Motivation

Problem: Find a linear decision surface (Hyperplane) that can separate two classes, and has the largest distance between border line classes.

- borderline class (sample) = support vectors
- largest distance or gap = margin

[1] Lemma-1 in https://www.svms.org/tutorials/Burges1998.pdf

Motivation

Problem: Find a linear decision surface (Hyperplane) that can separate two classes, and has the largest distance between border line classes.

- borderline class (sample) = support vectors
- largest distance or gap = margin

- Assume that a linear classifier exists
- If the convex hull of two classes form non-intersecting convex sets, then there exists a separating hyperplane[1]
- What if the data is not linearly separable?

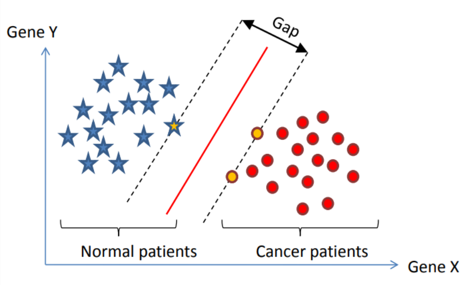
[1] Lemma-1 in https://www.svms.org/tutorials/Burges1998.pdf

Motivation

Problem: Find a linear decision surface (Hyperplane) that can separate two classes, and has the largest distance between border line classes.

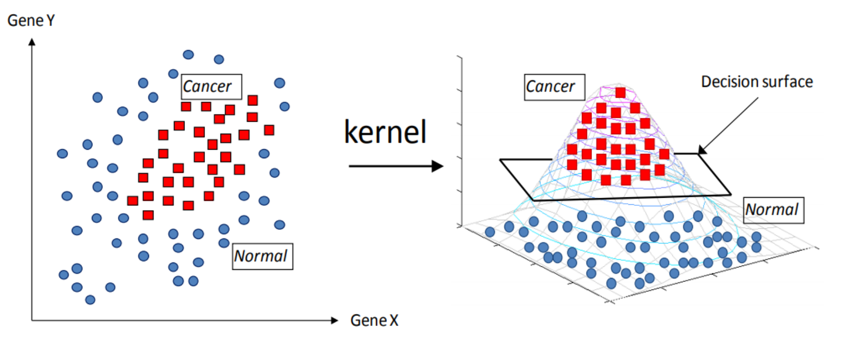
- borderline class (sample) = support vectors
- largest distance or gap = margin

- Assume that a linear classifier exists
- If the convex hull of two classes form non-intersecting convex sets, then there exists a separating hyperplane[1]
- What if the data is not linearly separable?

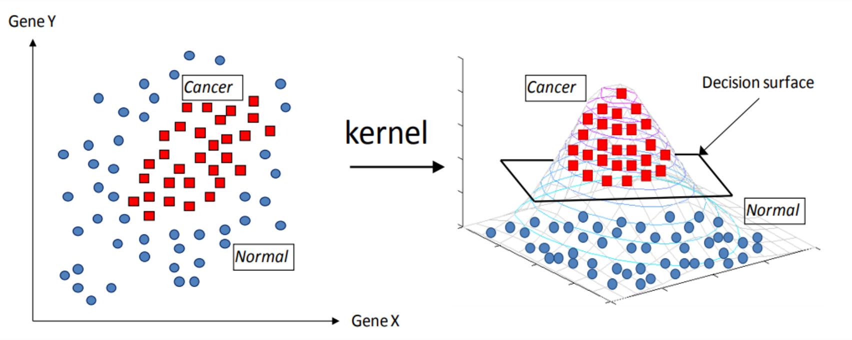


[1] Lemma-1 in https://www.svms.org/tutorials/Burges1998.pdf

When the Data is not linearly separable

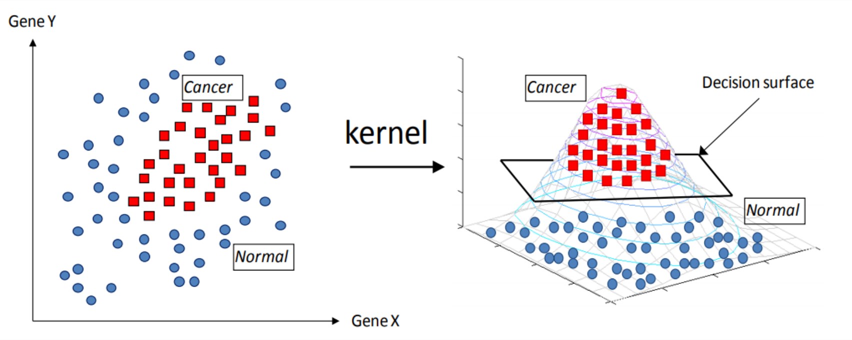


When the Data is not linearly separable



- Data is not linearly separable. That is, there does not exist a hyperplane that separates two classes

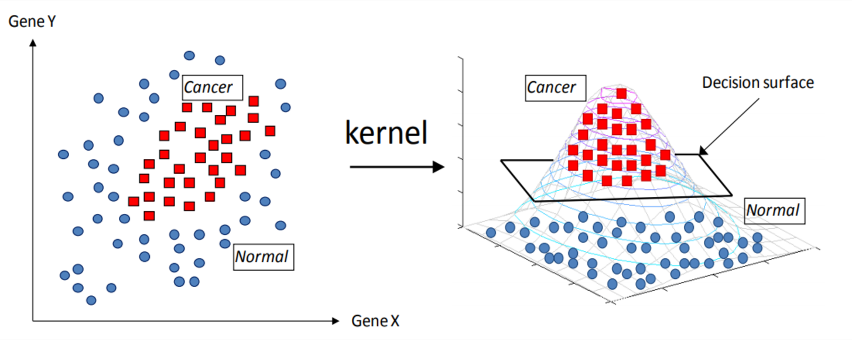
When the Data is not linearly separable



- Data is not linearly separable. That is, there does not exist a hyperplane that separates two classes

- After mapping the data is now linearly separable!

When the Data is not linearly separable



- Data is not linearly separable. That is, there does not exist a hyperplane that separates two classes

- After mapping the data is now linearly separable!

- Such a mapping is achieved using clever so-called Kernel trick!
- Note dimension of the space increases: price to pay for linear separation

History of SVM

- SVM have long history since 1960

History of SVM

- SVM have long history since 1960
- Use of SVMs have grown since...

History of SVM

- SVM have long history since 1960
- Use of SVMs have grown since...
- Main paper: ”A training algorithm for optimal margin classifiers”, by Boser, Guyon, Vapnik

History of SVM

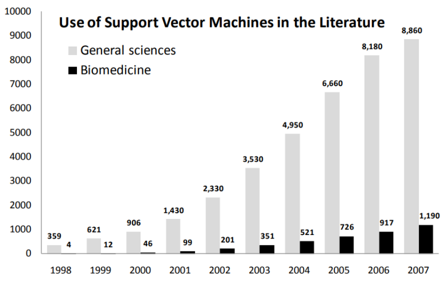
- SVM have long history since 1960
- Use of SVMs have grown since...
- Main paper: ”A training algorithm for optimal margin classifiers”, by Boser, Guyon, Vapnik
- It was discovered in 1960s, but its use started only around 1980s when Vapnik moved to USA

History of SVM

- SVM have long history since 1960
- Use of SVMs have grown since...
- Main paper: ”A training algorithm for optimal margin classifiers”, by Boser, Guyon, Vapnik
- It was discovered in 1960s, but its use started only around 1980s when Vapnik moved to USA
- Vapnik’s initial papers were rejected in NeurIPS!

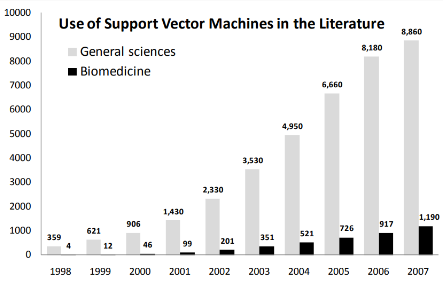
History of SVM

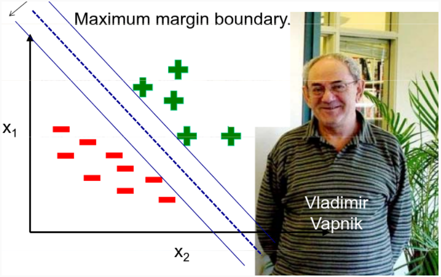
- SVM have long history since 1960
- Use of SVMs have grown since...
- Main paper: "A training algorithm for optimal margin classifiers", by Boser, Guyon, Vapnik
- It was discovered in 1960s, but its use started only around 1980s when Vapnik moved to USA
- Vapnik's initial papers were rejected in NeurIPS!



History of SVM

- SVM have long history since 1960
- Use of SVMs have grown since...
- Main paper: "A training algorithm for optimal margin classifiers", by Boser, Guyon, Vapnik
- It was discovered in 1960s, but its use started only around 1980s when Vapnik moved to USA
- Vapnik's initial papers were rejected in NeurIPS!





Representing the data

- SVMs are used for classifications and regressions

Representing the data

- SVMs are used for classifications and regressions
- How to represent data?

Representing the data

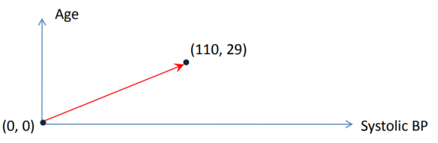
- SVMs are used for classifications and regressions
- How to represent data?
- A person aged 29 has BP of 110 is represented by a vector $(110, 29)$

Representing the data

- SVMs are used for classifications and regressions
- How to represent data?
- A person aged 29 has BP of 110 is represented by a vector $(110, 29)$
- vectors are tailed at origin $(0, 0)$

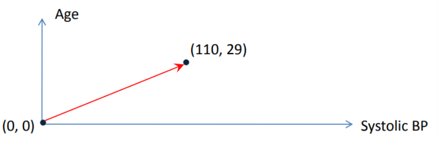
Representing the data

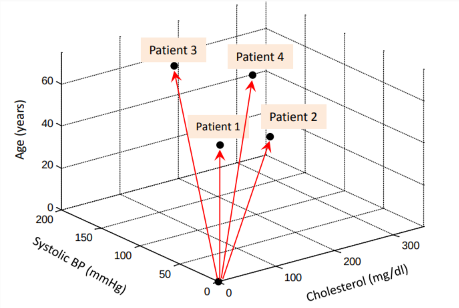
- SVMs are used for classifications and regressions
- How to represent data?
- A person aged 29 has BP of 110 is represented by a vector $(110, 29)$
- vectors are tailed at origin $(0, 0)$

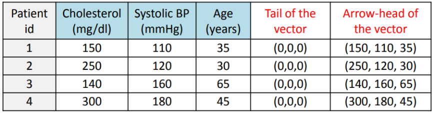


Representing the data

- SVMs are used for classifications and regressions
- How to represent data?
- A person aged 29 has BP of 110 is represented by a vector $(110, 29)$
- vectors are tailed at origin $(0, 0)$

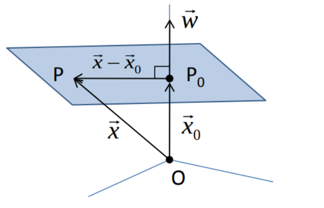


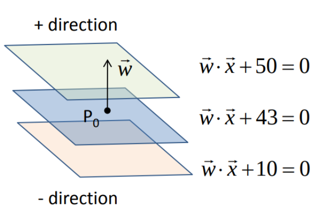




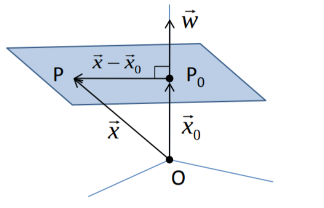
Patient id	Cholesterol (mg/dl)	Systolic BP (mmHg)	Age (years)	Tail of the vector	Arrow-head of the vector
1	150	110	35	(0,0,0)	(150, 110, 35)
2	250	120	30	(0,0,0)	(250, 120, 30)
3	140	160	65	(0,0,0)	(140, 160, 65)
4	300	180	45	(0,0,0)	(300, 180, 45)

