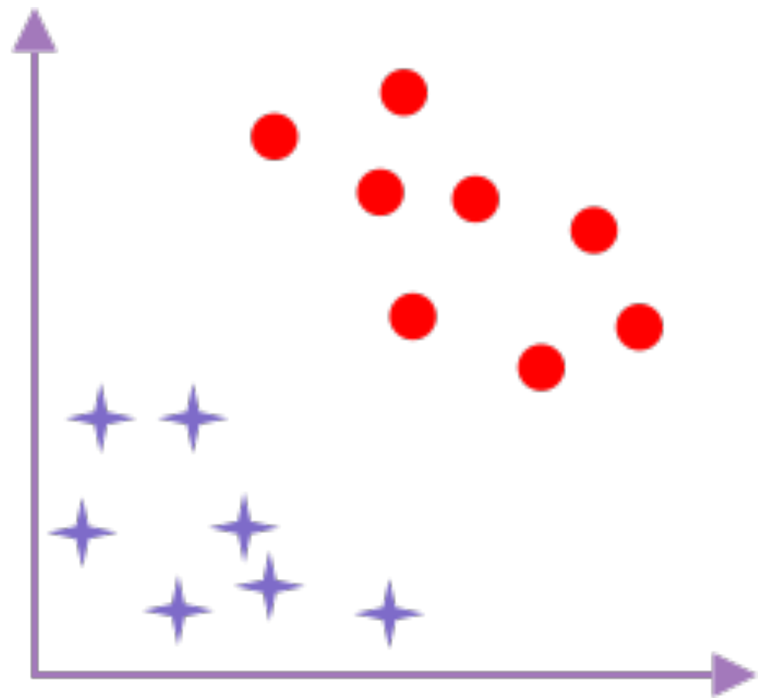
Support Vector Machine

Support Vector Machine (SVM)

- Supervised Machine Learning Algorithm
- Well known for Classification problems
- Can be applied for Regression problems
- Powerful tool in some contexts
  - Learning non-linear functions
  - Compared to Logistic Regression, ANN

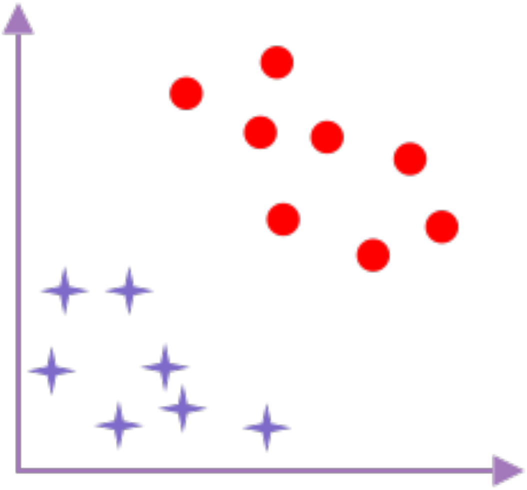
The Problem

- Assume we have a binary classification problem
  - Two dimensional input space



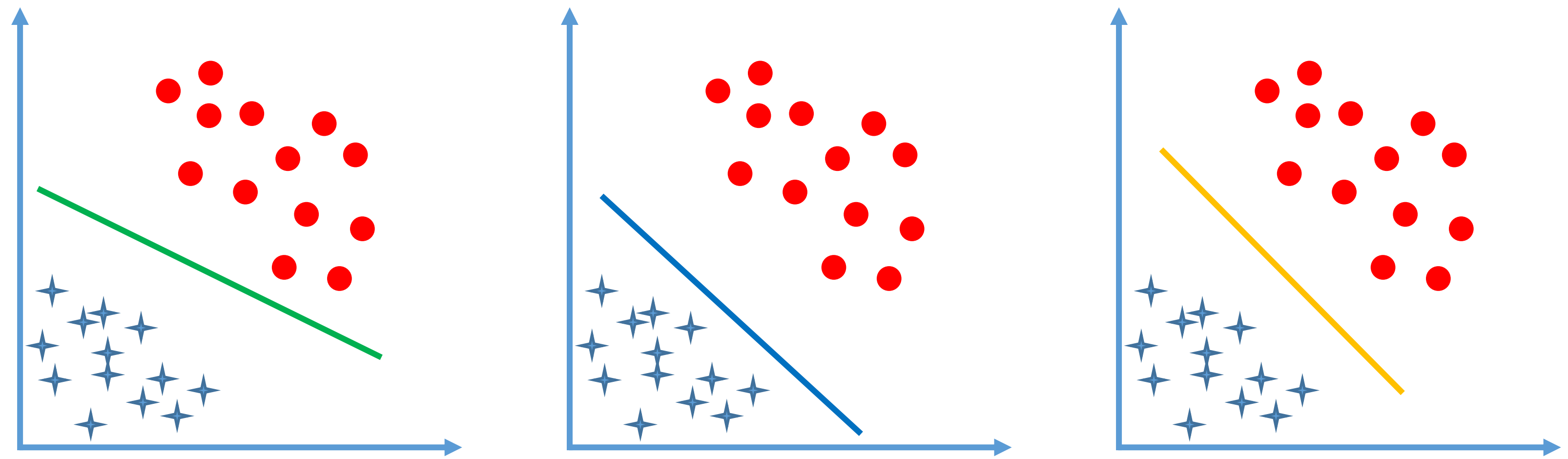
Optimal Decision Boundary

- Decision Boundary
  - More than one decision boundary is possible
- Optimal Decision Boundary
  - Separates two classes with the maximum margin
  - Generality of the model will be high



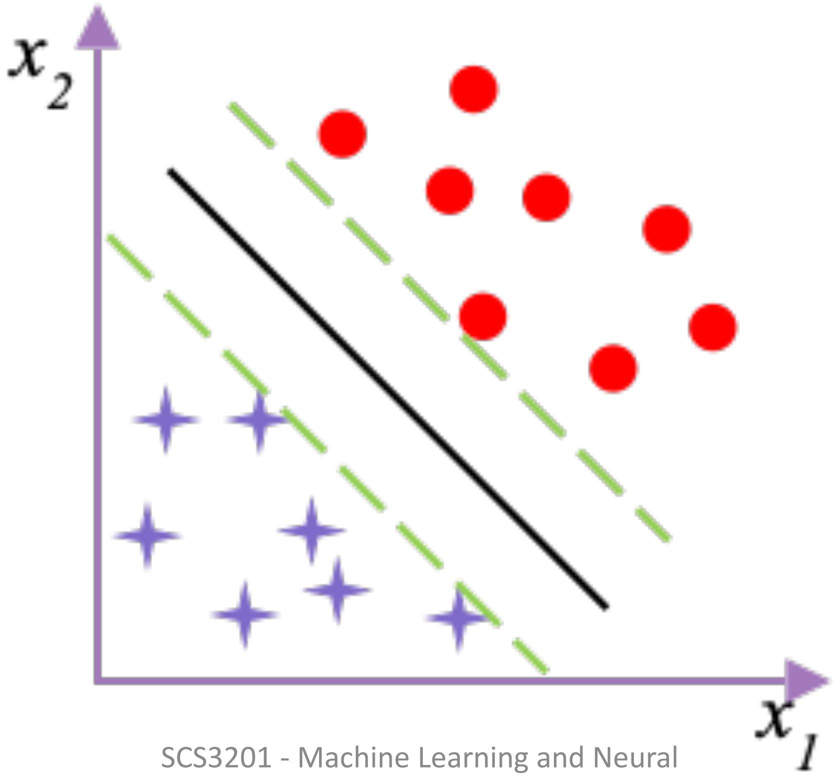
The Problem (Cont.)

