Support Vector Machines for Classification I

Swati Mishra

Applications of Machine Learning (4AL3)

Fall 2024

McMaster University

ENGINEERING

Review

- Multiclass classification
- Softmax Function
- Multinomial Logisitic Regression
- Accuracy, Precision, Recall, Sensitivity, Specifity

Classification Problem

- Diagnostic of Breast Cancer Wisconsin
- Classification goal to classify if a tissue is cancerous.

Variables Table

Variable Name	Role	Type
ID	ID	Categorical
Diagnosis	Target	Categorical
radius1	Feature	Continuous
texture1	Feature	Continuous
perimeter1	Feature	Continuous
area1	Feature	Continuous
smoothness1	Feature	Continuous
compactness1	Feature	Continuous
concavity1	Feature	Continuous
concave_points1	Feature	Continuous

Variables Table

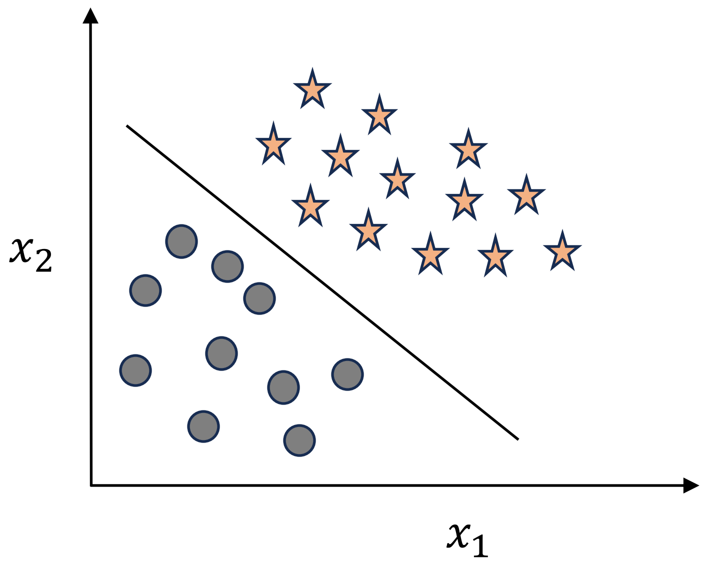
Variable Name	Role	Type
symmetry1	Feature	Continuous
fractal_dimension1	Feature	Continuous
radius2	Feature	Continuous
texture2	Feature	Continuous
perimeter2	Feature	Continuous
area2	Feature	Continuous
smoothness2	Feature	Continuous
compactness2	Feature	Continuous
concavity2	Feature	Continuous
concave_points2	Feature	Continuous

Variables Table

Variable Name	Role	Type
symmetry2	Feature	Continuous
fractal_dimension2	Feature	Continuous
radius3	Feature	Continuous
texture3	Feature	Continuous
perimeter3	Feature	Continuous
area3	Feature	Continuous
smoothness3	Feature	Continuous
compactness3	Feature	Continuous
concavity3	Feature	Continuous
concave_points3	Feature	Continuous

Concept of Hyperplane

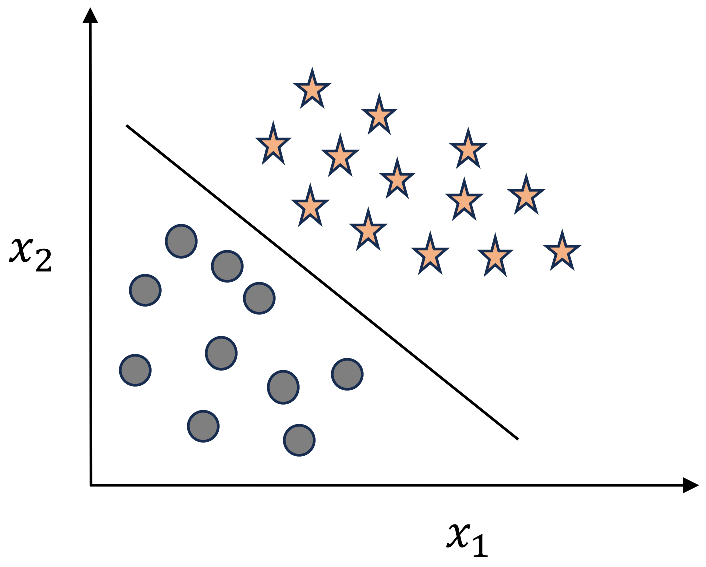
- In a p-dimensional space, a hyperplane is a flat affine subspace of $p-1$ dimension.



This is correct classification

Concept of Hyperplane

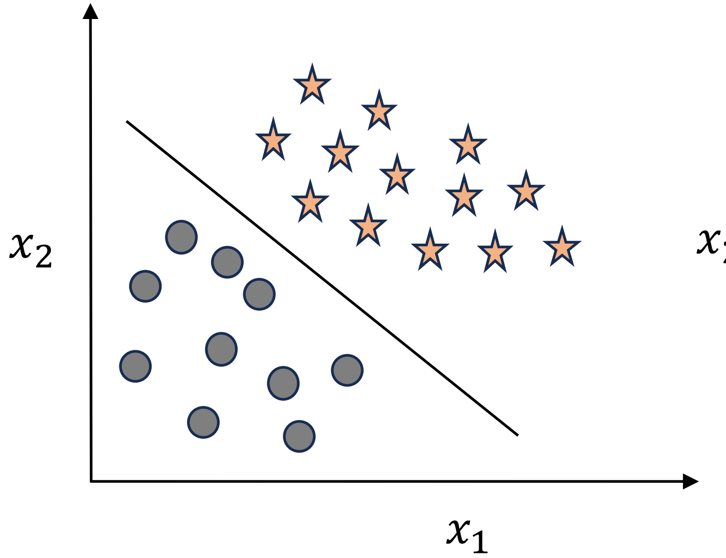
- In a p-dimensional space, a hyperplane is a flat affine subspace of $p-1$ dimension.
  - In two dimensions, a hyperplane is a flat one-dimensional subspace
  - In three dimensions, a hyperplane is a flat two-dimensional subspace—that is, a plane