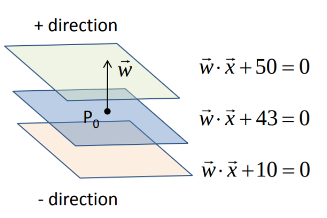
Recall: Equation of Hyperplane




+ direction
$\vec{w}$
$\vec{w} \cdot \vec{x} + 50 = 0$
P_0
$\vec{w} \cdot \vec{x} + 43 = 0$
$\vec{w} \cdot \vec{x} + 10 = 0$
- direction

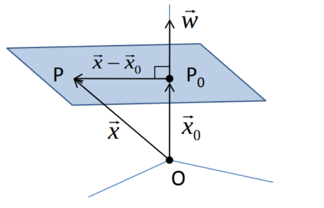
Recall: Equation of Hyperplane

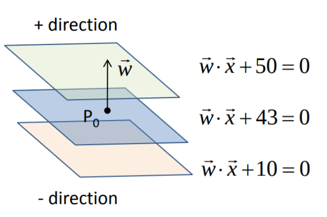



+ direction
$\vec{w}$
$\vec{w}\cdot\vec{x}+50=0$
P_0
$\vec{w}\cdot\vec{x}+43=0$
$\vec{w}\cdot\vec{x}+10=0$
- direction

- Let x_0 be any fixed point in plane P

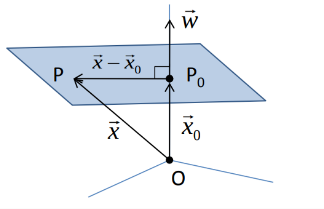
Recall: Equation of Hyperplane

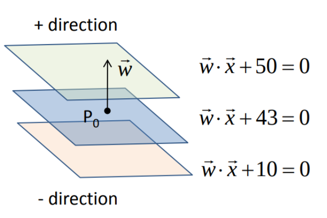



+ direction
$\vec{w}$
P_0
$\vec{w} \cdot \vec{x} + 50 = 0$
$\vec{w} \cdot \vec{x} + 43 = 0$
$\vec{w} \cdot \vec{x} + 10 = 0$
- direction

- Let x_0 be any fixed point in plane P
- Let w be a vector perpendicular to P

Recall: Equation of Hyperplane

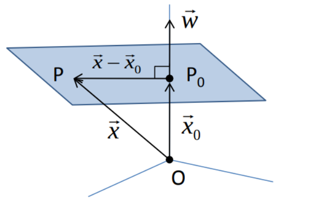


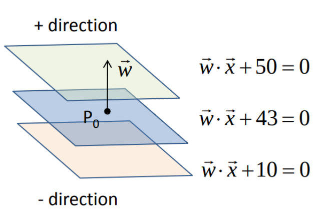

+ direction
$\vec{w}$
P_0
$\vec{w}\cdot\vec{x}+50=0$
$\vec{w}\cdot\vec{x}+43=0$
$\vec{w}\cdot\vec{x}+10=0$
- direction

- Let x_0 be any fixed point in plane P
- Let w be a vector perpendicular to P
- Let x be any point in P, then

$$(x - x_0) \perp w \implies (x - x_0)^T w = 0$$

Recall: Equation of Hyperplane





+ direction
$\vec{w}$
$\vec{w}\cdot\vec{x}+50=0$
P_0
$\vec{w}\cdot\vec{x}+43=0$
$\vec{w}\cdot\vec{x}+10=0$
- direction

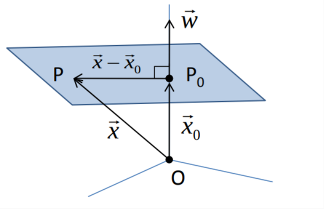
- Let x_0 be any fixed point in plane P
- Let w be a vector perpendicular to P
- Let x be any point in P, then

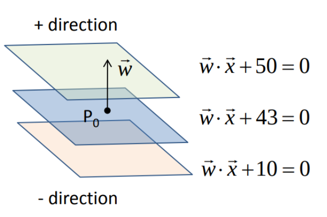
$$(x - x_0) \perp w \implies (x - x_0)^T w = 0$$

- Let $b = -w^T x$, then equation of plane is

$$w^T x + b = 0 \quad \text{or} \quad w \cdot x + b = 0$$

Recall: Equation of Hyperplane




+ direction
$\vec{w}$
$\vec{w}\cdot\vec{x}+50=0$
P_0
$\vec{w}\cdot\vec{x}+43=0$
$\vec{w}\cdot\vec{x}+10=0$
- direction

- Let x_0 be any fixed point in plane P
- Let w be a vector perpendicular to P
- Let x be any point in P, then

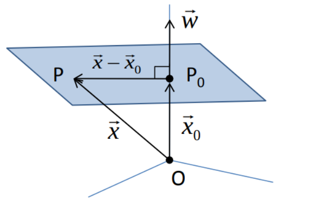
$$(x - x_0) \perp w \implies (x - x_0)^T w = 0$$

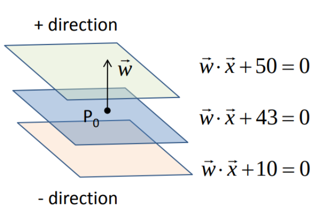
- Let $b = -w^T x$, then equation of plane is

$$w^T x + b = 0 \quad \text{or} \quad w \cdot x + b = 0$$

- Changing b, we get parallel planes, see figure on the right

Recall: Equation of Hyperplane





+ direction
$\vec{w}$
P_0
$\vec{w}\cdot\vec{x}+50=0$
$\vec{w}\cdot\vec{x}+43=0$
$\vec{w}\cdot\vec{x}+10=0$
- direction

- Let x_0 be any fixed point in plane P
- Let w be a vector perpendicular to P
- Let x be any point in P, then

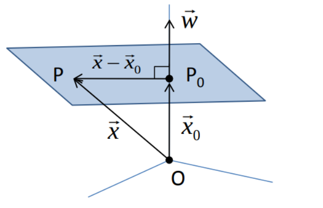
$$(x-x_0)\perp w \implies (x-x_0)^T w = 0$$

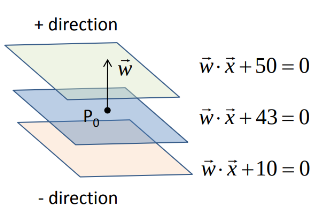
- Let $b=-w^T x$, then equation of plane is

$$w^T x + b = 0 \quad \text{or} \quad w\cdot x + b = 0$$

- Changing b, we get parallel planes, see figure on the right
  - increasing b moves in direction of w

Recall: Equation of Hyperplane




+ direction
$\vec{w}$
P_0
$\vec{w}\cdot\vec{x}+50=0$
$\vec{w}\cdot\vec{x}+43=0$
$\vec{w}\cdot\vec{x}+10=0$
- direction

- Let x_0 be any fixed point in plane P
- Let w be a vector perpendicular to P
- Let x be any point in P, then

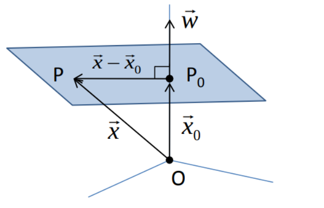
$$(x - x_0) \perp w \implies (x - x_0)^T w = 0$$

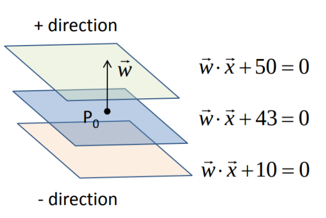
- Let $b = -w^T x$, then equation of plane is

$$w^T x + b = 0 \quad \text{or} \quad w \cdot x + b = 0$$

- Changing b, we get parallel planes, see figure on the right
  - increasing b moves in direction of w
  - decreasing b moves in direction of $-w$

Recall: Equation of Hyperplane





+ direction
$\vec{w}$
P_0
$\vec{w}\cdot\vec{x}+50=0$
$\vec{w}\cdot\vec{x}+43=0$
$\vec{w}\cdot\vec{x}+10=0$
- direction

- Let x_0 be any fixed point in plane P
- Let w be a vector perpendicular to P
- Let x be any point in P, then

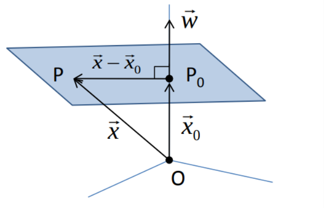
$$(x - x_0) \perp w \implies (x - x_0)^T w = 0$$

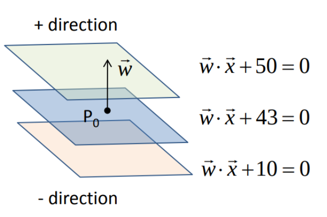
- Let $b = -w^T x$, then equation of plane is

$$w^T x + b = 0 \quad \text{or} \quad w \cdot x + b = 0$$

- Changing b, we get parallel planes, see figure on the right
  - increasing b moves in direction of w
  - decreasing b moves in direction of $-w$
- Distance between parallel planes $w^T x + b_1 = 0$ and $w^T x + b_2 = 0$ is

Recall: Equation of Hyperplane




+ direction
$\vec{w}$
$\vec{w}\cdot\vec{x}+50=0$
P_0
$\vec{w}\cdot\vec{x}+43=0$
$\vec{w}\cdot\vec{x}+10=0$
- direction

- Let x_0 be any fixed point in plane P
- Let w be a vector perpendicular to P
- Let x be any point in P, then

$$(x - x_0) \perp w \implies (x - x_0)^T w = 0$$

- Let $b = -w^T x$, then equation of plane is

$$w^T x + b = 0 \quad \text{or} \quad w \cdot x + b = 0$$

- Changing b, we get parallel planes, see figure on the right
  - increasing b moves in direction of w
  - decreasing b moves in direction of $-w$
- Distance between parallel planes $w^T x + b_1 = 0$ and $w^T x + b_2 = 0$ is

$$D = |b_1 - b_2| / \|w\|_2$$

SVMs for two class classification

- Consider the data:

$$X = \{x_1, x_2, \ldots, x_m \in \mathbb{R}^n\}$$
$$Y = \{y_1, y_2, \ldots, y_m \in \{+1, -1\}\}$$

SVMs for two class classification

- Consider the data:

$$X = \{x_1, x_2, \ldots, x_m \in \mathbb{R}^n\}$$

$$Y = \{y_1, y_2, \ldots, y_m \in \{+1, -1\}\}$$

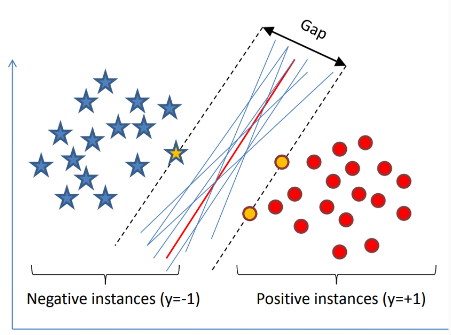
- Here Y is label. It takes $+1$ or -1 values. It is a two class problem