- What is the optimal decision boundary we can have?



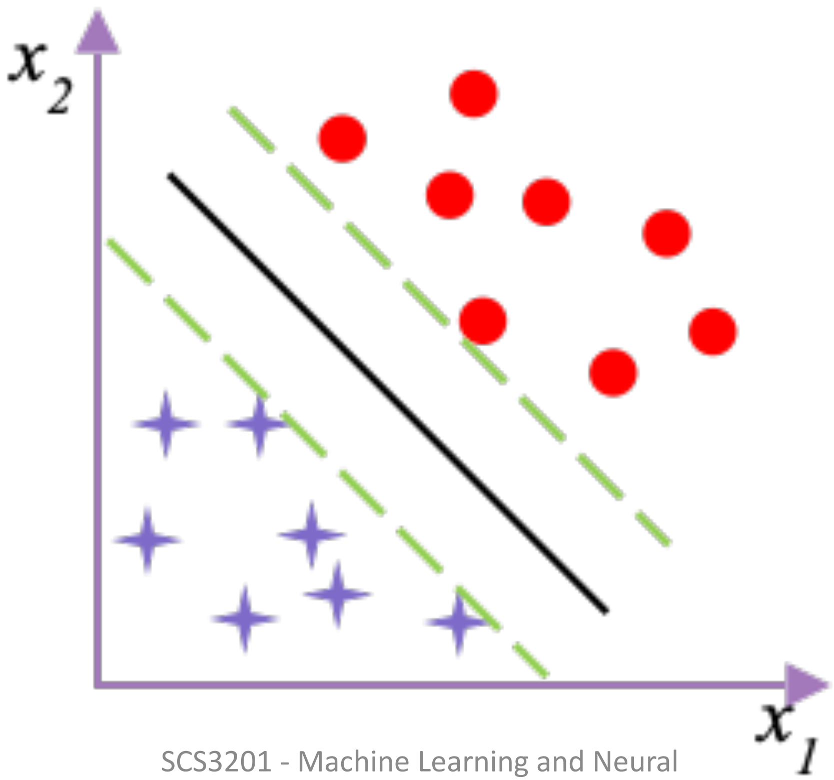
The Objective of SVM

- Find the decision boundary that gives the optimum margin in a binary classification problem.



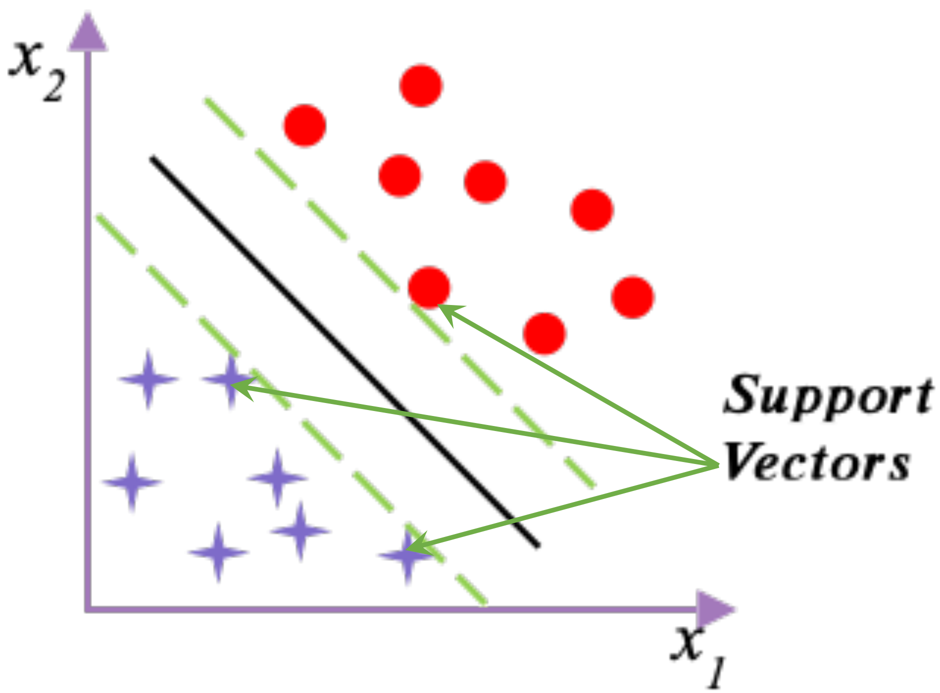
Support Vector Machine

- We can have an infinite number of hyperplanes.
- Select the hyperplane that maximizes the margin.



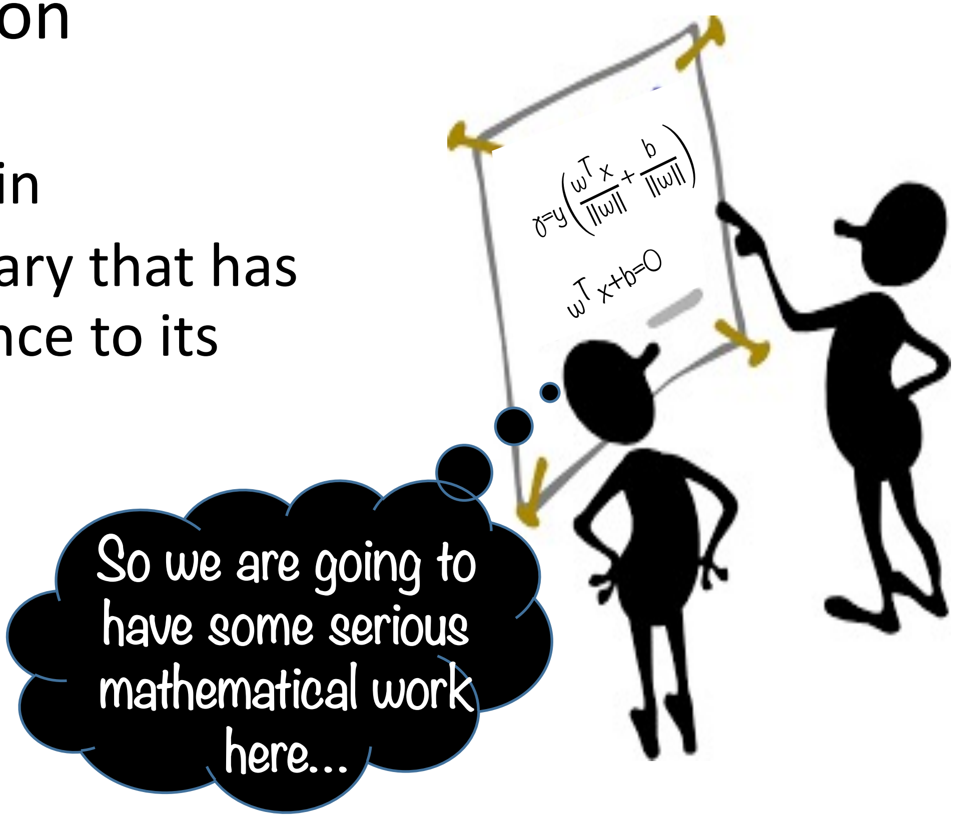
Support Vectors

- The data points that reside closest to the decision surface (or hyperplane)