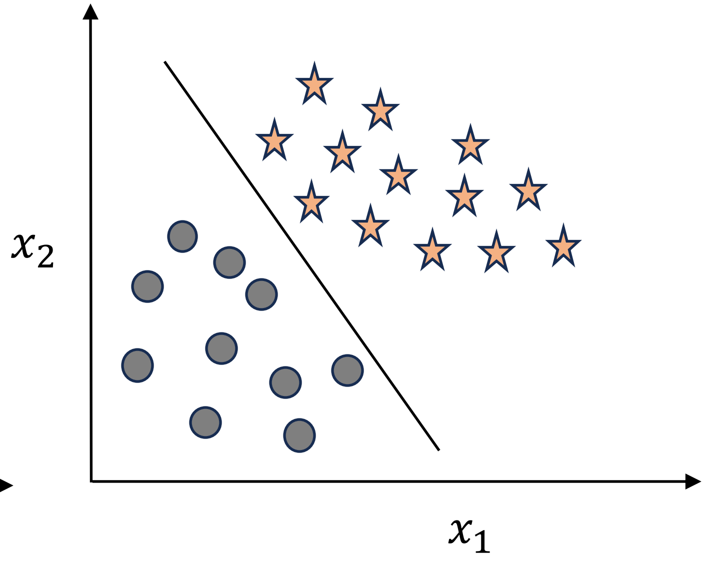
This is correct classification

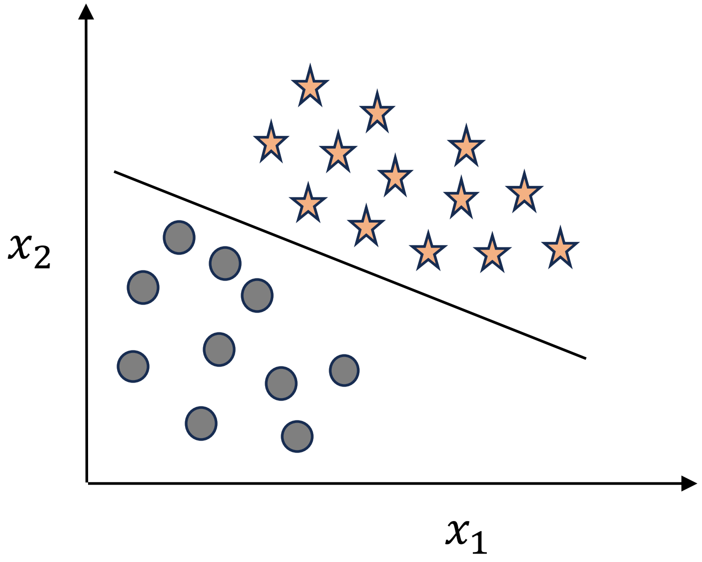
Concept of Hyperplane

- In a p-dimensional space, a hyperplane is a flat affine subspace of $p-1$ dimension.
  - In two dimensions, a hyperplane is a flat one-dimensional subspace
  - In three dimensions, a hyperplane is a flat two-dimensional subspace—that is, a plane



This is correct classification

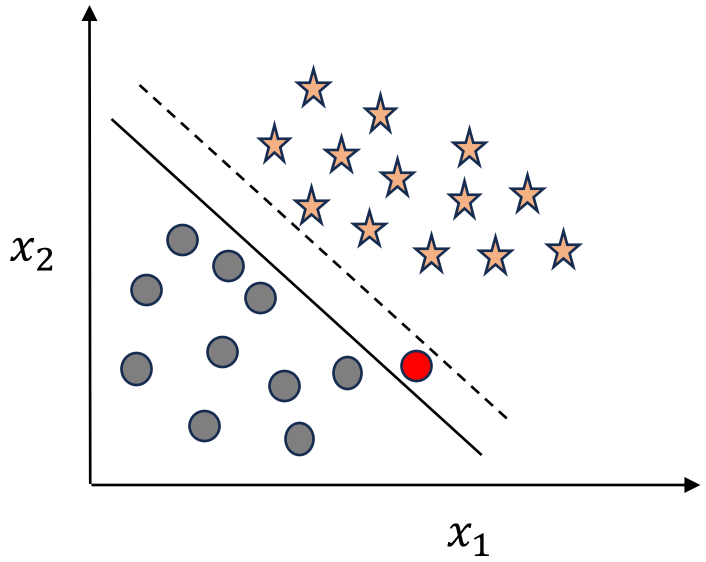


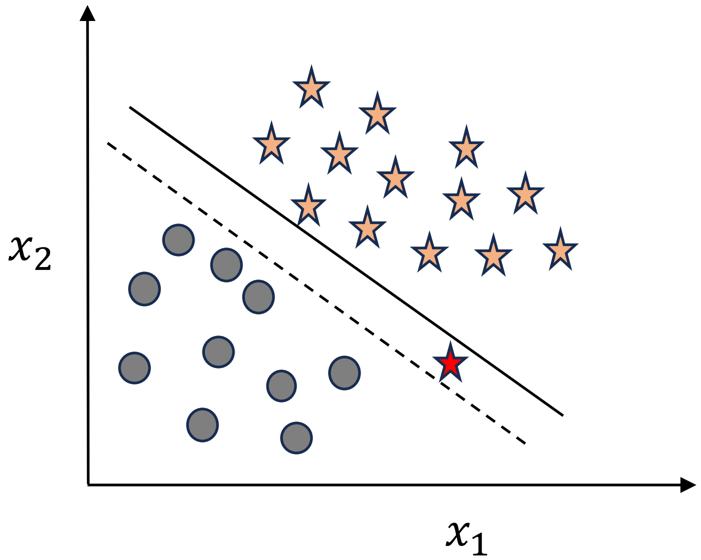


These are also correct classification

Concept of Hyperplane

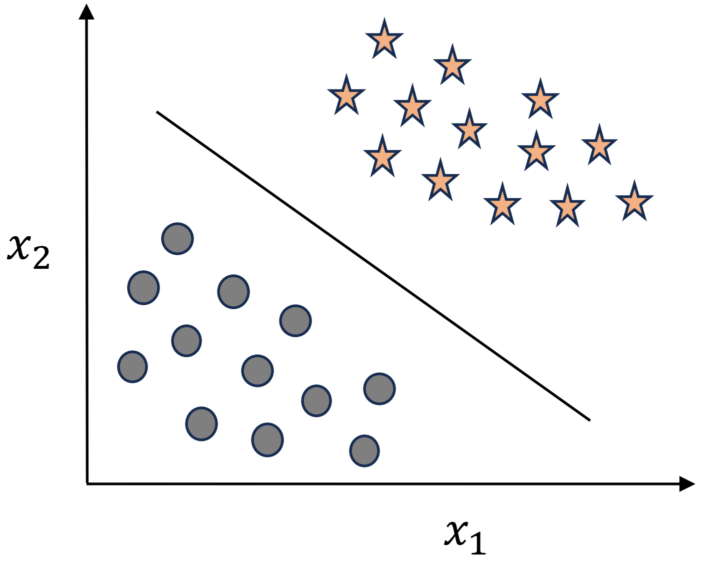
- When the hyperplane is too close to one of the classes, it fails to accurately classify new instances.





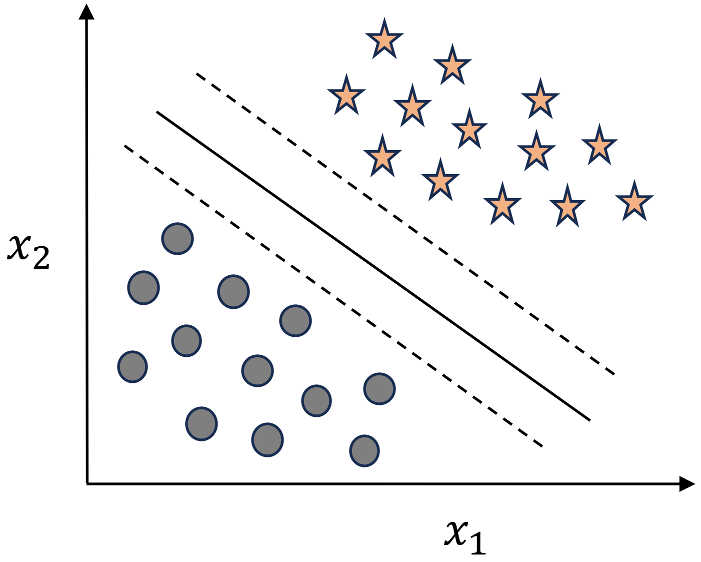
Support Vector Machine

- The goal of SVM which is to find the “separating hyperplane” that is furthest away from the training set



Support Vector Machine

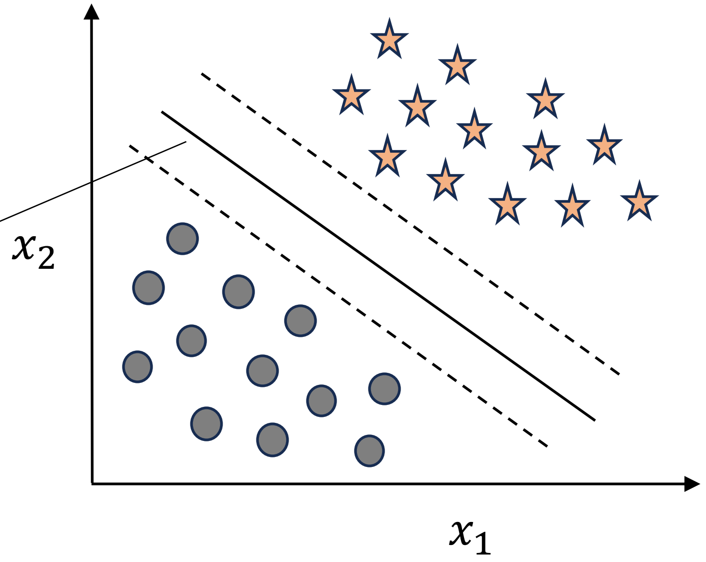
- The goal of SVM which is to find the “separating hyperplane” that is furthest away from the training set and defined by the optimal margin.