SVMs for two class classification

- Consider the data:

$$X = \{x_1, x_2, \ldots, x_m \in \mathbb{R}^n\}$$
$$Y = \{y_1, y_2, \ldots, y_m \in \{+1, -1\}\}$$

- Here Y is label. It takes $+1$ or -1 values. It is a two class problem



SVMs for two class classification

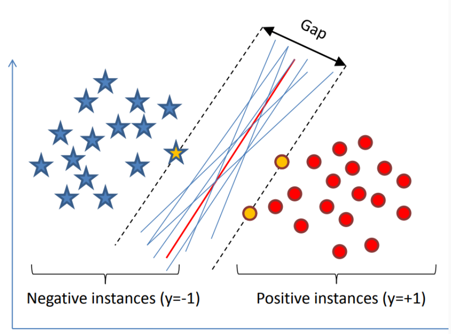
- Consider the data:

$$X = \{x_1, x_2, \dots, x_m \in \mathbb{R}^n\}$$

$$Y = \{y_1, y_2, \dots, y_m \in \{+1, -1\}\}$$

- Here Y is label. It takes $+1$ or -1 values. It is a two class problem

Goal: Find a classifier to separate negative instances from positive ones

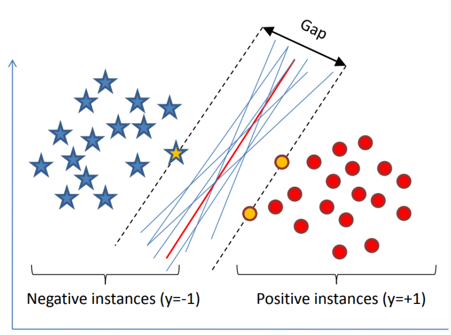


SVMs for two class classification

- Consider the data:

$$X = \{x_1, x_2, \ldots, x_m \in \mathbb{R}^n\}$$
$$Y = \{y_1, y_2, \ldots, y_m \in \{+1, -1\}\}$$

- Here Y is label. It takes $+1$ or -1 values. It is a two class problem



Goal: Find a classifier to separate negative instances from positive ones

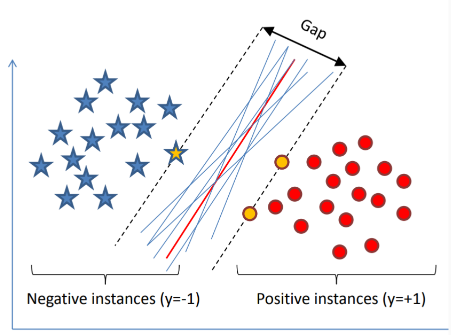
- As shown in figure, there exists infinite such hyperplanes (classifiers)

SVMs for two class classification

- Consider the data:

$$X = \{x_1, x_2, \ldots, x_m \in \mathbb{R}^n\}$$
$$Y = \{y_1, y_2, \ldots, y_m \in \{+1, -1\}\}$$

- Here Y is label. It takes $+1$ or -1 values. It is a two class problem



Goal: Find a classifier to separate negative instances from positive ones

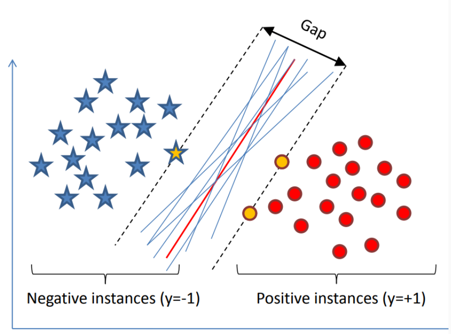
- As shown in figure, there exists infinite such hyperplanes (classifiers)
- Two hyperplanes (shown as dotted) pass through boundary points (support vectors)(yellow)

SVMs for two class classification

- Consider the data:

$$X = \{x_1, x_2, \ldots, x_m \in \mathbb{R}^n\}$$
$$Y = \{y_1, y_2, \ldots, y_m \in \{+1, -1\}\}$$

- Here Y is label. It takes $+1$ or -1 values. It is a two class problem



Goal: Find a classifier to separate negative instances from positive ones

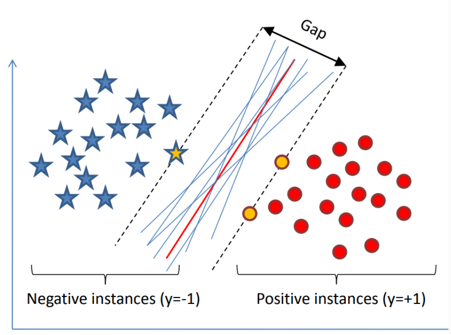
- As shown in figure, there exists infinite such hyperplanes (classifiers)
- Two hyperplanes (shown as dotted) pass through boundary points (support vectors)(yellow)
- SVMs finds hyperplanes that maximizes the gap bertween data points on boundary

SVMs for two class classification

- Consider the data:

$$X = \{x_1, x_2, \ldots, x_m \in \mathbb{R}^n\}$$
$$Y = \{y_1, y_2, \ldots, y_m \in \{+1, -1\}\}$$

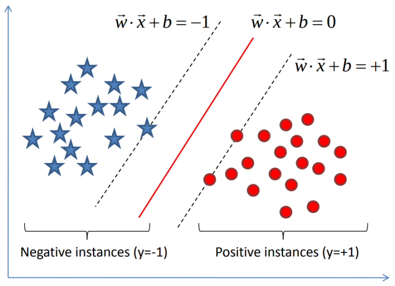
- Here Y is label. It takes $+1$ or -1 values. It is a two class problem



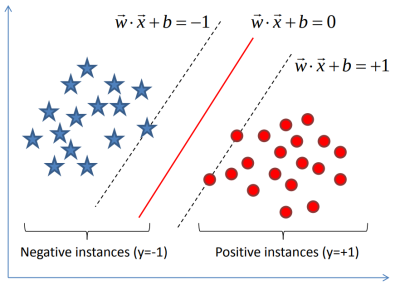
Goal: Find a classifier to separate negative instances from positive ones

- As shown in figure, there exists infinite such hyperplanes (classifiers)
- Two hyperplanes (shown as dotted) pass through boundary points (support vectors)(yellow)
- SVMs finds hyperplanes that maximizes the gap bertween data points on boundary
- If the support vectors are noisy, SVMs wont work well!

Particular Form of Hyperplanes Passing Through Support Vectors


$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

Particular Form of Hyperplanes Passing Through Support Vectors

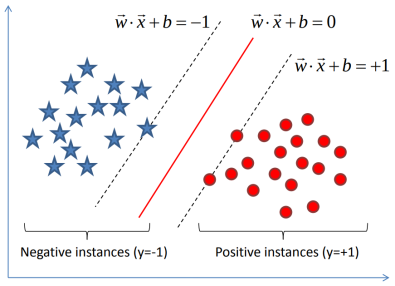

$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

- Quiz: Why dotted planes through support vectors are

$$w \cdot x + b = -1 \tag{1}$$

$$w \cdot x + b = 1 \tag{2}$$

Particular Form of Hyperplanes Passing Through Support Vectors


$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

- Quiz: Why dotted planes through support vectors are

$$w \cdot x + b = -1 \tag{1}$$

$$w \cdot x + b = 1 \tag{2}$$

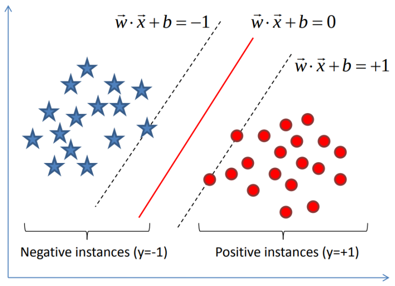
- Let middle red plane be

$$w \cdot x + b = 0 \tag{3}$$

- If we move c distance along w, we get to support vector of red class, then hyperplane is

$$w \cdot x + b = c \tag{4}$$

Particular Form of Hyperplanes Passing Through Support Vectors


$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

- Quiz: Why dotted planes through support vectors are

$$w \cdot x + b = -1 \tag{1}$$

$$w \cdot x + b = 1 \tag{2}$$

- Let middle red plane be

$$w \cdot x + b = 0 \tag{3}$$

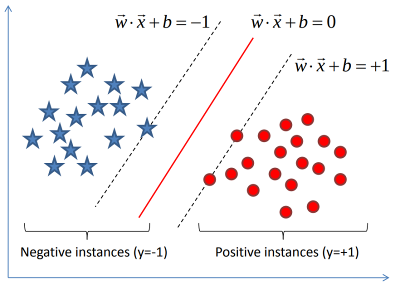
- If we move c distance along w, we get to support vector of red class, then hyperplane is

$$w \cdot x + b = c \tag{4}$$

- Similarly, if we move c distance along $-w$, we reach support vector of blue class

$$w \cdot x + b = -c \tag{5}$$

Particular Form of Hyperplanes Passing Through Support Vectors


$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

- Quiz: Why dotted planes through support vectors are

$$w \cdot x + b = -1 \tag{1}$$
$$w \cdot x + b = 1 \tag{2}$$

- Let middle red plane be

$$w \cdot x + b = 0 \tag{3}$$

- If we move c distance along w, we get to support vector of red class, then hyperplane is

$$w \cdot x + b = c \tag{4}$$

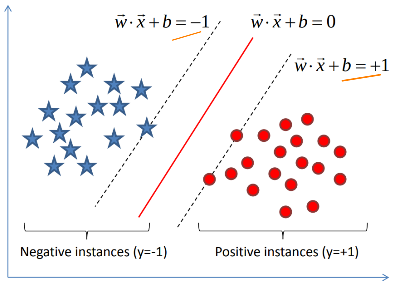
- Similarly, if we move c distance along $-w$, we reach support vector of blue class

$$w \cdot x + b = -c \tag{5}$$

- Divide $(3), (4), (5)$ by c to get

$$\tilde{w} \cdot x + \tilde{b} = 0$$
$$\tilde{w} \cdot x + \tilde{b} = -1$$
$$\tilde{w} \cdot x + \tilde{b} = 1$$

Particular Form of Hyperplanes Passing Through Support Vectors



$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

- Quiz: Why dotted planes through support vectors are

$$w \cdot x + b = -1 \tag{1}$$

$$w \cdot x + b = 1 \tag{2}$$

- Let middle red plane be

$$w \cdot x + b = 0 \tag{3}$$

- If we move c distance along w, we get to support vector of red class, then hyperplane is

$$w \cdot x + b = c \tag{4}$$

- Similarly, if we move c distance along $-w$, we reach support vector of blue class

$$w \cdot x + b = -c \tag{5}$$

- Divide $(3), (4), (5)$ by c to get

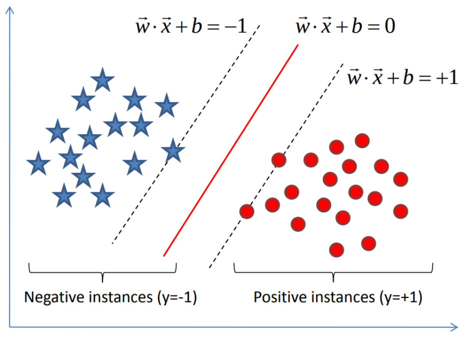
$$\tilde{w} \cdot x + \tilde{b} = 0$$

$$\tilde{w} \cdot x + \tilde{b} = -1$$

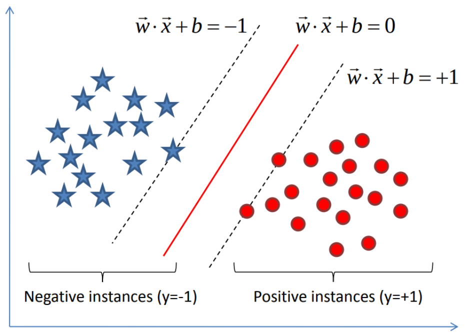
$$\tilde{w} \cdot x + \tilde{b} = 1$$

- Rename $\tilde{w}$ by w, and $\tilde{b}$ by b above!