Optimization Problem

- This is an Optimization Problem
  - The Optimum Margin
  - The decision boundary that has the maximum distance to its support vectors



Mathematical Basics

Hyperplane

- Hyperplane can be defined as

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

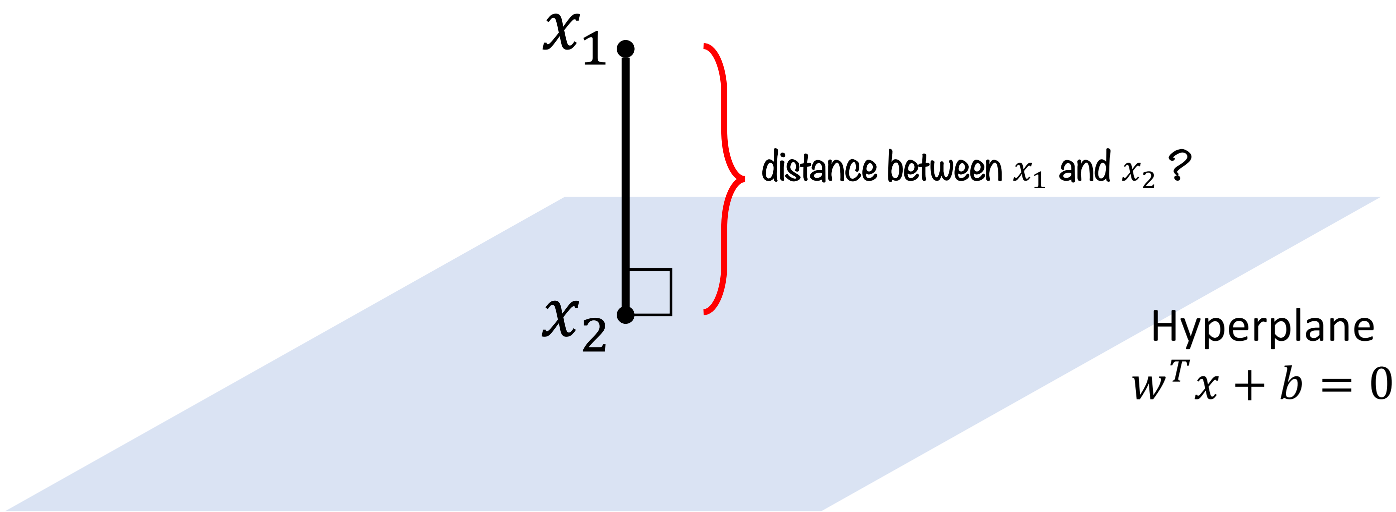
$$h(x) = \begin{cases} Class\ 1 & if\ w^T x + b > 0 \\ Class\ 2 & Otherwise \end{cases}$$

- Hyperplane can be declared if we find the values for $w\ and\ b$.

Decision Boundary is "Normal" to $\boldsymbol{w}$

- Assume a hyperplane $w^T x + b = 0$
- Let x_A and x_B are two points on the hyperplane
  - $w^T x_A + b = 0 \qquad (1)$
  - $w^T x_B + b = 0 \qquad (2)$
- The vector $(x_A - x_B)$ lies on the hyperplane
- From equation (1) and (2)
  - $w^T (x_A - x_B) = 0 \qquad (3)$
- $\therefore$ w is orthogonal to the decision boundary

Distance to the Hyperplane



Distance to the Hyperplane

- Let x_1 is a point **<u>on a side of the hyperplane</u>**
- Let x_2 is the closest point to x_1 on the hyperplane
- Then $(x_1 - x_2)$ is orthogonal to the hyperplane
- Therefore
  - $(x_1 - x_2) = cw$ $\qquad (4)$
- Since x_2 is on the hyperplane
  - $w^T x_2 + b = 0$ $\qquad (5)$
- From (4) and (5),
  - $\mathrm{c} = \frac{w^T x_1 + b}{\|w\|^2}$ $\qquad (6)$
- Substitute c in (4) for the normal distance to the hyperplane
  - $\|(x_1 - x_2)\| = \frac{w^T x_1 + b}{\|w\|}$ $\qquad (7)$

Simplification (Additional Slide)

- In (4), c is a constant. Therefore,

$$||(x_1 - x_2)|| = \mathrm{c}||w||$$

- From (6)

$$||(x_1 - x_2)|| = \frac{w^T x_1 + b}{||w||^2}\ ||w||$$

-

- Therefore,

$$||(x_1 - x_2)|| = \frac{w^T x_1 + b}{||w||}$$

Distance from the Origin

- Let x' is closest to the origin on the hyperplane
- Since x' is orthogonal to the hyperplane
  - $x' = kw$ $\qquad$ k is scalar $\qquad$ (8)
- Since $w^T x' + b = 0$,
  - $w^T(kw) + b = 0$ $\qquad$ (9)
  - $k\|w\|^2 + b = 0$ $\qquad$ (10)
  - $k = -\frac{b}{\|w\|^2}$ $\qquad$ (11)
- $\therefore \|x'\| = -\frac{b}{\|w\|^2}\|w\| = -\frac{b}{\|w\|}$ $\qquad$ (12)

An Optimization Problem

- Finding the maximum or minimum
  - Objective - function to be optimized
  - Variables - inputs
  - Constraints - limits on optimization

Optimization Problems (Cont.)

- Unconstrained Optimization

$$\min_{\theta} J(\theta)$$

$J(\theta)$ is called the obejective function

- Constrained Optimization
  - Single Equality Constraint

$$\min_{x} f(x)$$
$$s.t. \quad x^2 + 2x + 3 = 0$$

s.t. – Subject to

Optimization Problems (Cont.)

- Constrained Optimization
  - Multiple equality constraints

$$\min_{x} f(x)$$

$$s.t. \quad x^2 + 2x + 3 = 0$$

$$\sum_{i=1}^{n} x_i = 1$$

Optimization Problems (Cont.)