Support Vector Machine

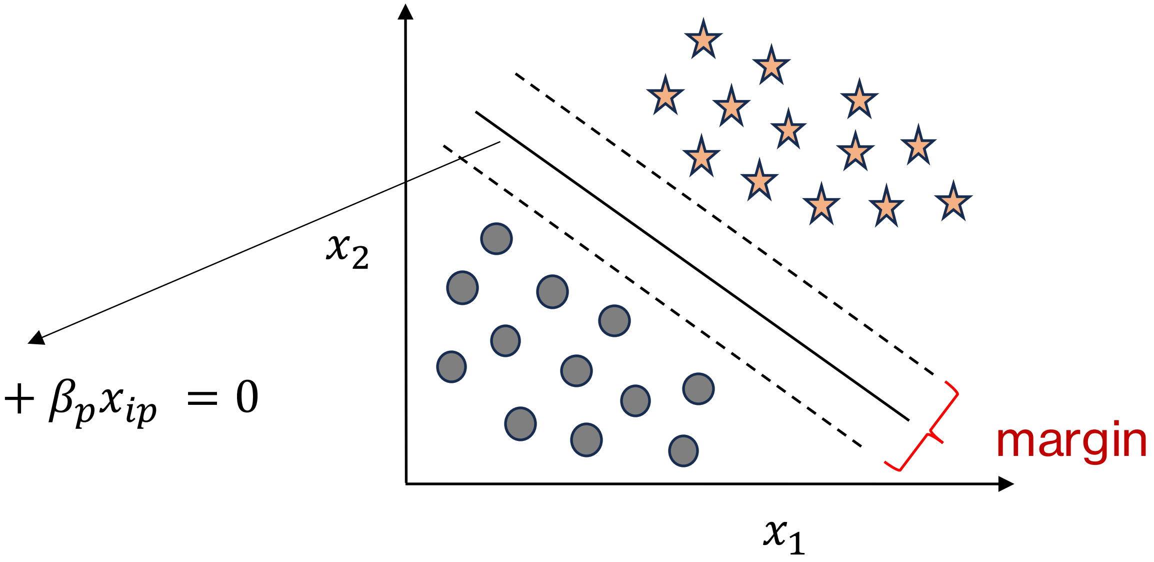
- The goal of SVM which is to find the “separating hyperplane” that is furthest away from the training set and defined by the optimal margin.

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = 0$$



Support Vector Machine

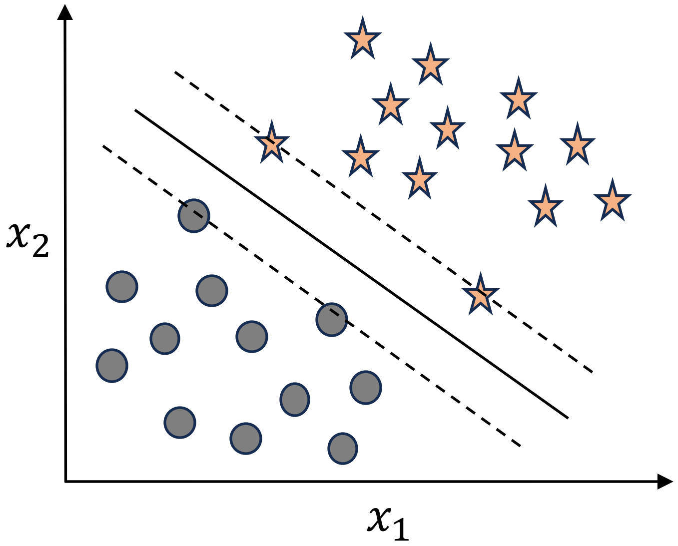
- The goal of SVM which is to find the “separating hyperplane” that is furthest away from the training set and defined by the optimal margin.
- SVM is the generalization of maximum margin classifier, where margin is the distance between the observations and the hyperplane.



$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = 0$$

Support Vector Machine

- A test observation is defined based on which side of the hyperplane it is.

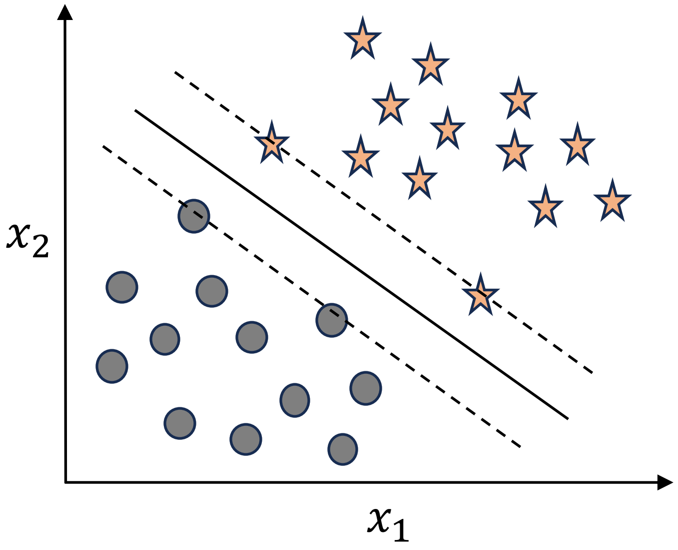


Support Vector Machine

- A test observation is defined based on which side of the hyperplane it is.

$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} > 0$ if $y_i = 1$ Class ☆

$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} > 0$ if $y_i = -1$ Class ●



Support Vector Machine

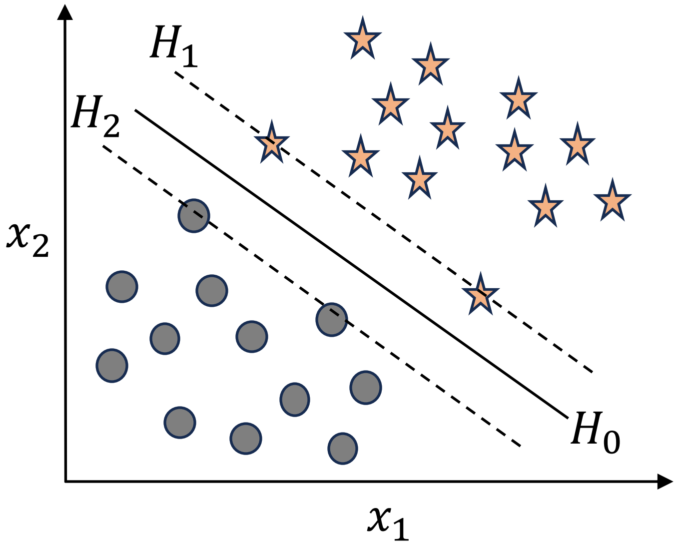
- A test observation is defined based on which side of the hyperplane it is.

Let's say there is a hyperplane H_1

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \geq 1 \text{ when } y_i = 1$$

Let's say there is hyperplane H_2

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \leq -1 \text{ when } y_i = -1$$



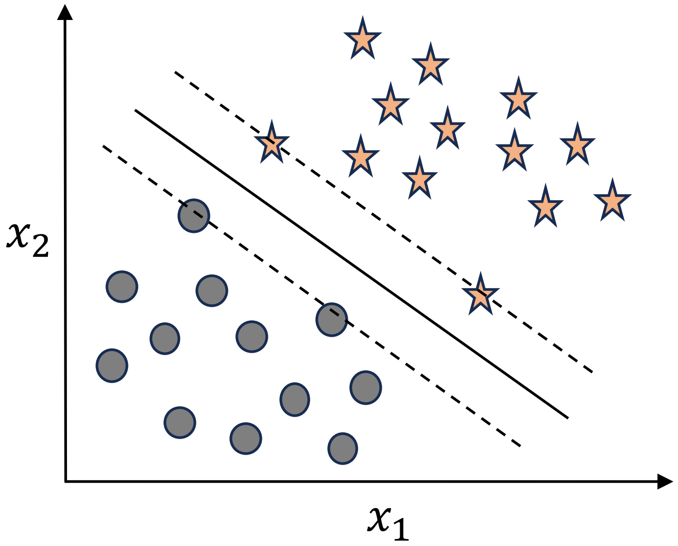
Support Vector Machine

- A test observation is defined based on which side of the hyperplane it is.