SVM idea: Maximize the Margin

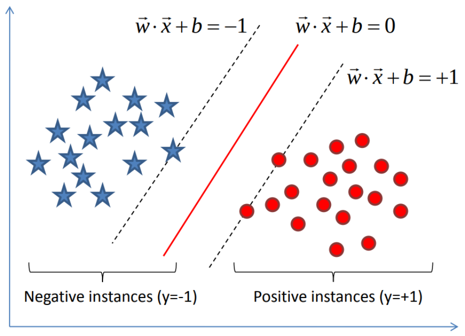

$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

SVM idea: Maximize the Margin


$\vec{w} \cdot \vec{x} + b = -1$
$\vec{w} \cdot \vec{x} + b = 0$
$\vec{w} \cdot \vec{x} + b = +1$
Negative instances (y=-1)
Positive instances (y=+1)

Margin/Gap: Distance between parallel planes passing through support vectors

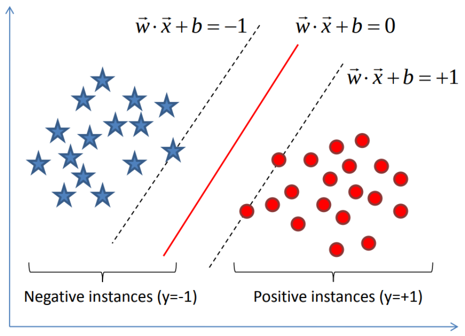
SVM idea: Maximize the Margin


$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

Margin/Gap: Distance between parallel planes passing through support vectors

Goal: Maximize the margin!

SVM idea: Maximize the Margin


$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

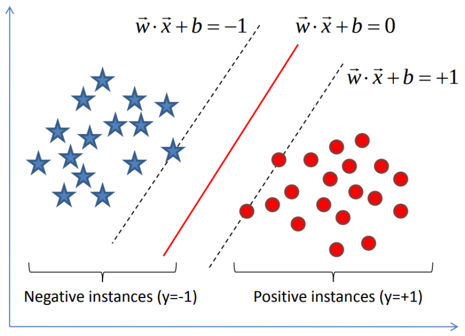
Margin/Gap: Distance between parallel planes passing through support vectors

Goal: Maximize the margin!

- We have

$$w\cdot x+b=-1$$
$$w\cdot x+b=1$$

SVM idea: Maximize the Margin



$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

Margin/Gap: Distance between parallel planes passing through support vectors

Goal: Maximize the margin!

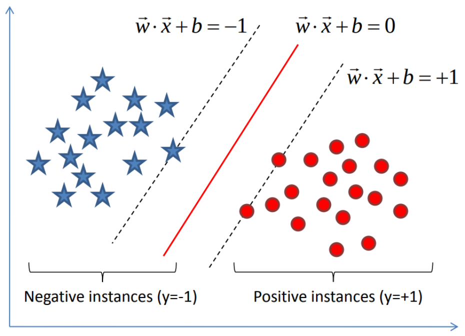
- We have

$$w \cdot x + b = -1$$
$$w \cdot x + b = 1$$

or equivalently,

$$w \cdot x + b + 1 = 0$$
$$w \cdot x + b - 1 = 0$$

SVM idea: Maximize the Margin


$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

Margin/Gap: Distance between parallel planes passing through support vectors

Goal: Maximize the margin!

- We have

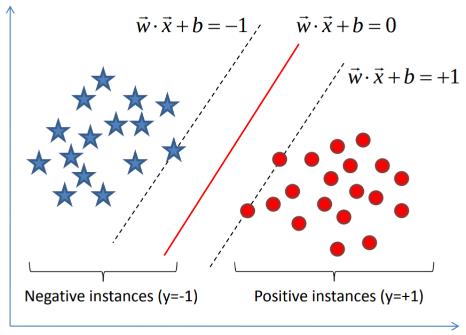
$$w \cdot x + b = -1$$
$$w \cdot x + b = 1$$

or equivalently,

$$w \cdot x + b + 1 = 0$$
$$w \cdot x + b - 1 = 0$$

- Distance/Margin $D = 2/\|w\|$

SVM idea: Maximize the Margin


$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

Margin/Gap: Distance between parallel planes passing through support vectors

Goal: Maximize the margin!

- We have

$$w \cdot x + b = -1$$
$$w \cdot x + b = 1$$

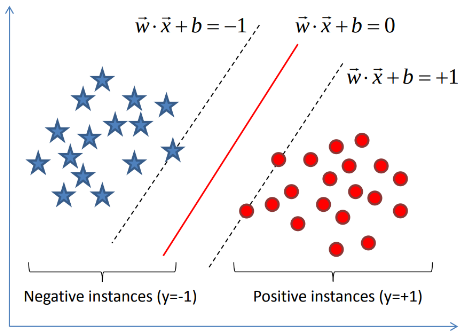
or equivalently,

$$w \cdot x + b + 1 = 0$$
$$w \cdot x + b - 1 = 0$$

- Distance/Margin $D = 2/\|w\|$
- To maximize the gap:

$$\boxed{\text{minimize} \quad \frac{1}{2}\|w\|^2}$$

SVM idea: Maximize the Margin



$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

Margin/Gap: Distance between parallel planes passing through support vectors

Goal: Maximize the margin!

- We have

$$w \cdot x + b = -1$$
$$w \cdot x + b = 1$$

or equivalently,

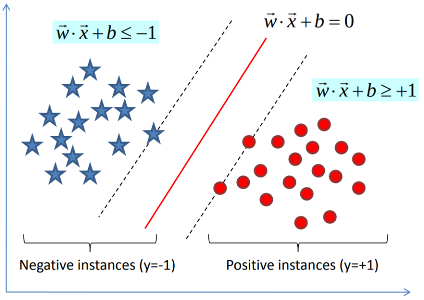
$$w \cdot x + b + 1 = 0$$
$$w \cdot x + b - 1 = 0$$

- Distance/Margin $D = 2/\|w\|$
- To maximize the gap:

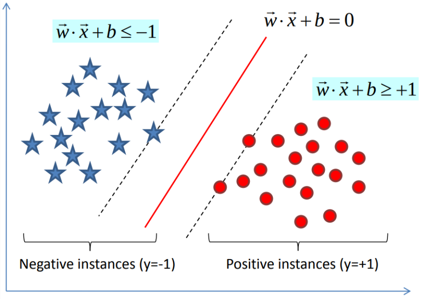
$$\boxed{\text{minimize} \quad \frac{1}{2}\|w\|^2}$$

- Are there constraints?

Impose Constraints, Optimization Model, Prediction

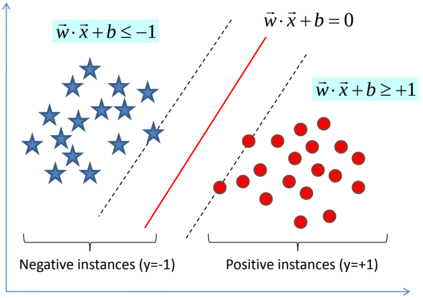

$\vec{w}\cdot\vec{x}+b\leq -1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b\geq +1$
Negative instances (y=-1)
Positive instances (y=+1)

Impose Constraints, Optimization Model, Prediction


$\vec{w}\cdot\vec{x}+b\leq -1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b\geq +1$
Negative instances (y=-1)
Positive instances (y=+1)

Goal: Impose constraints such that all the data poitns are correctly classified.

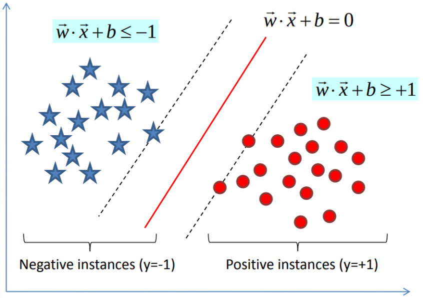
Impose Constraints, Optimization Model, Prediction


$\vec{w}\cdot\vec{x}+b\leq -1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b\geq +1$
Negative instances (y=-1)
Positive instances (y=+1)

Goal: Impose constraints such that all the data poitns are correctly classified.

- For all the blue data to be correctly classified, we must have

Impose Constraints, Optimization Model, Prediction

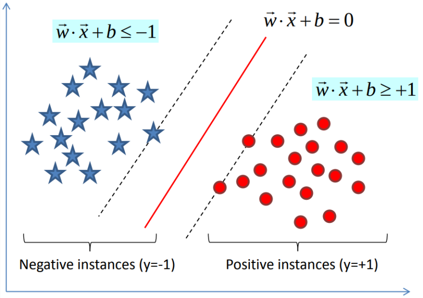

$\vec{w}\cdot\vec{x}+b\leq -1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b\geq +1$
Negative instances (y=-1)
Positive instances (y=+1)

Goal: Impose constraints such that all the data poitns are correctly classified.

- For all the blue data to be correctly classified, we must have

$$w\cdot x+b \leq -1, \text{ if } y_i=-1$$

Impose Constraints, Optimization Model, Prediction



$\vec{w}\cdot\vec{x}+b\leq -1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b\geq +1$
Negative instances (y=-1)
Positive instances (y=+1)

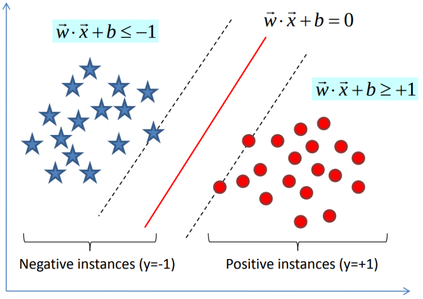
Goal: Impose constraints such that all the data poitns are correctly classified.

- For all the blue data to be correctly classified, we must have

$$w \cdot x + b \leq -1, \text{ if } y_i = -1$$

- Similarly, to classify all red data correctly, we must have

Impose Constraints, Optimization Model, Prediction


$\vec{w}\cdot\vec{x}+b\le -1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b\ge +1$
Negative instances (y=-1)
Positive instances (y=+1)

Goal: Impose constraints such that all the data poitns are correctly classified.

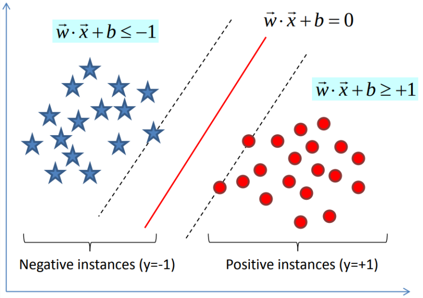
- For all the blue data to be correctly classified, we must have

$$w \cdot x + b \leq -1, \text{ if } y_i = -1$$

- Similarly, to classify all red data correctly, we must have

$$w \cdot x + b \geq +1, \text{ if } y_i = +1$$

Impose Constraints, Optimization Model, Prediction


$\vec{w}\cdot\vec{x}+b\le -1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b\ge +1$
Negative instances (y=-1)
Positive instances (y=+1)

Goal: Impose constraints such that all the data poitns are correctly classified.