- Constrained Optimization
  - Inequality constraints

$$\min_{x} f(x)$$
$$s.t. \quad x_i \geq 0 \quad \forall i \in N$$

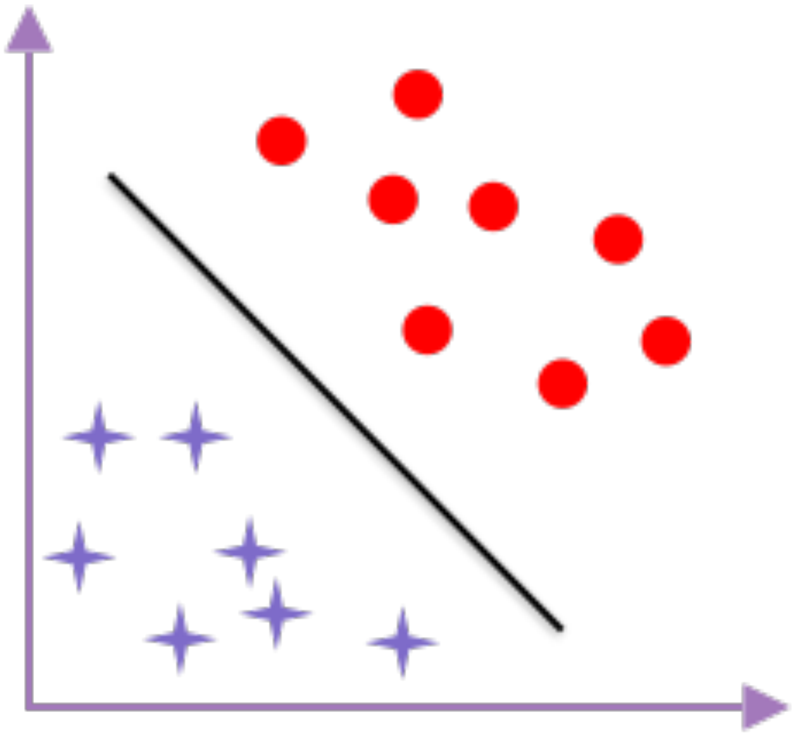
Next…

Margin Measurement

Margin Measurement

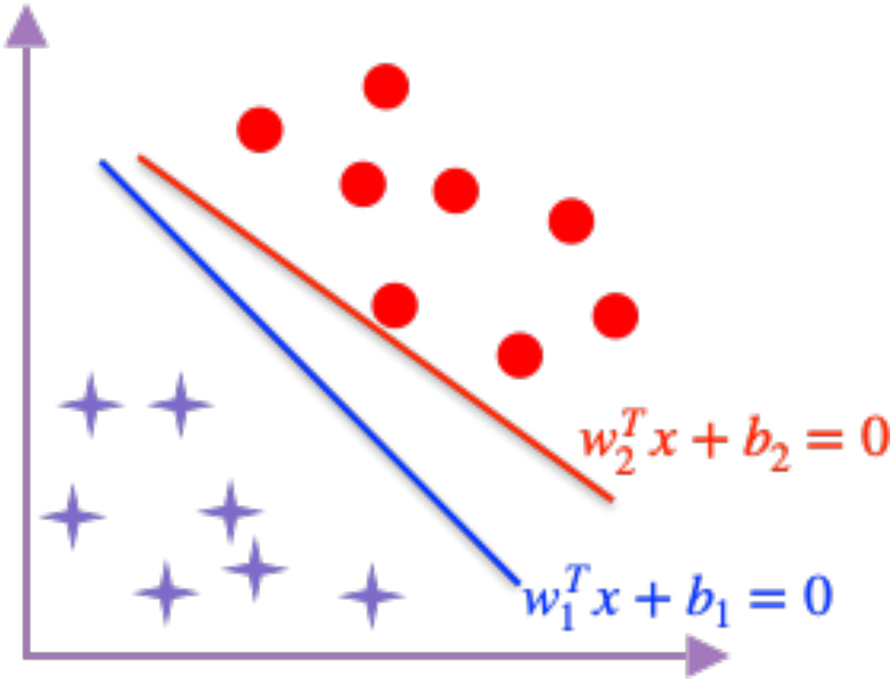
Assumption

- The data is linearly separable



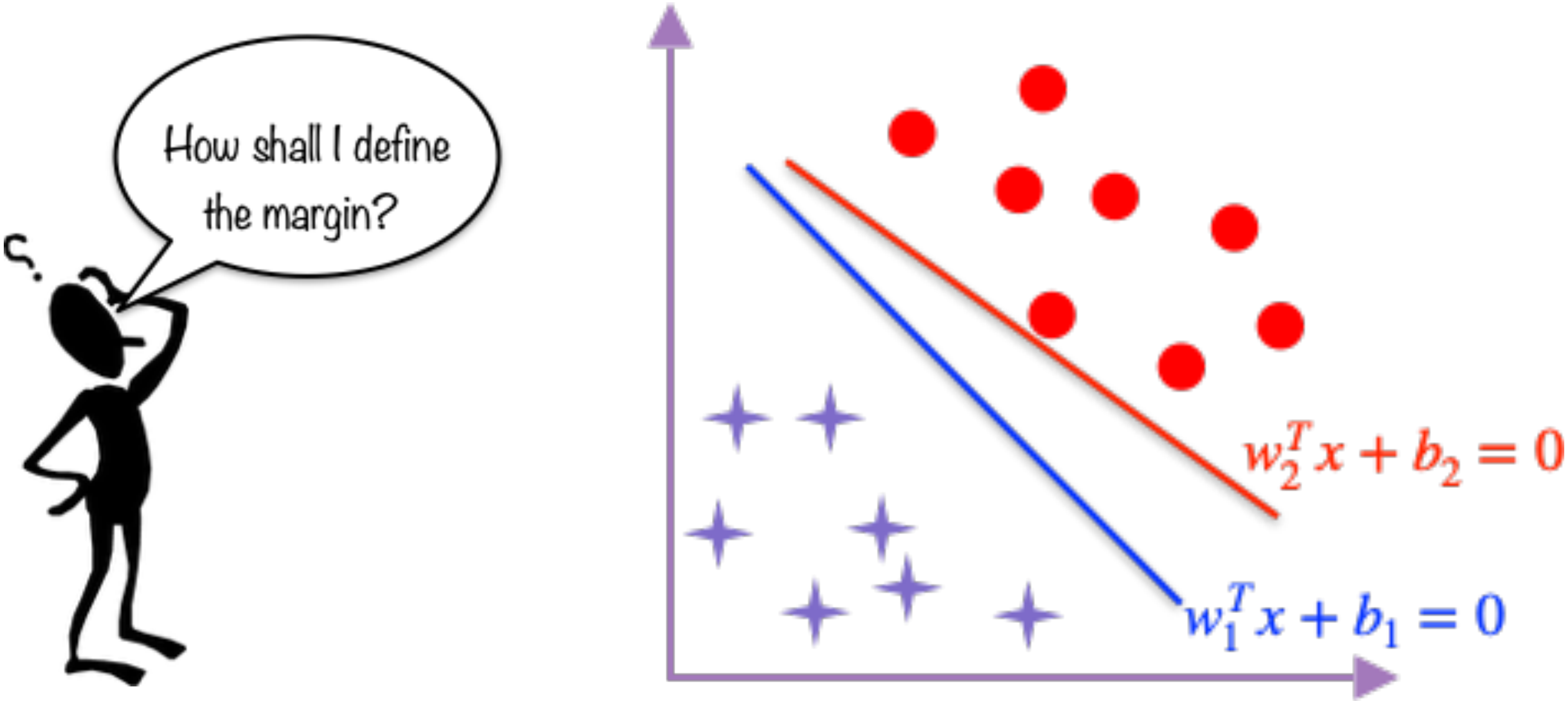
Hyperplane Comparison

- We need a measurement to compare hyperplanes and pick the best one.