Let's define two of our hyperplanes

$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \geq 1$ when $y_i = 1$

$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \leq -1$ when $y_i = -1$



Support Vector Machine

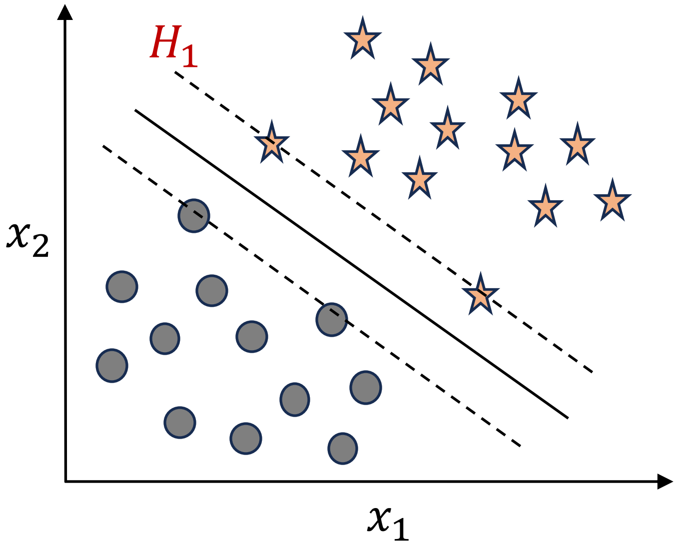
- A test observation is defined based on which side of the hyperplane it is.

Let's define two of our hyperplanes

$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \geq 1$ when $y_i = 1$

$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \leq -1$ when $y_i = -1$

Then hyperplane H_1 $\quad \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = 1$



Support Vector Machine

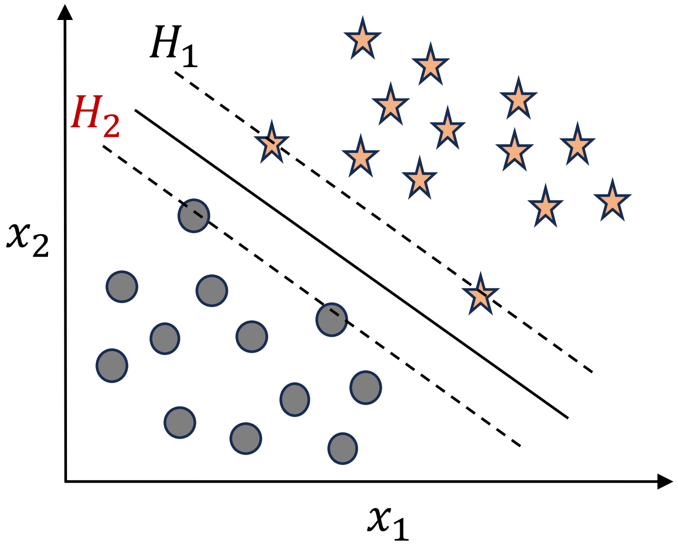
- A test observation is defined based on which side of the hyperplane it is.

Let's define two of our hyperplanes

$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \geq 1$ when $y_i = 1$

$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \leq -1$ when $y_i = -1$

Then hyperplane H_1 $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = 1$



Then hyperplane H_2 $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = -1$

Support Vector Machine

- A test observation is defined based on which side of the hyperplane it is.

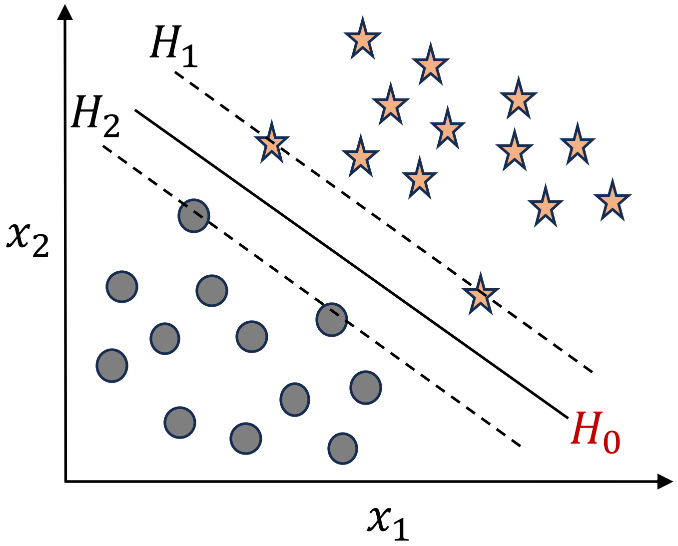
Let's define two of our hyperplanes

$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \geq 1$ when $y_i = 1$

$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \leq -1$ when $y_i = -1$

Then hyperplane H_1 $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = 1$

Then hyperplane H_2 $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = -1$



...and $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = 0$ is our H_0

Support Vector Machine

- A test observation is defined based on which side of the hyperplane it is.

Let's define two of our hyperplanes

$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \geq 1$ when $y_i = 1$

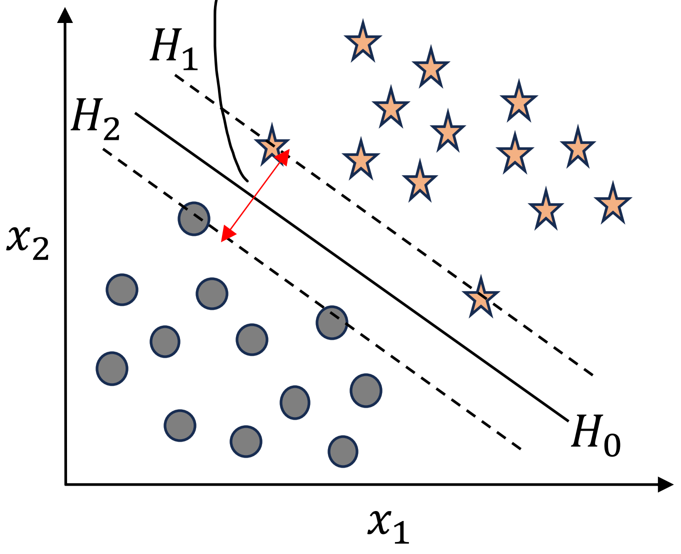
$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \leq -1$ when $y_i = -1$