- For all the blue data to be correctly classified, we must have

$$w \cdot x + b \leq -1, \text{ if } y_i = -1$$

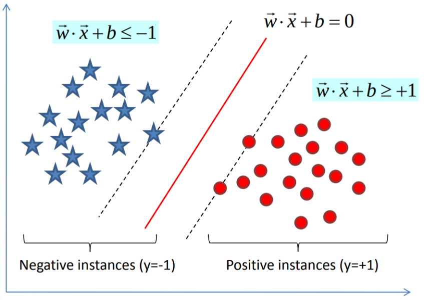
- Similarly, to classify all red data correctly, we must have

$$w \cdot x + b \geq +1, \text{ if } y_i = +1$$

- Equivalently, we have

$$y_i(w \cdot x + b) \geq 1$$

Impose Constraints, Optimization Model, Prediction



$\vec{w}\cdot\vec{x}+b\leq -1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b\geq +1$
Negative instances (y=-1)
Positive instances (y=+1)

Goal: Impose constraints such that all the data poitns are correctly classified.

Optimization model of SVM (training):

$$\text{minimize } \frac{1}{2}\|w\|^2$$

$$\text{subject to } y_i(w\cdot x+b)\geq 1,\ i=1,\ldots,m.$$

- For all the blue data to be correctly classified, we must have

$$w\cdot x+b\leq -1,\ \text{if } y_i=-1$$

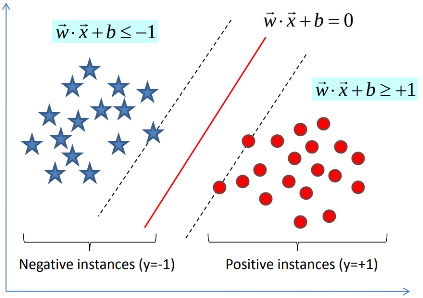
- Similarly, to classify all red data correctly, we must have

$$w\cdot x+b\geq +1,\ \text{if } y_i=+1$$

- Equivalently, we have

$$y_i(w\cdot x+b)\geq 1$$

Impose Constraints, Optimization Model, Prediction


$\vec{w}\cdot\vec{x}+b\le -1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b\ge +1$
Negative instances (y=-1)
Positive instances (y=+1)

Goal: Impose constraints such that all the data poitns are correctly classified.

Optimization model of SVM (training):

$$\underset{w}{\text{minimize}} \ \frac{1}{2}\|w\|^2$$

$$\text{subject to } y_i(w\cdot x+b)\ge 1,\ i=1,\dots,m.$$

- For all the blue data to be correctly classified, we must have

$$w\cdot x+b \le -1, \text{ if } y_i=-1$$

- Similarly, to classify all red data correctly, we must have

$$w\cdot x+b \ge +1, \text{ if } y_i=+1$$

- Equivalently, we have

$$y_i(w\cdot x+b)\ge 1$$

Prediction:

$$f(x)=\text{sign}(w\cdot x+b)$$

Dual Formulation of SVM Optimization Problem

Standard form:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & -y_i(w \cdot x + b) + 1 \leq 0, \quad i = 1, \ldots, m. \end{aligned}$$

Dual Formulation of SVM Optimization Problem

Standard form:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & -y_i(w \cdot x + b) + 1 \leq 0, \quad i = 1, \ldots, m. \end{aligned}$$

✓ This is primal problem in standard form

Dual Formulation of SVM Optimization Problem

Standard form:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & -y_i(w \cdot x + b) + 1 \leq 0, \quad i = 1, \ldots, m. \end{aligned}$$

- ✓ This is primal problem in standard form
- ✓ It is convex quadratic programming (QP) because objective is quardatic, and constraints are linear

Dual Formulation of SVM Optimization Problem

Standard form:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & -y_i(w \cdot x + b) + 1 \leq 0, \quad i = 1, \ldots, m. \end{aligned}$$

- ✓ This is primal problem in standard form
- ✓ It is convex quadratic programming (QP) because objective is quardatic, and constraints are linear
- ✓ The problem can be recast into dual form

Dual Formulation of SVM Optimization Problem

Standard form:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & -y_i(w \cdot x + b) + 1 \leq 0, \quad i = 1, \ldots, m. \end{aligned}$$

1. Define the Lagrangian

✓ This is primal problem in standard form

✓ It is convex quadratic programming (QP) because objective is quardatic, and constraints are linear

✓ The problem can be recast into dual form

Dual Formulation of SVM Optimization Problem

Standard form:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & -y_i(w \cdot x + b) + 1 \leq 0, \quad i = 1, \ldots, m. \end{aligned}$$

- ✓ This is primal problem in standard form
- ✓ It is convex quadratic programming (QP) because objective is quardatic, and constraints are linear
- ✓ The problem can be recast into dual form

1. Define the Lagrangian

$$L(w, b, \lambda) = \underline{\frac{1}{2}\|w\|_2^2} + \sum_{i=1}^{m} \lambda_i(-\underline{y_i(w \cdot x} + b) + 1)$$

Dual Formulation of SVM Optimization Problem

Standard form:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & -y_i(w \cdot x + b) + 1 \leq 0, \quad i = 1, \dots, m. \end{aligned}$$

- ✓ This is primal problem in standard form
- ✓ It is convex quadratic programming (QP) because objective is quardatic, and constraints are linear
- ✓ The problem can be recast into dual form

1. Define the Lagrangian

$$L(w, b, \lambda) = \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{m} \lambda_i(-y_i(w \cdot x + b) + 1)$$

2. The dual function is

Dual Formulation of SVM Optimization Problem

Standard form:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & -y_i(w \cdot x + b) + 1 \leq 0, \quad i = 1, \ldots, m. \end{aligned}$$

- ✓ This is primal problem in standard form
- ✓ It is convex quadratic programming (QP) because objective is quardatic, and constraints are linear
- ✓ The problem can be recast into dual form

1. Define the Lagrangian

$$L(w, b, \lambda) = \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{m} \lambda_i(-y_i(w \cdot x + b) + 1)$$

2. The dual function is

$$g(\lambda) = \inf_{w,b} L(w, b, \lambda)$$

Dual Formulation of SVM Optimization Problem

Standard form:

$$\begin{aligned} \underset{x,b}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & -y_i(w\cdot x+b)+1\le 0, \quad i=1,\dots,m. \end{aligned}$$

- ✓ This is primal problem in standard form
- ✓ It is convex quadratic programming (QP) because objective is quardatic, and constraints are linear
- ✓ The problem can be recast into dual form

1. Define the Lagrangian

$$L(w,b,\lambda)=\frac{1}{2}\|w\|_2^2+\sum_{i=1}^{m}\lambda_i(-y_i(w\cdot x+b)+1)$$

2. The dual function is

$$g(\lambda)=\inf_{\underset{x}{\boxed{w,b}}} L(\underbrace{w,b}_{x},\lambda)$$

3. $\underbrace{L(w,b,\lambda)}$ convex in w, b, minima given by

$$\nabla_{w,b}L(w,b,\lambda)=0$$

SVM-Dual Formulation

1. Setting the derivative of L w.r.t b to zero,

SVM-Dual Formulation

1. Setting the derivative of L w.r.t b to zero,

$$\frac{\partial L(w, b, \lambda)}{\partial b} = 0 \implies \sum_{i=1}^{m} \lambda_i y_i = 0$$

SVM-Dual Formulation

1. Setting the derivative of L w.r.t b to zero,

$$\frac{\partial L(w, b, \lambda)}{\partial b} = 0 \implies \sum_{i=1}^{m} \lambda_i y_i = 0$$

2. Setting the derivative of L w.r.t. w to zero,

SVM-Dual Formulation

1. Setting the derivative of L w.r.t b to zero,

$$\frac{\partial L(w,b,\lambda)}{\partial b} = 0 \implies \sum_{i=1}^{m} \lambda_i y_i = 0$$

2. Setting the derivative of L w.r.t. w to zero,

$$\frac{\partial L(w,b,\lambda)}{\partial w} = 0 \implies w = \sum_{j=1}^{m} \lambda_j y_j x_j$$

SVM-Dual Formulation

1. Setting the derivative of L w.r.t b to zero,

$$\frac{\partial L(w, b, \lambda)}{\partial b} = 0 \implies \sum_{i=1}^{m} \lambda_i y_i = 0$$

2. Setting the derivative of L w.r.t. w to zero,

$$\frac{\partial L(w, b, \lambda)}{\partial w} = 0 \implies w = \sum_{j=1}^{m} \lambda_j y_j x_j$$

3. Substituting in $\underline{L(w, b,} \lambda) = \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{m} \lambda_i(-y_i(w \cdot x_i + b) + 1)$,

SVM-Dual Formulation

1. Setting the derivative of L w.r.t b to zero,

$$\frac{\partial L(w,b,\lambda)}{\partial b} = 0 \implies \sum_{i=1}^{m} \lambda_i y_i = 0$$

2. Setting the derivative of L w.r.t. w to zero,

$$\frac{\partial L(w,b,\lambda)}{\partial w} = 0 \implies w = \sum_{j=1}^{m} \lambda_j y_j x_j$$

3. Substituting in $L(w,b,\lambda) = \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{m} \lambda_i(-y_i(w \cdot x_i + b) + 1)$,

$$g(\lambda) = \sum_{i=1}^{m} \lambda_i - \frac{1}{2}\sum_{i=1}^{m} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j)$$

$$-\sum \lambda_i \lambda_j y_i y_j (x_i \cdot x_j)$$

$$+\sum \lambda_i + \cancel{\sum_{i=1}^{m} \lambda_i b y_i} \to 0$$

=

because

$$\boxed{\sum \lambda_i y_i = 0}$$

$$\frac{1}{2}\|w\|_2^2 = \frac{1}{2} w^T w = \frac{1}{2}\Big(\sum_{i=1}^{m} \lambda_i y_i x_i\Big)^T \Big(\sum_{j=1}^{m} \lambda_j y_j x_j\Big)$$

$$\frac{1}{2}\sum_{i=1}^{m} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j)$$

SVM-Dual Formulation

1. Setting the derivative of L w.r.t b to zero,

$$\frac{\partial L(w,b,\lambda)}{\partial b} = 0 \implies \sum_{i=1}^{m} \lambda_i y_i = 0$$

2. Setting the derivative of L w.r.t. w to zero,

$$\frac{\partial L(w,b,\lambda)}{\partial w} = 0 \implies w = \sum_{j=1}^{m} \lambda_j y_j x_j$$

3. Substituting in $L(w,b,\lambda) = \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{m} \lambda_i(-y_i(w \cdot x_i + b) + 1)$,

$$g(\lambda) = \sum_{i=1}^{m} \lambda_i - \frac{1}{2}\sum_{i=1}^{m} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j)$$

SVM Dual Problem:

maximize $\displaystyle \sum_{i=1}^{m} \lambda_i - \frac{1}{2}\sum_{i=1}^{m} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j)$

subject to $\underline{\lambda_i \geq 0}, \quad \displaystyle \sum_{i=1}^{m} \lambda_i y_i = 0, \quad i = 1, \ldots, m$

SVM Primal Versus Dual Problem

SVM Primal Problem:

$$\text{minimize } \frac{1}{2}\|w\|^2, \quad w \in \mathbb{R}^n$$

$$\text{subject to } -y_i(w \cdot x + b) + 1 \leq 0, \; i = 1, \ldots, m.$$

SVM Primal Versus Dual Problem

SVM Primal Problem:

$$\text{minimize } \frac{1}{2}\|w\|^2, \quad w \in \mathbb{R}^n$$

$$\text{subject to } -y_i(w \cdot x + b) + 1 \leq 0, \ i = 1, \ldots, m.$$

SVM Dual Problem:

$$\text{maximize } \quad \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j)$$

$$\text{subject to } \quad \lambda_i \geq 0, \quad \sum_{i=1}^{m} \lambda_i y_i = 0, \ i = 1, \ldots, m.$$