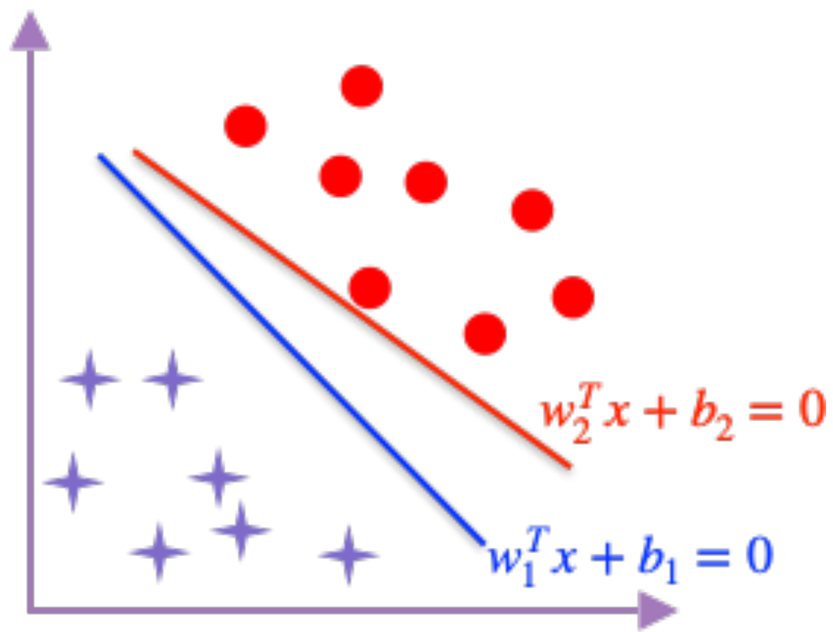
Intution: Margins

- Which hyperplane has the better “margin”?



Intuition : Margins (Cont.)

- Let's measure the distance from a datapoint to the hyperplane
- How far the closest datapoint (x_1, x_2) reside from the hyperplane?



Intuition : Margins (Cont.)

$$h(x) = \begin{cases} Class\ 1 & if\ w^T x + b > 0 \\ Class\ 2 & Otherwise \end{cases}$$

- We know that $w^T x + b = 0$ when a given datapoint x resides on the hyperplane.
- What if x is not on the hyperplane?
- What does value $h(x)$ say when x is not on the hyperplane?

Intuition : Margins (Cont.)

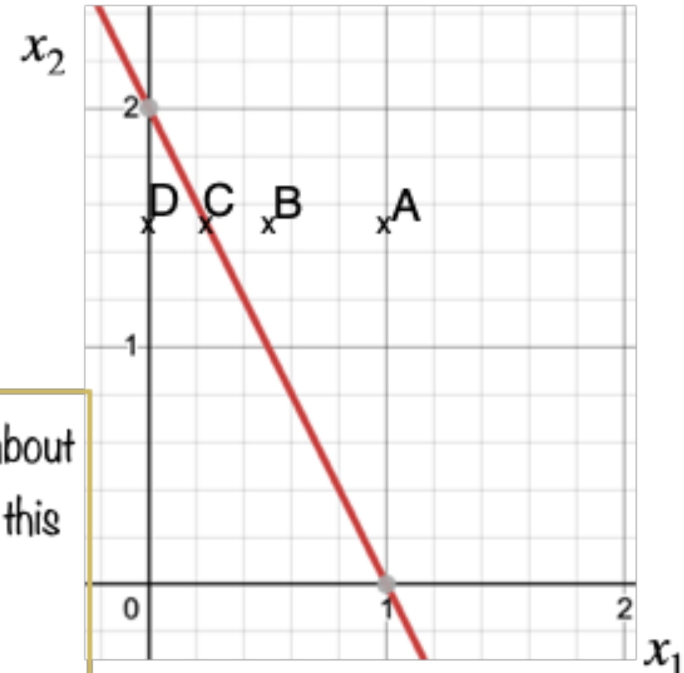
- $f(x) = w^T x + 2$
- $w = [-2, -1], x = [x_1, x_2]$
- A = [1, 1.5]
- B = [0.5, 1.5]
- C = [0.25, 1.5]
- D = [0, 1.5]

- $f(A) = -1.5$
- $f(B) = -0.5$
- $f(C) = 0$
- $f(D) = 0.5$

Thus, $w^T x + 2$ speaks about the distance. If we take this statement as α

$$\alpha = w^T x + 2$$

$$f(x_1, x_2) = -2x_1 - x_2 + 2$$



Intuition : Margins (Cont.)

- Let's define our distance measurement as α

$$\alpha = w^T x + b$$

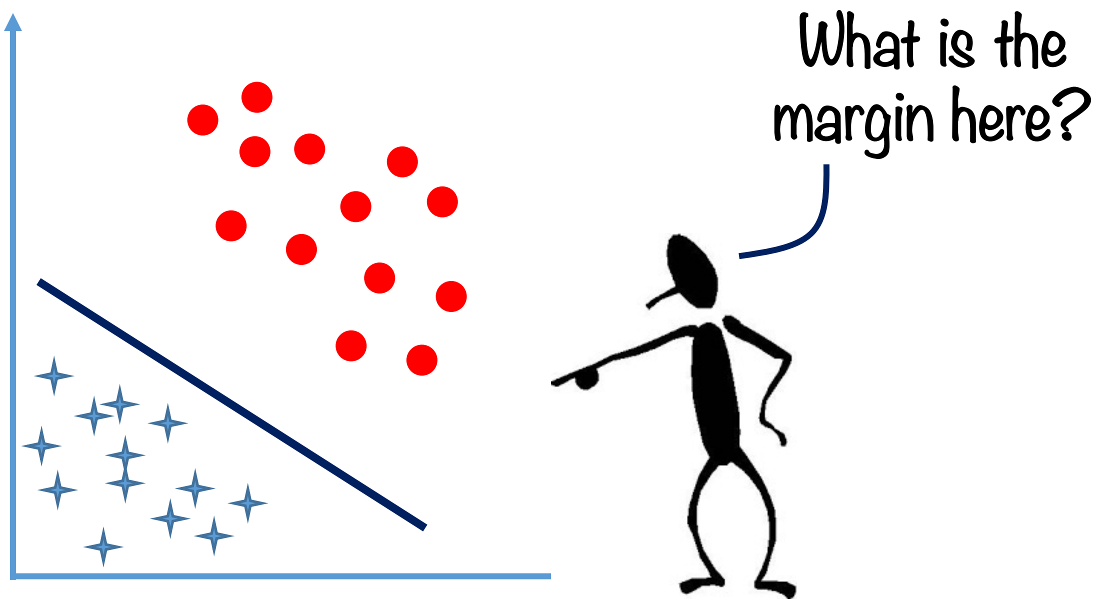
- If α is our distance measurement, then what defines the margin?

$$Margin = ?$$

How can we define the margin using the distance measurement?

Intuition : Margins (Cont.)

- If the dataset is with N samples
$$\{(x^1, y^1), (x^2, y^2), \dots, (x^N, y^N)\}$$



Intuition : Margins (Cont.)