Then hyperplane H_1 $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = 1$

Then hyperplane H_2 $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = -1$

...and $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = 0$ is our H_0

Training objective
widest possible gap



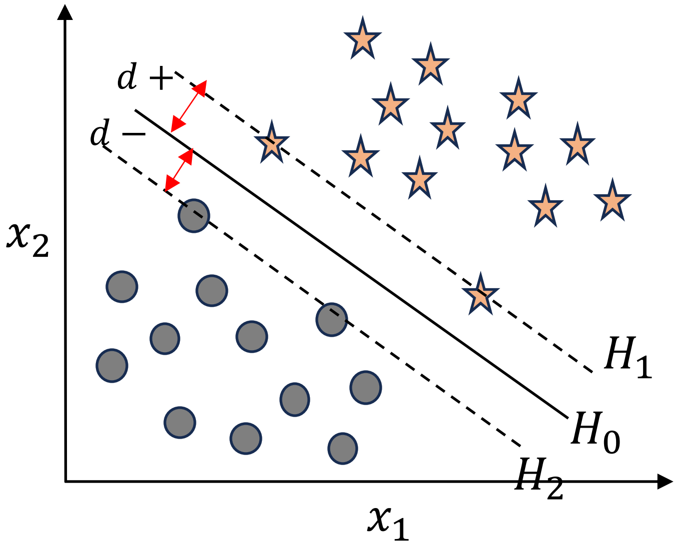
Support Vector Machine

- Some properties:

d^{+} It is the shortest distance closest to the positive point

d^{-} It is the shortest distance closest to the negative point

Margin = $d^{+} + d^{-}$

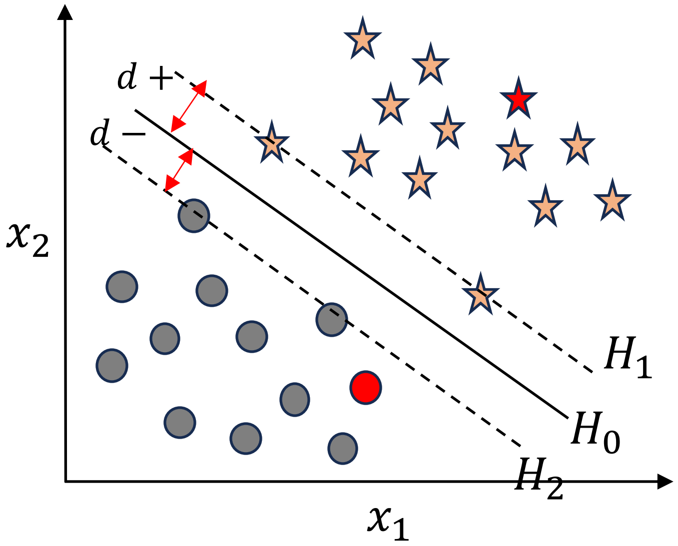


McMaster University

Support Vector Machine

- Some properties:

Moving other vectors has no effect on the margin

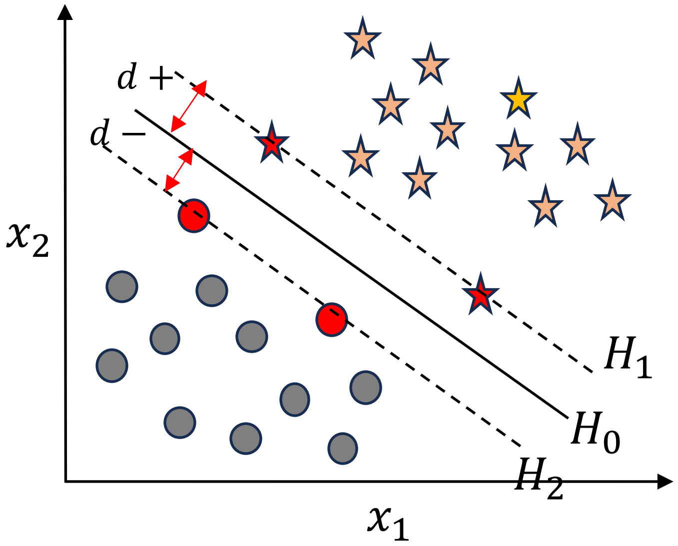


Support Vector Machine

- Some properties:

Moving other vectors has no effect on the margin

Moving support vectors has effect on the margin



Our optimization algorithm therefore must learn the weights in such a way that only the support vectors determine the weights and thus the boundary .

Optimization Problem

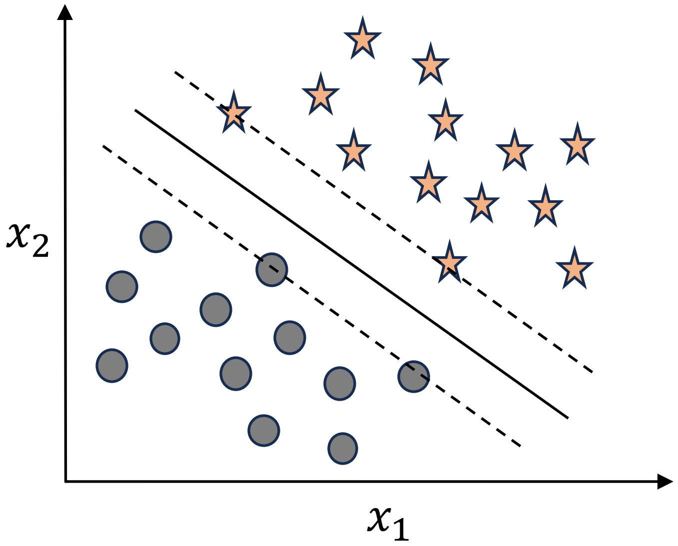
- The optimization goal of the SVM is:

$$\underset{\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots \epsilon_n, M}{maximize} \quad M$$

subject to $\sum_{j=1}^{p} \beta_j^2 = 1,$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C,$$



Optimization Problem

Let's unpack this! 

- The optimization goal of the SVM is:

$$\underset{\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots \epsilon_n, M}{\text{maximize}} \; M$$

subject to $\displaystyle\sum_{j=1}^{p} \beta_j^2 = 1,$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C,$$

Optimization Problem

- The optimization goal of the SVM is:

$$\underset{\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots \epsilon_n, M}{maximize} \quad M$$

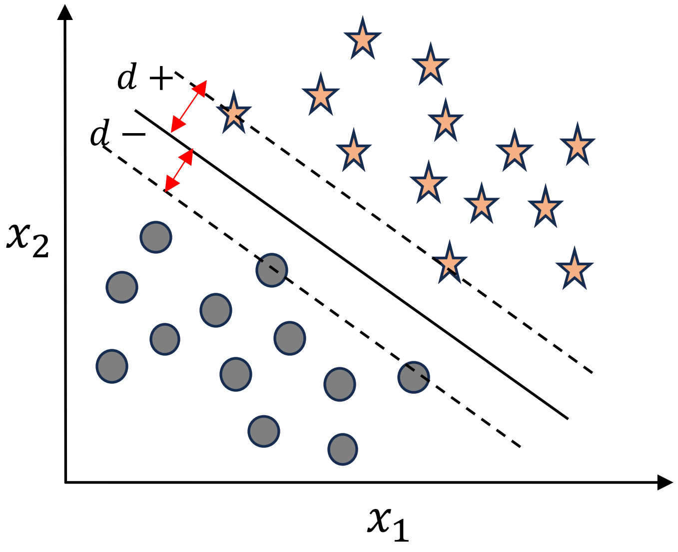
subject to $\sum_{j=1}^{p} \beta_j^2 = 1,$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C,$$

What we know

Distance from a point (x_0, y_0) to a line $ax + by + c = 0$ is $\frac{|ax+by+c|}{\sqrt{a^2+b^2}}$



Optimization Problem

What we know

Distance from a point (x_0, y_0) to a line $ax + by + c = 0$ is $\frac{|ax+by+c|}{\sqrt{a^2+b^2}}$

- The optimization goal of the SVM is:

$$\underset{\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots \epsilon_n, M}{maximize} \quad M$$

subject to $\sum_{j=1}^{p} \beta_j^2 = 1,$

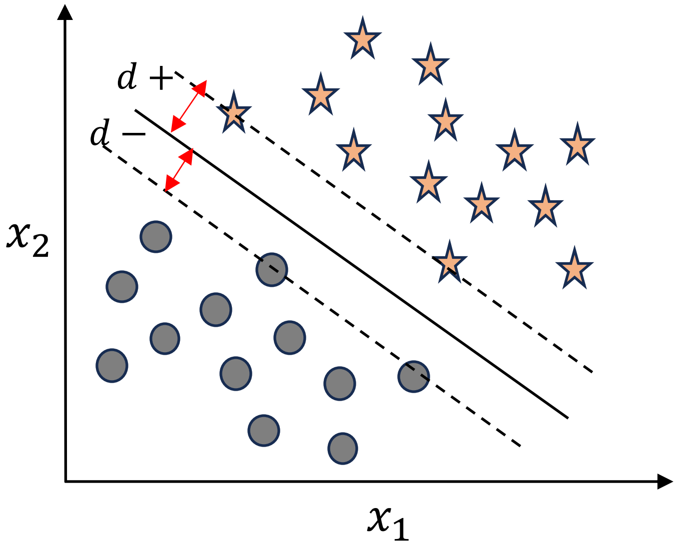
$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C,$$