SVM Primal Versus Dual Problem

SVM Primal Problem:

$$\text{minimize } \frac{1}{2}\|w\|^2, \quad w \in \mathbb{R}^n$$

$$\text{subject to } -y_i(w \cdot x + b) + 1 \leq 0, \; i = 1, \ldots, m.$$

SVM Dual Problem:

$$\text{maximize } \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j)$$

$$\text{subject to } \lambda_i \geq 0, \quad \sum_{i=1}^{m} \lambda_i y_i = 0, \; i = 1, \ldots, m.$$

Primal:

- Variables are $\{\underline{w_1}, \underline{w_2}, \ldots, \underline{w_n}\}$

SVM Primal Versus Dual Problem

SVM Primal Problem:

$$\text{minimize } \frac{1}{2}\|w\|^2, \quad w \in \mathbb{R}^n$$

$$\text{subject to } -y_i(w \cdot x + b) + 1 \leq 0, \; i = 1, \ldots, m.$$

SVM Dual Problem:

$$\text{maximize } \quad \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j)$$

$$\text{subject to } \quad \lambda_i \geq 0, \quad \sum_{i=1}^{m} \lambda_i y_i = 0, \; i = 1, \ldots, m.$$

Primal:

- Variables are $\{w_1, w_2, \ldots, w_n\}$
- n is number of features

SVM Primal Versus Dual Problem

SVM Primal Problem:

$$\text{minimize } \frac{1}{2}\|w\|^2, \quad w \in \mathbb{R}^n$$

$$\text{subject to } -y_i(w \cdot x + b) + 1 \leq 0, \; i = 1, \ldots, m.$$

$\frac{1}{2} w^T w$

$\boxed{P \succeq 0}$

constraint + objective fn all convex

SVM Dual Problem:

$$\text{maximize} \quad \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j)$$

$$\text{subject to} \quad \lambda_i \geq 0, \quad \sum_{i=1}^{m} \lambda_i y_i = 0, \; i = 1, \ldots, m.$$

Primal:

- Variables are $\{w_1, w_2, \ldots, w_n\}$
- n is number of features
- Primal is convex

SVM Primal Versus Dual Problem

SVM Primal Problem:

$$\text{minimize } \frac{1}{2}\|w\|^2, \quad w \in \mathbb{R}^n$$

$$\text{subject to } -y_i(w \cdot x + b) + 1 \leq 0, \; i = 1, \dots, m.$$

SVM Dual Problem:

$$\text{maximize} \quad \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j)$$

$$\text{subject to} \quad \lambda_i \geq 0, \quad \sum_{i=1}^{m} \lambda_i y_i = 0, \; i = 1, \dots, m.$$

Primal:

- Variables are $\{w_1, w_2, \dots, w_n\}$
- n is number of features
- Primal is convex

Dual:

- Variables are $\underline{\{\lambda_1, \lambda_2, \dots, \lambda_m\}}$

SVM Primal Versus Dual Problem

SVM Primal Problem:

$$\text{minimize } \frac{1}{2}\|w\|^2, \quad w \in \mathbb{R}^n$$

$$\text{subject to } -y_i(w \cdot x + b) + 1 \leq 0,\ i = 1, \ldots, m.$$

SVM Dual Problem:

$$\text{maximize} \quad \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j)$$

$$\text{subject to} \quad \lambda_i \geq 0, \quad \sum_{i=1}^{m} \lambda_i y_i = 0,\ i = 1, \ldots, m.$$

Primal:

- Variables are $\{w_1, w_2, \ldots, w_n\}$
- n is number of features
- Primal is convex

Dual:

- Variables are $\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$
- m is number of data samples

SVM Primal Versus Dual Problem

SVM Primal Problem:

$$\text{minimize } \frac{1}{2}\|w\|^2, \quad w \in \mathbb{R}^n$$

$$\text{subject to } -y_i(w \cdot x + b) + 1 \leq 0, \; i = 1, \ldots, m.$$

SVM Dual Problem:

$$\text{maximize} \quad \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j)$$

$$\text{subject to} \quad \lambda_i \geq 0, \quad \sum_{i=1}^{m} \lambda_i y_i = 0, \; i = 1, \ldots, m.$$

$-\lambda_i \leq 0$

minimiz

$-\left(\quad (obj) \quad \right)$

Primal:

- Variables are $\{w_1, w_2, \ldots, w_n\}$
- n is number of features
- Primal is convex

Dual:

- Variables are $\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$
- m is number of data samples
- Dual is convex

SVM Primal Versus Dual Problem

SVM Primal Problem:

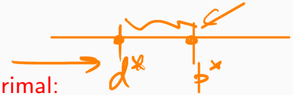
$$\text{minimize } \frac{1}{2}\|w\|^2, \quad w \in \mathbb{R}^n$$

$$\text{subject to } -y_i(w \cdot x + b) + 1 \leq 0, \; i = 1, \ldots, m.$$

SVM Dual Problem:

$$\text{maximize } \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j)$$

$$\text{subject to } \lambda_i \geq 0, \quad \sum_{i=1}^{m} \lambda_i y_i = 0, \; i = 1, \ldots, m.$$



Primal:

- Variables are $\{w_1, w_2, \ldots, w_n\}$
- n is number of features
- Primal is convex

Dual:

- Variables are $\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$
- m is number of data samples
- Dual is convex

Recommendation: Use dual when the number of samples m are significantly less relative to the number of features n, otherwise use primal.

Strong Duality and KKT Conditions for SVM

SVM Primal Problem:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & -y_i(w \cdot x + b) + 1 \leq 0, \quad i = 1, \ldots, m. \end{aligned}$$

Strong Duality and KKT Conditions for SVM

SVM Primal Problem:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & -y_i(w \cdot x + b) + 1 \leq 0, \quad i = 1, \ldots, m. \end{aligned}$$

Recall Slater's condition from slide ~~12~~

Slater's condition when some f_i are affine. There exists $x \in \operatorname{int} \mathcal{D}$ with

$$\underbrace{f_i(x) \leq 0, \quad i = 1, \ldots, k,}_{\text{affine}} \quad \underbrace{f_i(x) < 0, \quad i = k+1, \ldots, m,}_{\text{non-affine}} \quad Ax = b.$$

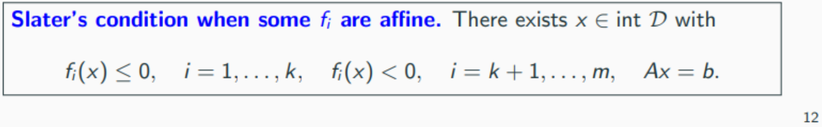
12

Strong Duality and KKT Conditions for SVM

SVM Primal Problem:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & -y_i(w \cdot x + b) + 1 \leq 0, \quad i = 1, \ldots, m. \end{aligned}$$

Recall Slater's condition from slide 12



Slater's condition when some f_i are affine. There exists $x \in \operatorname{int} \mathcal{D}$ with

$$f_i(x) \leq 0, \quad i = 1, \ldots, k, \quad f_i(x) < 0, \quad i = k+1, \ldots, m, \quad Ax = b.$$

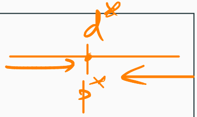
12

For SVM primal problem

- objective is quadratic and convex, and all inequality constraints are affine
- Slater's condition holds trivially, hence, strong duality holds

SVM Dual Problem:

$$\left.\begin{aligned}
\text{maximize} \quad & \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j) \\
\text{subject to} \quad & \lambda_i \geq 0, \quad i = 1, \ldots, m, \\
& \sum_{i=1}^{m} \lambda_i y_i = 0.
\end{aligned}\right\}$$



SVM Dual Problem:

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j) \\ \text{subject to} \quad & \lambda_i \geq 0, \quad i = 1, \ldots, m, \\ & \sum_{i=1}^{m} \lambda_i y_i = 0. \end{aligned}$$

Solving primal from dual:

- Assume that the dual problem is solved to obtain the dual optimal $\underline{\lambda^*}$
- Recall: Setting the derivative of L w.r.t. w to zero, we got

SVM Dual Problem:

$$
\begin{aligned}
\underset{\lambda}{\text{maximize}} \quad & \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j) \\
\text{subject to} \quad & \lambda_i \geq 0, \quad i = 1, \ldots, m, \\
& \sum_{i=1}^{m} \lambda_i y_i = 0.
\end{aligned}
$$

Solving primal from dual:

- Assume that the dual problem is solved to obtain the dual optimal λ^*
- Recall: Setting the derivative of L w.r.t. w to zero, we got

$$\frac{\partial L(w, b, \lambda)}{\partial w} = 0 \implies w = \sum_{i=1}^{m} \lambda_i y_i x_i$$

SVM Dual Problem:

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j) \\ \text{subject to} \quad & \lambda_i \geq 0, \quad i = 1, \ldots, m, \\ & \sum_{i=1}^{m} \lambda_i y_i = 0. \end{aligned}$$

Solving primal from dual:

- Assume that the dual problem is solved to obtain the dual optimal λ^*
- Recall: Setting the derivative of L w.r.t. w to zero, we got

$$\frac{\partial L(w, b, \lambda)}{\partial w} = 0 \implies w = \sum_{i=1}^{m} \lambda_i y_i x_i$$

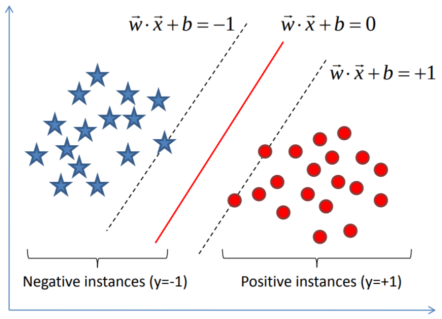
- Hence we get primal optimal w^* as

$$w^* = \sum_{i=1}^{m} \lambda_i^* y_i x_i$$

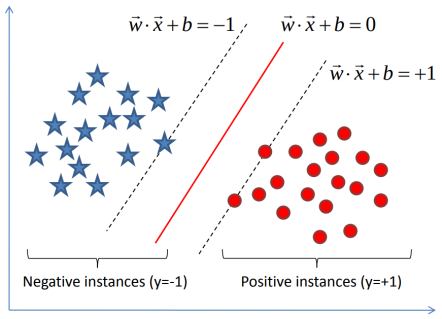
- The primal intercept b^* is

$$b^* = -\frac{\max_{i:y^{(i)}=-1} {w^*}^T x^{(i)} + \min_{i:y^{(i)}=+1} {w^*}^T x^{(i)}}{2}$$

Computing the optimal intercept b^*

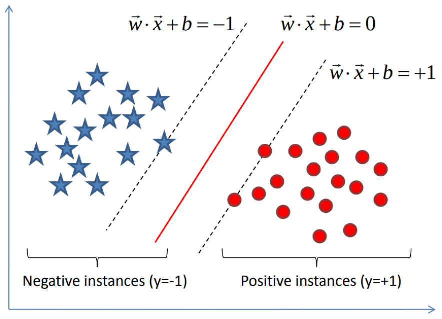

$\vec{w} \cdot \vec{x} + b = -1$
$\vec{w} \cdot \vec{x} + b = 0$
$\vec{w} \cdot \vec{x} + b = +1$
Negative instances (y=-1)
Positive instances (y=+1)

Computing the optimal intercept b^*


$\vec{w} \cdot \vec{x} + b = -1$
$\vec{w} \cdot \vec{x} + b = 0$
$\vec{w} \cdot \vec{x} + b = +1$
Negative instances (y=-1)
Positive instances (y=+1)