- When the given dataset

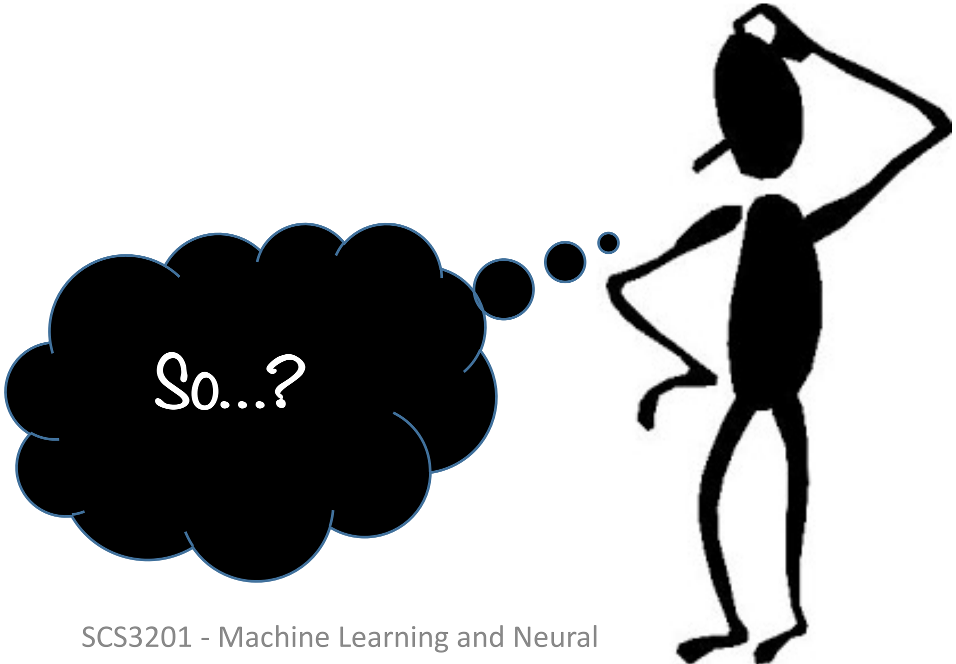
$$S = \{(x^i, y^i) | x^i \in \mathbb{R}^d, y^i \in \{+1, -1\}\}$$