What we know

$\beta_1, \beta_2 \ldots, \beta_p$ = weight vector **w**

β_0 = bias vector **b**



Optimization Problem

- The optimization goal of the SVM is:

$$\underset{\beta_0, \beta_1, \dots, \beta_p, \epsilon_1, \dots \epsilon_n, M}{maximize} \quad M$$

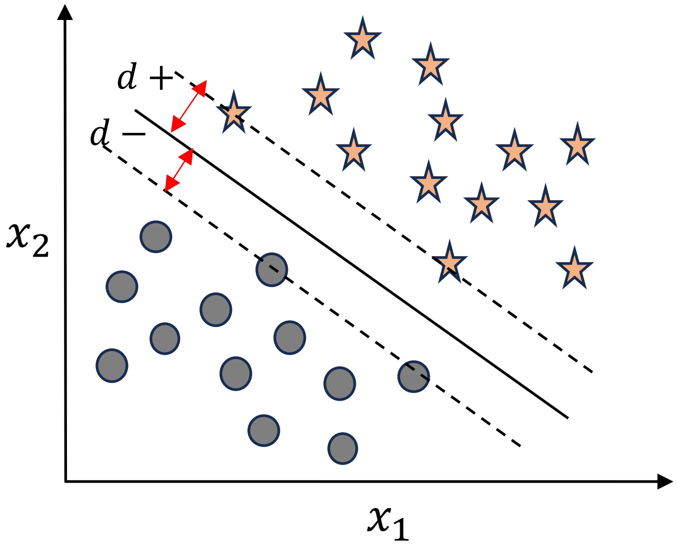
$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C,$$

With these knowledge points,

If weight vector is represented by **w**, then $d^+ = \frac{1}{||\boldsymbol{w}||}$ and $\mathrm{M} = \frac{2}{||\boldsymbol{w}||}$



Optimization Problem

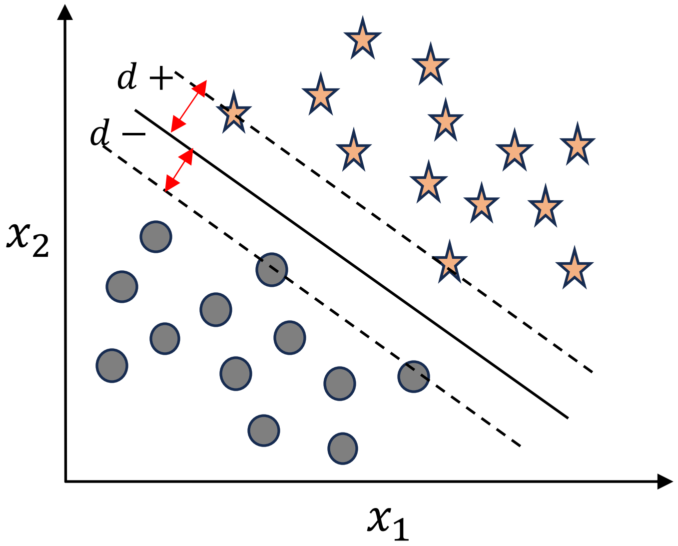
- The optimization goal of the SVM is:

$$\underset{b, w, \epsilon_1, \ldots \epsilon_n, M}{\text{maximize}} \frac{2}{||\boldsymbol{w}||}$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C,$$



Note 1: To maximize M, we need to minimize $||\boldsymbol{w}||$

Optimization Problem

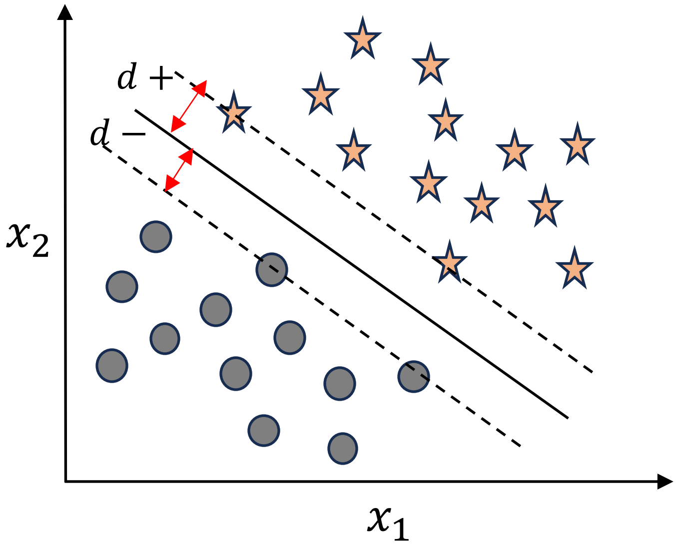
- The optimization goal of the SVM is:

$$\underset{b, w, \epsilon_1, \ldots \epsilon_n, M}{maximize} \ \frac{2}{||\boldsymbol{w}||}$$

$$\text{subject to} \ \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0 \, , \sum_{i=1}^{n} \epsilon_i \leq C,$$



Note 2: Minimizing $||\boldsymbol{w}||$ is same as minimizing $\frac{1}{2}||\boldsymbol{w}||^2$

Optimization Problem

- The optimization goal of the SVM is:

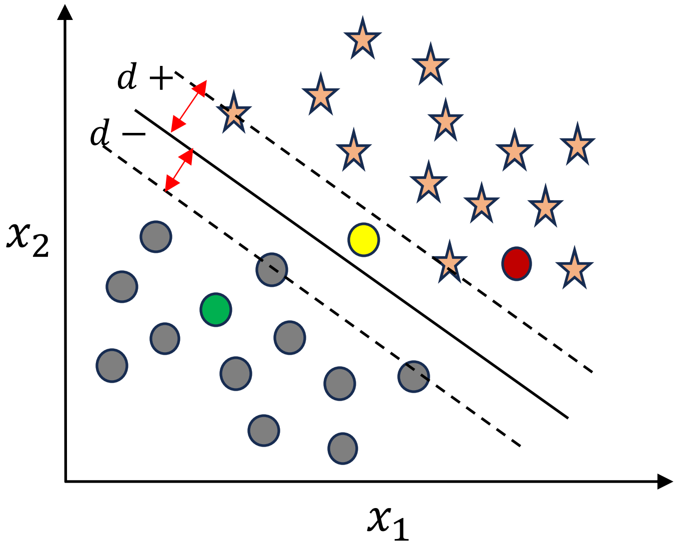
$$\underset{\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots \epsilon_n, M}{maximize} \quad M$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C,$$

What if observations are not linearly separable?