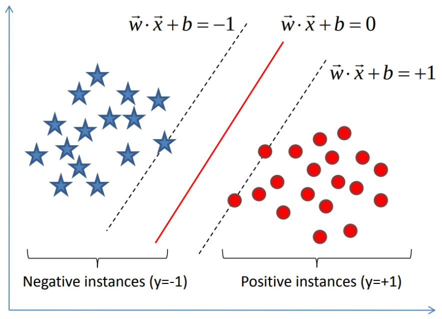
- For all x in red plane:$w^* \cdot x = -b$

Computing the optimal intercept b^*


$\vec{w} \cdot \vec{x} + b = -1$
$\vec{w} \cdot \vec{x} + b = 0$
$\vec{w} \cdot \vec{x} + b = +1$
Negative instances (y=-1)
Positive instances (y=+1)

- For all x in red plane: $w^* \cdot x = -b$
- Optimal b^* is such that two support vectors are equal distance

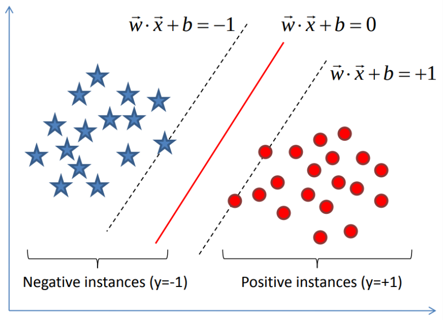
Computing the optimal intercept b^*


$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

- For all x in red plane:$w^* \cdot x = -b$
- Optimal b^* is such that two support vectors are equal distance

- We recall that b^* is such that $w^*x + b^* = 0$ is the middle plane

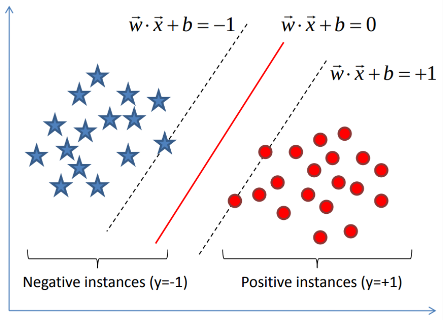
Computing the optimal intercept b^*


$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

- For all x in red plane:$w^* \cdot x = -b$
- Optimal b^* is such that two support vectors are equal distance

- We recall that b^* is such that $w^* x + b^* = 0$ is the middle plane
- Assume that the normal w^* points in positive class direction

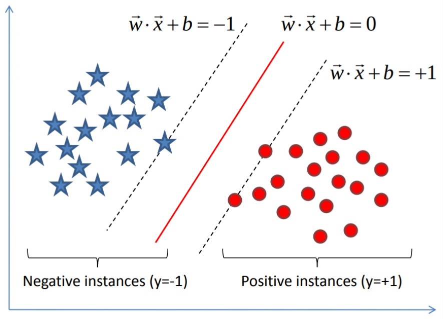
Computing the optimal intercept b^*


$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

- For all x in red plane:$w^* \cdot x = -b$
- Optimal b^* is such that two support vectors are equal distance

- We recall that b^* is such that $w^* x + b^* = 0$ is the middle plane
- Assume that the normal w^* points in positive class direction
- Take projections of $x^{(i)}$ of positive class on w^*, and find the least farthest

Computing the optimal intercept b^*

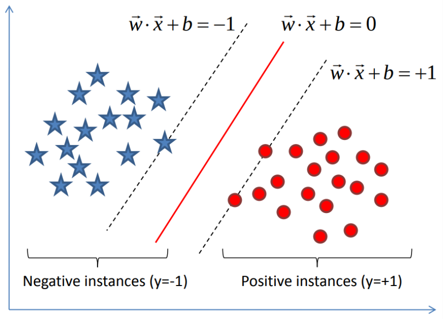

$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

- For all x in red plane:$w^* \cdot x = -b$
- Optimal b^* is such that two support vectors are equal distance

- We recall that b^* is such that $w^* x + b^* = 0$ is the middle plane
- Assume that the normal w^* points in positive class direction
- Take projections of $x^{(i)}$ of positive class on w^*, and find the least farthest

$$\min_{i:y^{(i)}=+1} \quad {w^*}^T x^{(i)}$$

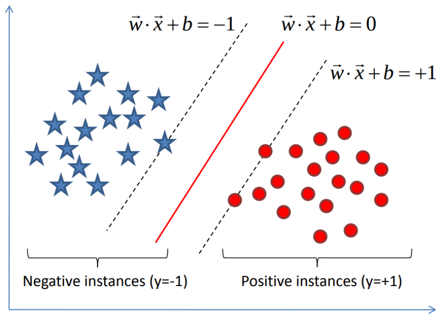
Computing the optimal intercept b^*


$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

- For all x in red plane:$w^* \cdot x = -b$
- Optimal b^* is such that two support vectors are equal distance

- We recall that b^* is such that $w^* x + b^* = 0$ is the middle plane
- Assume that the normal w^* points in positive class direction
- Take projections of $x^{(i)}$ of positive class on w^*, and find the least farthest
$$\min_{i:y^{(i)}=+1} \quad {w^*}^T x^{(i)}$$
- Similarly, take projections of $x^{(i)}$ of negative class along $-w^*$, and find the farthest point

Computing the optimal intercept b^*

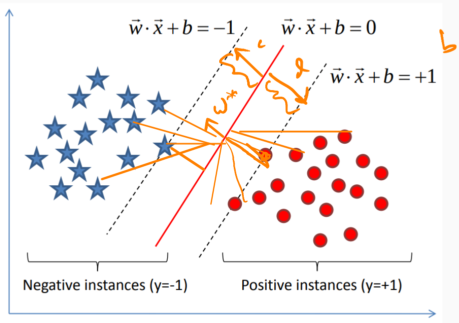

$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
Negative instances (y=-1)
Positive instances (y=+1)

- For all x in red plane:$w^* \cdot x = -b$
- Optimal b^* is such that two support vectors are equal distance

- We recall that b^* is such that $w^* x + b^* = 0$ is the middle plane
- Assume that the normal w^* points in positive class direction
- Take projections of $x^{(i)}$ of positive class on w^*, and find the least farthest
$$\min_{i:y^{(i)}=+1} \quad {w^*}^T x^{(i)}$$
- Similarly, take projections of $x^{(i)}$ of negative class along $-w^*$, and find the farthest point

$$\min_{i:y^{(i)}=-1} \; -{w^*}^T x^{(i)} = \max_{i:y^{(i)}=-1} \; {w^*}^T x^{(i)}$$

Computing the optimal intercept b^*



$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=0$
$\vec{w}\cdot\vec{x}+b=+1$
c
d
w^*
Negative instances (y=-1)
Positive instances (y=+1)

- For all x in red plane:$w^* \cdot x = -b$
- Optimal b^* is such that two support vectors are equal distance

$b = \frac{c+d}{2}$

- We recall that b^* is such that $w^* x + b^* = 0$ is the middle plane
- Assume that the normal w^* points in positive class direction
- Take projections of $x^{(i)}$ of positive class on w^*, and find the least farthest

$$\min_{i:y^{(i)}=+1} {w^*}^T x^{(i)}$$

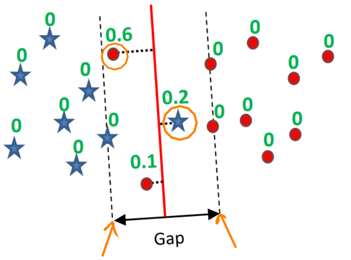
- Similarly, take projections of $x^{(i)}$ of negative class along $-w^*$, and find the farthest point

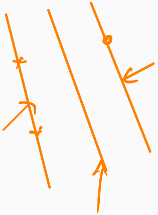
$$\min_{i:y^{(i)}=-1} -{w^*}^T x^{(i)} = \max_{i:y^{(i)}=-1} {w^*}^T x^{(i)}$$

$$\text{Hence, optimal intercept } b^* = -\frac{\overbrace{\max_{i:y^{(i)}=-1} {w^*}^T x^{(i)}}^{c} + \overbrace{\min_{i:y^{(i)}=+1} {w^*}^T x^{(i)}}^{d}}{2}$$

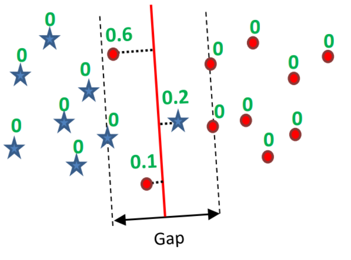
Linearly Separable with Noisy Data: Soft-Margin SVM

noisy?



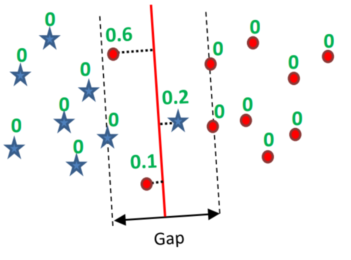


Linearly Separable with Noisy Data: Soft-Margin SVM



Two red and one blue samples noisy.

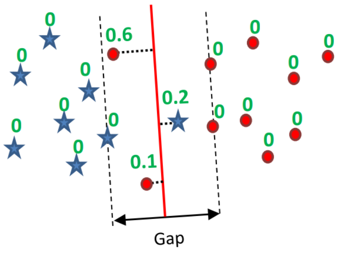
Linearly Separable with Noisy Data: Soft-Margin SVM



Two red and one blue samples noisy.

Goal: The data is noisy and it is not linearly separable. Modify SVM optimization model to allow for noisy data into account. Formulate optimization problem so that some noisy data is allowed bewteen gap/margin (region between dotted lines).

Linearly Separable with Noisy Data: Soft-Margin SVM

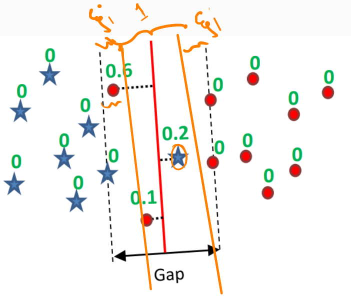


Two red and one blue samples noisy.

Goal: The data is noisy and it is not linearly separable. Modify SVM optimization model to allow for noisy data into account. Formulate optimization problem so that some noisy data is allowed bewteen gap/margin (region between dotted lines).

- Want to allow for noisy data, i.e., allow some sample x_i to be between dotted planes

Linearly Separable with Noisy Data: Soft-Margin SVM



$y_i(w\cdot x_i+b)\geq 1$

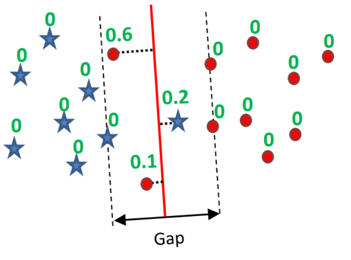
Two red and one blue samples noisy.

Goal: The data is noisy and it is not linearly separable. Modify SVM optimization model to allow for noisy data into account. Formulate optimization problem so that some noisy data is allowed bewteen gap/margin (region between dotted lines).

- Want to allow for noisy data, i.e., allow some sample x_i to be between dotted planes

$$y_i(w\cdot x_i+b)\geq 1-\xi_i$$

Linearly Separable with Noisy Data: Soft-Margin SVM



Two red and one blue samples noisy.

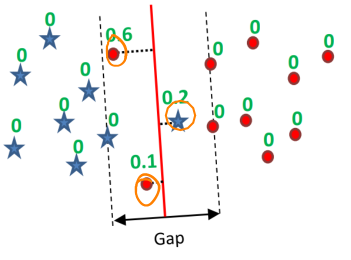
Goal: The data is noisy and it is not linearly separable. Modify SVM optimization model to allow for noisy data into account. Formulate optimization problem so that some noisy data is allowed bewteen gap/margin (region between dotted lines).