- Margin is the minimum distance
  - Ex. Minimum α out of all α^i

$$Margin_\alpha = \min_{i=1..N} \alpha_i$$

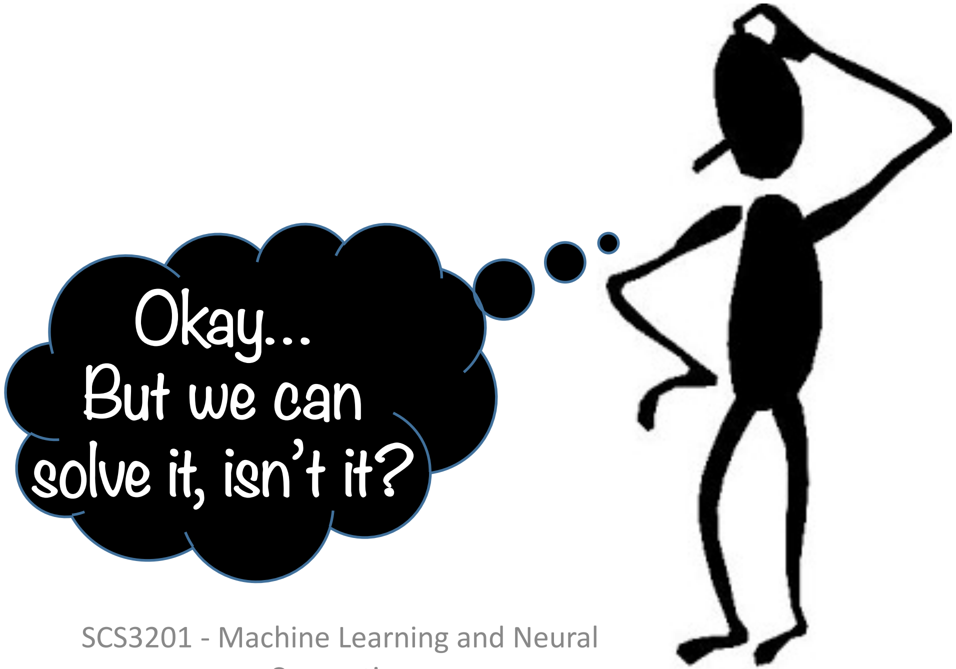
Problems: α as the Measurement

- α can be negative



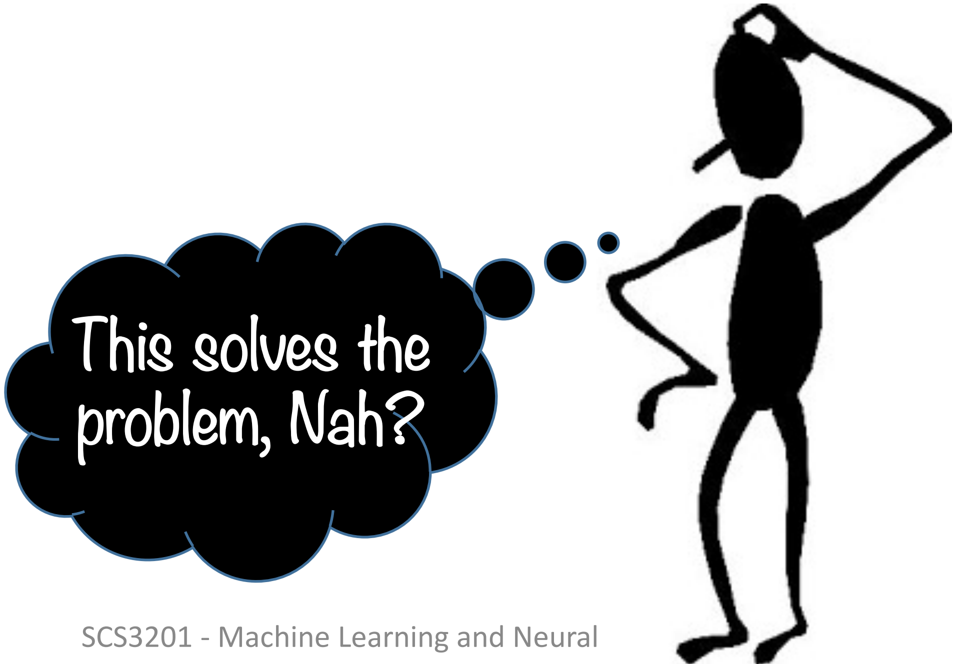
Problems: α as the Measurement

- α can be negative
  - Negative α will be selected over any positive value



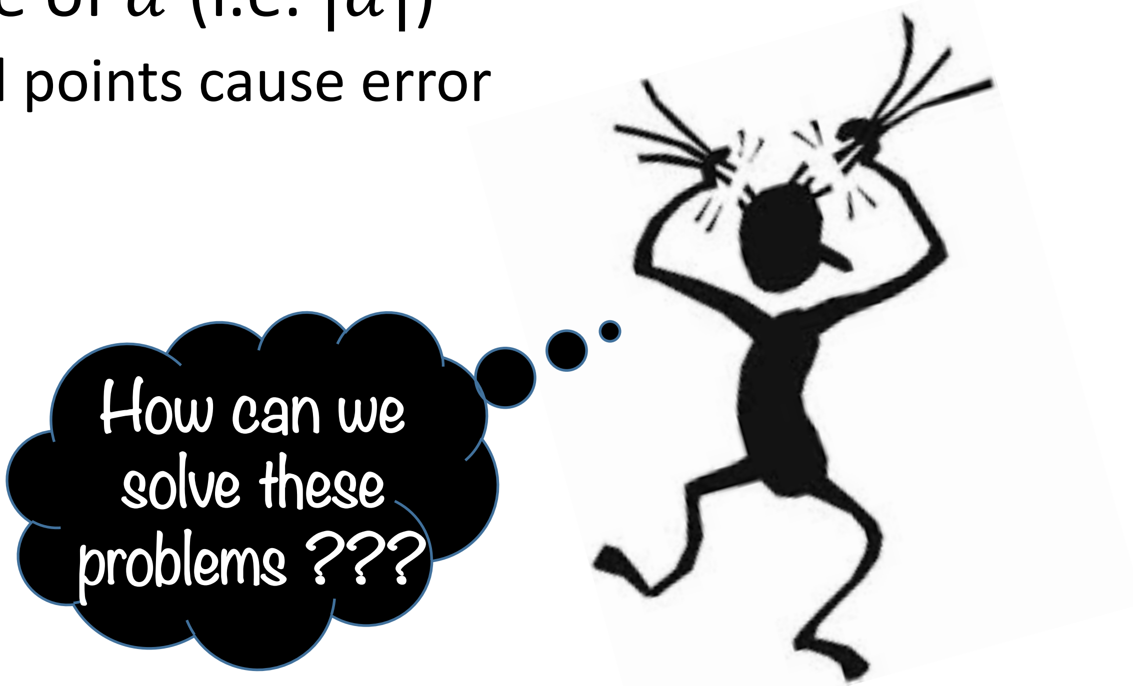
Problems: α as the Measurement

- α can be negative
  - Negative α will be selected over any positive value
- Absolute value of α (i.e. $|\alpha|$)



Problems: α as the Measurement

- α can be negative
  - Negative α will be selected over any positive value
- Absolute value of α (i.e. $|\alpha|$)
  - Misclassified points cause error



Functional Distance - β

- Define our distance measurement as,

$$\beta^i = y^i(w^T x^i + b)$$

- Compare β against α

$$\alpha^i = w^T x^i + b$$

- β is called $Functional\ Distance$ in this context

Functional Distance - β (Cont.)

- Because of the way the *Functional Distance* has been defined, it is easier if we define our class labels +1 and -1.
- Positive Case: When $w^T x^i + b > 0$

$$\beta^i = +1 \times (w^T x^i + b)$$

- Negative Case: When $w^T x^i + b < 0$

$$\beta^i = -1 \times (w^T x^i + b)$$

Have we found it ?

Have we found it ?

- Functional Distance is not "$scale - invariant$"
- Assume two hyperplanes A and B, defined as,
  - $w^T x + b = 0$ -- hyperplane A
  - $5(w^T x + b) = 0$ -- hyperplane B
- For a given data point i, β_B^i will be five times larger than β_A^i

$$\beta_B^i = 5\beta_A^i$$

- But A and B are the same hyperplane

Geometric Distance - γ

- We can make it scale invariant by defining γ

$$\gamma = \frac{\beta}{||w||} = y\left(\frac{w^T x}{||w||} + \frac{b}{||w||}\right)$$

- Now the Margin can be measured

$$Margin_\gamma = \min_{i=1..N} \gamma^i$$

- This is called the “Geometric Margin” - $Margin_\gamma$

Important Note..!

- Functional Distance

$$\beta = y(w^T x + b)$$

- Geometric Distance

$$\gamma = y\left(\frac{w^T x}{\|w\|} + \frac{b}{\|w\|}\right) = \frac{\beta}{\|w\|}$$

We can scale up or down β by appropriately selecting w and b as we wish, without changing γ

Next…

Margin Optimization

Margin Optimization

Margin Selection

- How do we pick the better hyperplane using margins?
- **Question**: If hyperplanes A and B have M^A and M^B respectively, what is the better hyperplane, as our decision boundary ?
- **Answer**:

Margin Selection

- How do we pick the better hyperplane using margins?
- **Question**: If hyperplanes A and B have M^A and M^B respectively, what is the better hyperplane, as our decision boundary ?
- **Answer**: We have to pick the hyperplane with larger margin.

SVM: Optimal Hyperplane

- Select the hyperplane which gives the largest margin for the given dataset S.
- This is an optimization (maximization) problem.
  - Maximize margin M with respect to w and b

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

SVM: Optimization Problem

- SVM Optimization Problem can be defined as,

$$\max_{w,b} \gamma_{min}$$

$$s.t. \quad \gamma^i \geq \gamma_{min} \quad i = 1..N$$

Maximize the margin subject to given inequality $\gamma^i \geq \gamma_{min}$ for all $i = 1..N$

SVM: Optimization Problem (Cont.)

- SVM Optimization Problem can be defined as,

$$\max_{w,b} \gamma_{min}$$

$$s.t. \quad \gamma^i \geq \gamma_{min} \quad i = 1..N$$

Now we are going to simplify this optimization problem...

SVM: Optimization Problem (Cont.)

- We know that $\gamma = \frac{\beta}{\|w\|}$, so our optimization problem can be written as,

$$\max_{w,b} \gamma_{min}$$

$$s.t. \quad \frac{\beta^i}{\|w\|} \geq \frac{\beta_{min}}{\|w\|} \quad i = 1..N$$

SVM: Optimization Problem (Cont.)

- We can further simplify the constraint by removing $||w||$ from both sides,

$$\max_{w,b} \gamma_{min}$$

$$s.t. \quad \beta^i \geq \beta_{min} \quad i = 1..N$$

SVM: Optimization Problem (Cont.)

- We know that we can scale up or down β without affecting to γ. So here we scale $\beta_{min} = 1$ and it will not affect to the problem [see slide 41].

$$\max_{w,b} \gamma_{min}$$

$$s.t. \quad \beta^i \geq 1 \quad i = 1..N$$

SVM: Optimization Problem (Cont.)

- Since constraints are in Functional Distance form, we can change the objective function to the same form using $\gamma = \frac{\beta}{\|w\|}$

$$\max_{w,b} \frac{\beta_{min}}{\|w\|}$$

$$s.t. \quad \beta^i \geq 1 \quad i = 1..N$$

SVM: Optimization Problem (Cont.)

- Since we scaled $\beta_{min} = 1$, the problem becomes,

$$\max_{w,b} \frac{1}{\|w\|}$$

$$s.t. \quad \beta^i \geq 1 \quad i = 1..N$$

SVM: Optimization Problem (Cont.)

- Since we scaled $\beta_{min} = 1$, the problem becomes

Maximum of function $\frac{1}{||w||}$ is equal to the minimum of function $||w||$.
$\therefore$ the problem can be written as a minimization problem…

$$\max_{w,b} \frac{1}{||w||}$$
$$s.t. \quad \beta^i \geq 1 \quad i = 1..N$$

SVM: Optimization Problem (Cont.)

- Optimization problem in minimization form,

$$\min_{w,b} \ ||w||$$

$$s.t. \quad \beta^i \geq 1 \quad i = 1..N$$

SVM: Optimization Problem (Cont.)

- To transfer the problem to an easily solvable form,

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

$$s.t. \quad \beta^i \geq 1, \quad i = 1..N$$

SVM: Optimization Problem (Cont.)