Optimization Problem

- The optimization goal of the SVM is:

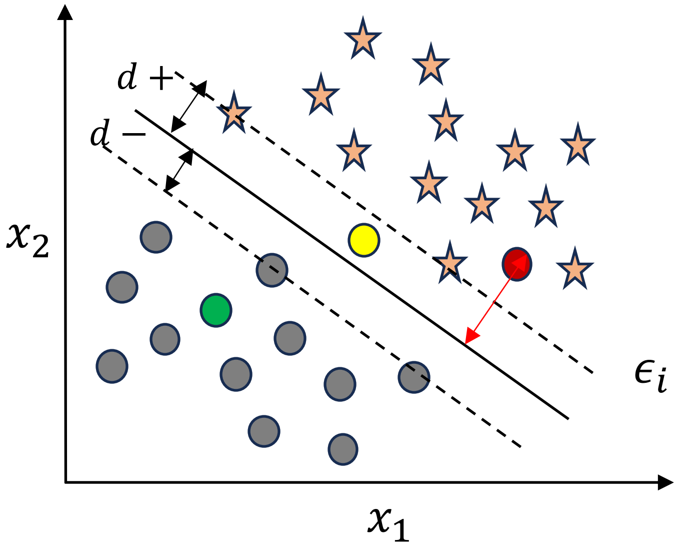
$$\underset{\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots \epsilon_n, M}{maximize} \quad M$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C,$$

- This observation is on correct side
- This observation is on wrong side of margin
- This observation is on wrong side of hyperplane



Optimization Problem

- The optimization goal of the SVM is:

$\epsilon_i = 0$

$\epsilon_i > 0$

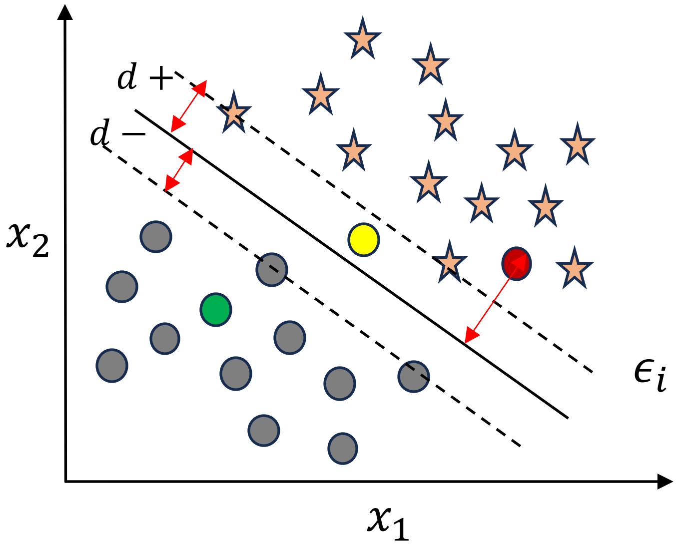
$\epsilon_i > 1$

$$\underset{\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots \epsilon_n, M}{\text{maximize}} \quad M$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C,$$



Problem corresponds to soft margin!

Optimization Problem

- The optimization goal of the SVM is:

$\epsilon_i = 0$

$\epsilon_i > 0$

$\epsilon_i > 1$

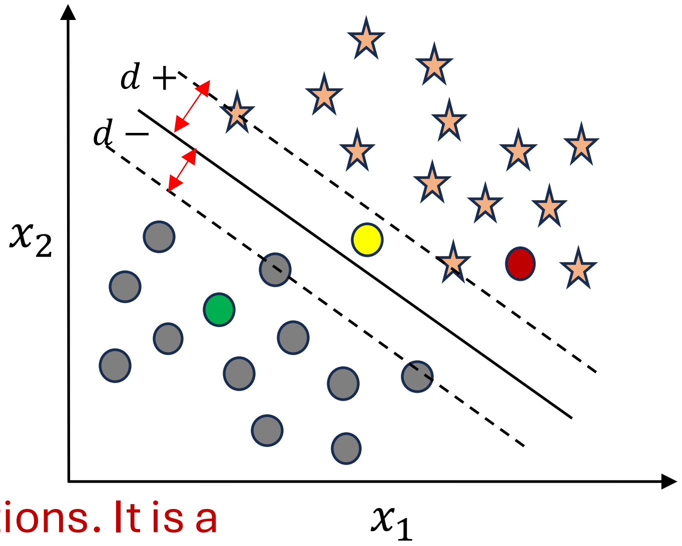
$$\underset{\beta_0, \beta_1, \dots, \beta_p, \epsilon_1, \dots \epsilon_n, M}{maximize} \quad M$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C,$$

C is the allowable budget for violations. It is a tuning parameter. It represents the amount that the margin can be violated by the n observations.



Optimization Problem

What if C=0?

- The optimization goal of the SVM is:

(green dot) $\epsilon_i = 0$

(yellow dot) $\epsilon_i > 0$

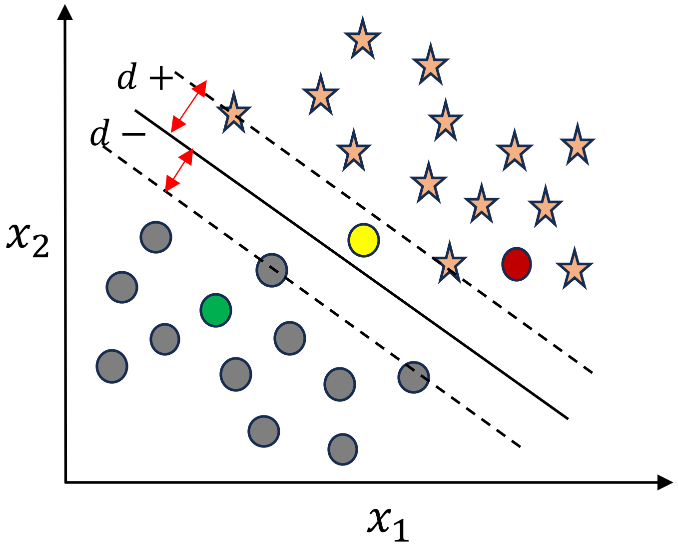
(red dot) $\epsilon_i > 1$

$$\underset{\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots \epsilon_n, M}{maximize} \quad M$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C,$$



Optimization Problem

How do we choose C?

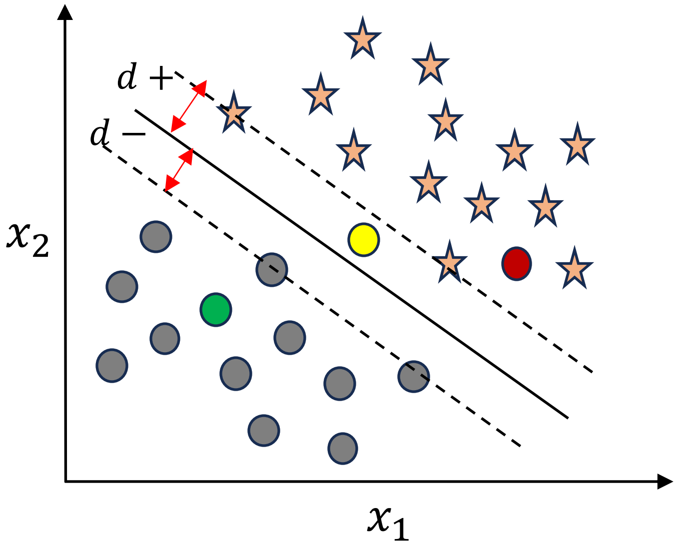
- The optimization goal of the SVM is:

$$\underset{\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots \epsilon_n, M}{maximize} \quad M$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C,$$



Optimization Problem

- The optimization goal of the SVM is:

Represents our hyperplane

$\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p\, x_{ip}$ is a hyperplane

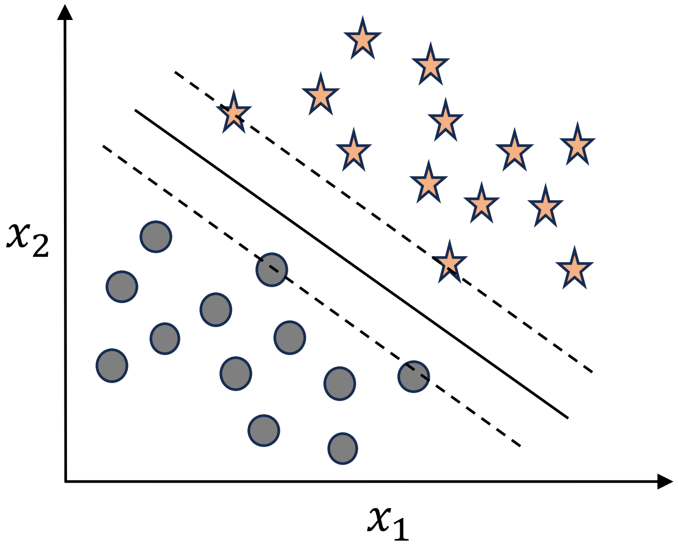
$k\,(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p\, x_{ip})$ is also a hyperplane if $k\neq 0$