- Final optimization problem,

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

$$s.t. \quad y^i\left(w^T x^i + b\right) - 1 \geq 0, \quad i = 1..N$$

SVM: Optimization Problem

- The optimization problem in SVM is a quadratic programming problem with inequality constraints.
- To find solutions here, the method of $Lagrange$ $multipliers$ is used.
- As the solution it will find a series of $Lagrange$ $multipliers$ $(\alpha^1, \alpha^2, \dots, \alpha^N)$.
- Typically, the most of these $\alpha^i = 0$ (inputs which are not support vectors)
- For all the support vectors, corresponding $\alpha^i > 0$.

α^i is not the distance measure we used earlier.

SVM: Optimization Problem

- Using series of α^i values, desired outcomes (y^i) and inputs, we can derive w.

$$w = \sum_{i=1}^{N} \alpha^i y^i x^i = \sum_{x^i \in SV} \alpha^i y^i x^i$$

- Thereafter, b can be dreived.
  - For any support vector,

$$y^i\left(w^T x^i + b\right) = 1$$

SVM Training Summary

- SVM solves this optimization problem with the given data set.
- Finds w and b which defines the optimal hyperplane.

Next…

Soft Margin SVM

Soft Margin SVM

Support Vectors and the Hyperplane

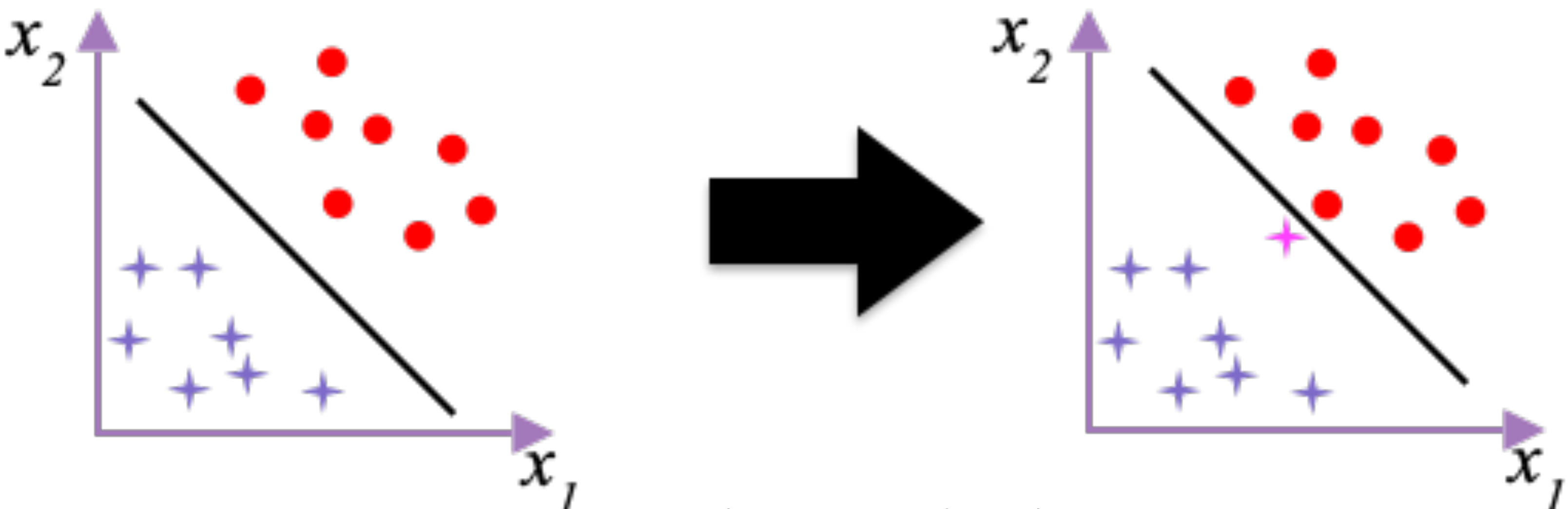
- Support vectors decides the decision boundary.
- Removal of support vector will change the decision boundary
- Removal of an ordinary vector will not change the decision boundary

Hard Margin SVM

- The model we formulated only works when the dataset is linearly separable.
- There are two types of outliers in this context
  - Narrow Margin Outliers
  - Misclassified Outliers

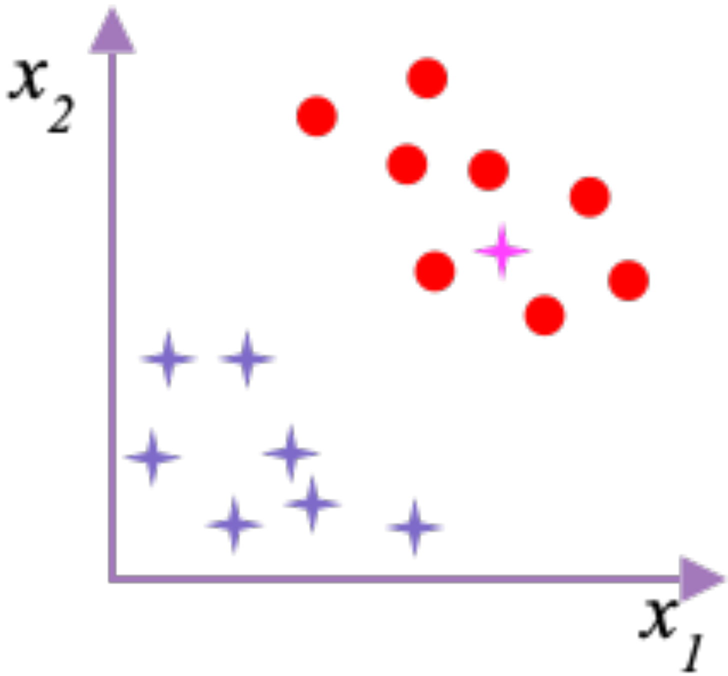
Narrow Margin Outliers

- Resides on the correct side
- Closer to the other class data
- Margin will be narrowed
- Loss of generality



Misclassified Outliers

- Classifier is incapable of finding a hyperplane.
- Single misclassification can cause this.



To Have a Soft Margin

Original Form

$$\min_{w,b} \frac{1}{2}\|w\|^2$$
$$s.t. \quad y^i(w^T x^i + b) - 1 \geq 0, \quad i = 1..N$$

- Soft Margin to avoid Hard Margin problems
- Modification to the constraint

$$y^i(w^T x^i + b) \geq 1 - \zeta^i$$

- Parameter Zeta (ζ) helps to recover from a misclassification

Soft Margin SVM

- By setting ζ large for every data point this constraint will be fulfilled
- To negate being large, we include ζ into the objective function as well

$$\min_{w,b} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{N} \zeta^i$$

$$s.t. \quad y^i(w^T x^i + b) \geq 1 - \zeta^i, \quad i = 1..N$$

Soft Margin

- Few more modifications to the Soft Margin version
- To prevent setting ζ negative, add $\zeta \geq 0$ constraint

  (large negative ζ will minimize the objective function)

- Added penalty $\sum_{i=1}^{N} \zeta^{i}$ is like regularization, to overall control all ζ parameters we add regularization parameter C to the objective function

SVM: Optimization for Soft Margin

- Now our optimization problem is defined as,

$$\min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\zeta^i$$

$$s.t. \quad y^i\left(w^T x^i + b\right) \geq 1 - \zeta^i$$

$$\zeta^i \geq 0, \quad i = 1..N$$

Thank You..!

Q&A