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots\epsilon_n,M}{maximize}\quad M$$

subject to $\displaystyle\sum_{j=1}^{p}\beta_j^2=1,$

$$y_i(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p\, x_{ip})\geq M(1-\epsilon_i),$$

$$\epsilon_i\geq 0\,,\sum_{i=1}^{n}\epsilon_i\leq C,$$



Optimization Problem

- The optimization goal of the SVM is:

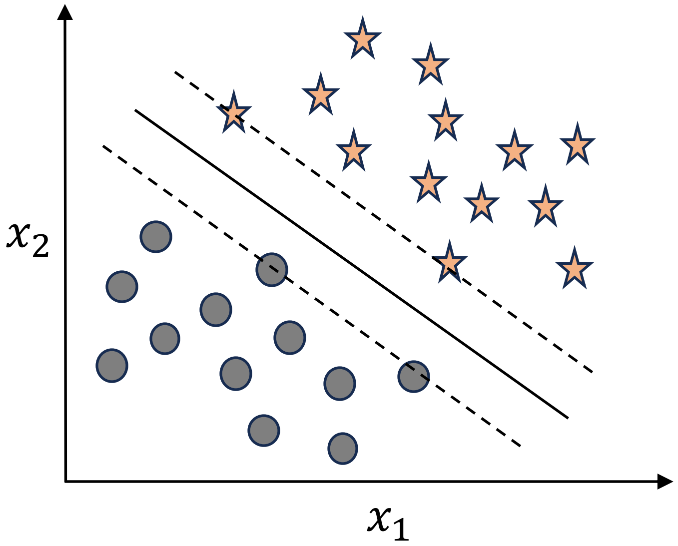
$y_i\,(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p\, x_{ip})$ alone does not make much sense, but...

$$\underset{\beta_0, \beta_1, \dots, \beta_p, \epsilon_1, \dots \epsilon_n, M}{maximize} \quad M$$

subject to $\sum_{j=1}^{p} \beta_j^2 = 1,$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p\, x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0\,, \sum_{i=1}^{n} \epsilon_i \leq C,$$



Optimization Problem

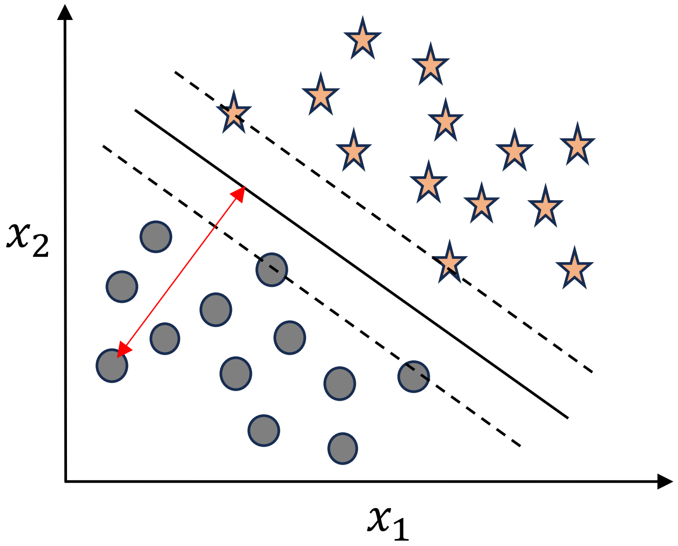
- The optimization goal of the SVM is:

$y_i\,(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p\, x_{ip})$ alone does not make much sense, but…

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots\epsilon_n,M}{\text{maximize}}\quad M$$

subject to $\displaystyle\sum_{j=1}^{p}\beta_j^2=1,$

With this constraint, it represents the perpendicular distance of i^{th} observation to the hyperplane



$$y_i(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p\, x_{ip})\geq M(1-\epsilon_i),$$

$$\epsilon_i\geq 0\,,\sum_{i=1}^{n}\epsilon_i\leq C,$$

Optimization Problem

When will this fail?

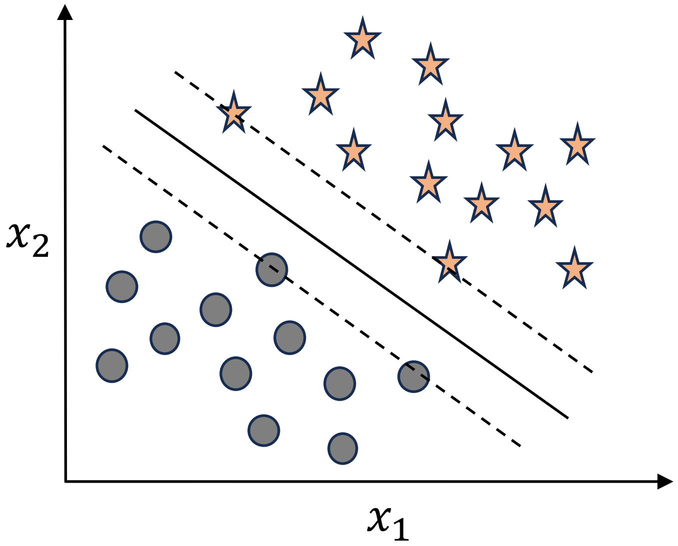
- SVM cannot be solved using a straightforward equation.
- The optimization goal of the SVM is:

$$\underset{\beta_0, \beta_1, \dots, \beta_p, \epsilon_1, \dots \epsilon_n, M}{maximize} \quad M$$

subject to $\sum_{j=1}^{p} \beta_j^2 = 1,$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C,$$



Optimization Problem

- The previous optimization problem can be re-written as:

$$\underset{\beta_0, \beta_1, \ldots, \beta_p}{minimize} \left\{ \sum_{i=1}^{n} \max\left[0, 1 - y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})\right] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

That is how it is introduced in ISLP

Optimization Problem

- The previous optimization problem can be re-written as:

$$\underset{\beta_0, \beta_1, \dots, \beta_p}{minimize} \left\{ \underbrace{\sum_{i=1}^{n} \max\left[0, 1 - y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})\right]}_{\text{Loss}} + \overbrace{\lambda \sum_{j=1}^{p} \beta_j^2}^{\text{Regularization / Penalty}} \right\}$$

Optimization Problem

- The previous optimization problem can be re-written as:

$$\underset{\beta_0, \beta_1, \ldots, \beta_p}{\text{minimize}} \left\{ \sum_{i=1}^{n} \max\left[0, 1 - y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})\right] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

- Above equation is same as:

$$\underset{b, w}{\text{minimize}} \left\{ C \sum_{i=1}^{n} \max[0, 1 - y_i(b + w.x)] + \frac{1}{2} ||w||^2 \right\}$$

That is how it is introduced in MML.

Optimization Problem

- The previous optimization problem can be re-written as:

$$\underset{\beta_0, \beta_1, \ldots, \beta_p}{minimize} \left\{ \sum_{i=1}^{n} \max[0, 1 - y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

- Above equation is same as:

λ plays the inverse role of C

$$\underset{b, w}{minimize} \left\{ C \sum_{i=1}^{n} \max[0, 1 - y_i(b + w.x)] + \frac{1}{2} ||w||^2 \right\}$$

Readings

***Required Readings*:**

Introduction to Statistical Learning
1. Chapter 9 – Section 9.1 – 9.3 Page 367 – 382

Supplemental Readings (Not required but recommended):

Mathematics for Machine Learning
1. Chapter 12 – Section 12.1 – 12.3 Page 370 – 383

Thank You

McMaster University