- Want to allow for noisy data, i.e., allow some sample x_i to be between dotted planes

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i$$

- Our margin is no longer strict or hard, it is now softened

Linearly Separable with Noisy Data: Soft-Margin SVM

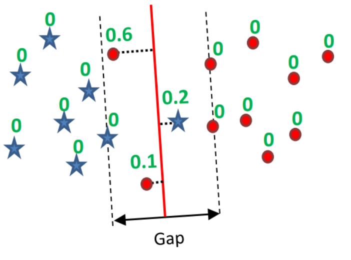


Two red and one blue samples noisy.

Goal: The data is noisy and it is not linearly separable. Modify SVM optimization model to allow for noisy data into account. Formulate optimization problem so that some noisy data is allowed bewteen gap/margin (region between dotted lines).

- Want to allow for noisy data, i.e., allow some sample x_i to be between dotted planes
$$y_i(w \cdot x_i + b) \geq 1 - \xi_i$$
- Our margin is no longer strict or hard, it is now softened
- Also, don't want too many x_i to be between dotted planes, introduce penalty in objective

Linearly Separable with Noisy Data: Soft-Margin SVM



Two red and one blue samples noisy.

Goal: The data is noisy and it is not linearly separable. Modify SVM optimization model to allow for noisy data into account. Formulate optimization problem so that some noisy data is allowed bewteen gap/margin (region between dotted lines).

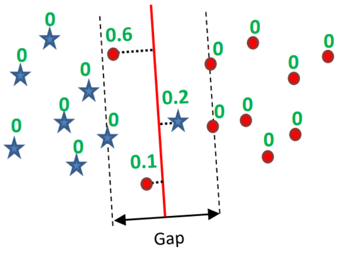
- Want to allow for noisy data, i.e., allow some sample x_i to be between dotted planes

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i$$

- Our margin is no longer strict or hard, it is now softened
- Also, don't want too many x_i to be between dotted planes, introduce penalty in objective

$$\min \quad \underbrace{\frac{1}{2}\|w\|^2} + C \underbrace{\sum_{i=1}^{m} \xi_i}$$

Linearly Separable with Noisy Data: Soft-Margin SVM



Two red and one blue samples noisy.

Goal: The data is noisy and it is not linearly separable. Modify SVM optimization model to allow for noisy data into account. Formulate optimization problem so that some noisy data is allowed bewteen gap/margin (region between dotted lines).

- Want to allow for noisy data, i.e., allow some sample x_i to be between dotted planes
$$y_i(w \cdot x_i + b) \geq 1 - \xi_i$$
- Our margin is no longer strict or hard, it is now softened
- Also, don't want too many x_i to be between dotted planes, introduce penalty in objective
$$\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{m} \xi_i$$
- Since it is a minimization, large values of ξ_i will be discouraged

Primal and Dual Formulation of Soft-Margin SVM

Primal Soft-Margin SVM:

$$\begin{aligned}\text{minimize} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{subject to} \quad & y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m.\end{aligned}$$

Primal and Dual Formulation of Soft-Margin SVM

Primal Soft-Margin SVM:

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\
&\text{subject to} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m.
\end{aligned}$$

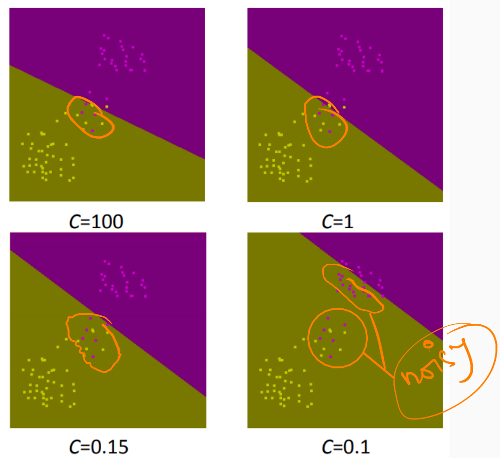
Dual Soft-Margin SVM:

$$\begin{aligned}
&\text{Maximize} \quad \sum_{i=1}^{m}\lambda_i - \frac{1}{2}\sum_{i,j=1}^{m}\lambda_i\lambda_j\, y_i y_j\, x_i \cdot x_j \\
&\text{subject to} \quad 0 \leq \lambda_i \leq C, \quad i = 1, \ldots, m, \\
&\qquad\qquad\quad \sum_{i=1}^{m}\lambda_i y_i = 0.
\end{aligned}$$

Quiz: How did we get the dual form? Try.

Effect of Parameter on soft-margin SVM

Minimize $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i$ subject to $y_i(w\cdot x_i + b) \geq 1 - \xi_i$ for $i = 1,\ldots,m$.



C=100

C=1

C=0.15

C=0.1

- For C very large, soft margin is equivalent to hard margin.
- When C is small, we allow misclassification.
- Here C is a hyperparameter.
- In practice, cross-validations can be used.

Primal and Dual Formulation of Soft-Margin SVM

Primal Soft-Margin SVM:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{subject to} \quad & y_i(w \cdot x_i + b) \geq 1 - \xi_i, \\ & i = 1, \ldots, m. \end{aligned}$$

Primal and Dual Formulation of Soft-Margin SVM

Primal Soft-Margin SVM:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \\ \text{subject to} \quad & y_i(w \cdot x_i + b) \geq 1 - \xi_i, \\ & i = 1, \ldots, m. \end{aligned}$$

Dual Soft-Margin SVM:

$$\begin{aligned} \text{Maximize} \quad & \sum_{i=1}^{m} \lambda_i - \frac{1}{2}\sum_{i,j=1}^{m} \lambda_i\lambda_j \, y_i y_j \, x_i \cdot x_j \\ \text{subject to} \quad & 0 \leq \lambda_i \leq C, \quad i = 1, \ldots, m, \\ & \sum_{i=1}^{m} \lambda_i y_i = 0. \end{aligned}$$

Additional advantages of primal over dual

- We see $x_i \cdot x_j$ in dual.

Primal and Dual Formulation of Soft-Margin SVM

Primal Soft-Margin SVM:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \\ \text{subject to} \quad & y_i(w \cdot x_i + b) \geq 1 - \xi_i, \\ & i = 1, \ldots, m. \end{aligned}$$

Dual Soft-Margin SVM:

$$\begin{aligned} \text{Maximize} \quad & \sum_{i=1}^{m} \lambda_i - \frac{1}{2}\sum_{i,j=1}^{m} \lambda_i\lambda_j \, y_i y_j \, \underline{x_i \cdot x_j} \\ \text{subject to} \quad & 0 \leq \lambda_i \leq C, \quad i = 1, \ldots, m, \\ & \sum_{i=1}^{m} \lambda_i y_i = 0. \end{aligned}$$

Additional advantages of ~~primal~~ dual over ~~dual~~ primal

- We see $\underline{x_i \cdot x_j}$ in dual.
- Can we make use of this term?

Primal and Dual Formulation of Soft-Margin SVM

Primal Soft-Margin SVM:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{subject to} \quad & y_i(w \cdot x_i + b) \geq 1 - \xi_i, \\ & i = 1, \ldots, m. \end{aligned}$$

Dual Soft-Margin SVM:

$$\begin{aligned} \text{Maximize} \quad & \sum_{i=1}^{m}\lambda_i - \frac{1}{2}\sum_{i,j=1}^{m}\lambda_i\lambda_j\, y_i y_j\, x_i \cdot x_j \\ \text{subject to} \quad & 0 \leq \lambda_i \leq C, \quad i = 1, \ldots, m, \\ & \sum_{i=1}^{m}\lambda_i y_i = 0. \end{aligned}$$

Additional advantages of primal over dual

- We see $x_i \cdot x_j$ in dual.
- Can we make use of this term?
- Sample x_i also appears in constraint in primal. Is that useful?

Primal and Dual Formulation of Soft-Margin SVM

Primal Soft-Margin SVM:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{subject to} \quad & y_i(w \cdot x_i + b) \geq 1 - \xi_i, \\ & i = 1, \ldots, m. \end{aligned}$$

Dual Soft-Margin SVM:

$$\begin{aligned} \text{Maximize} \quad & \sum_{i=1}^{m}\lambda_i - \frac{1}{2}\sum_{i,j=1}^{m}\lambda_i\lambda_j \, y_i y_j \, x_i \cdot x_j \\ \text{subject to} \quad & 0 \leq \lambda_i \leq C, \quad i = 1, \ldots, m, \\ & \sum_{i=1}^{m}\lambda_i y_i = 0. \end{aligned}$$

Additional advantages of primal over dual

- We see $x_i \cdot x_j$ in dual.
- Can we make use of this term?
- Sample x_i also appears in constraint in primal. Is that useful?
- Fact that $x_i \cdot x_j$ appears only in dual objective and that too as dot product is useful.

Primal and Dual Formulation of Soft-Margin SVM

Primal Soft-Margin SVM:

$$
\begin{aligned}
\text{minimize} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \\
\text{subject to} \quad & y_i(w \cdot x_i + b) \geq 1 - \xi_i, \\
& i = 1, \ldots, m.
\end{aligned}
$$

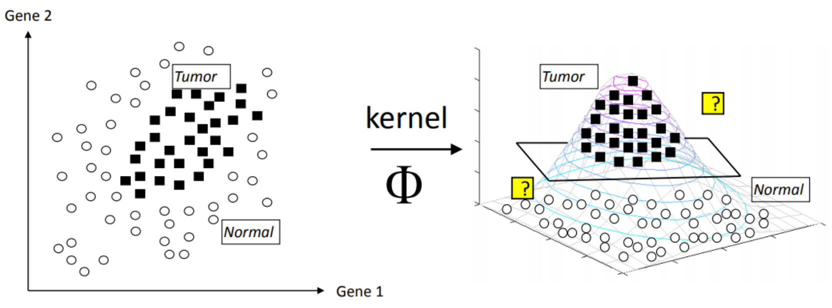
Dual Soft-Margin SVM:

$$
\begin{aligned}
\text{Maximize} \quad & \sum_{i=1}^{m} \lambda_i - \frac{1}{2}\sum_{i,j=1}^{m} \lambda_i \lambda_j \, y_i y_j \, \underline{x_i \cdot x_j} \\
\text{subject to} \quad & 0 \leq \lambda_i \leq C, \quad i = 1, \ldots, m, \\
& \sum_{i=1}^{m} \lambda_i y_i = 0.
\end{aligned}
$$

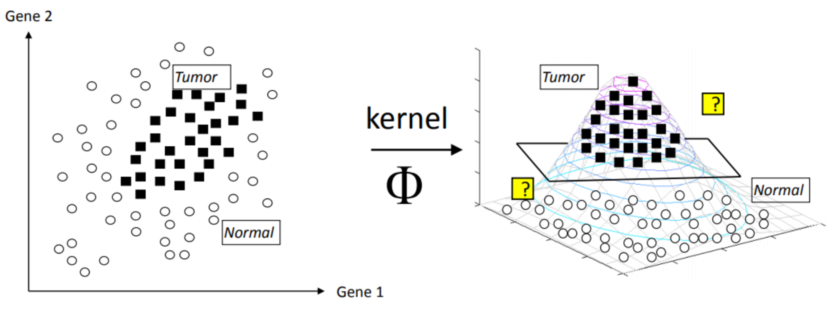
Additional advantages of primal over dual

- We see $x_i \cdot x_j$ in dual.
- Can we make use of this term?
- Sample x_i also appears in constraint in primal. Is that useful?
- Fact that $x_i \cdot x_j$ appears only in dual objective and that too as dot product is useful.
- This allows mapping samples to a space where they may be linearly separated.

Non-Linearly Separable Data, Kernel Trick

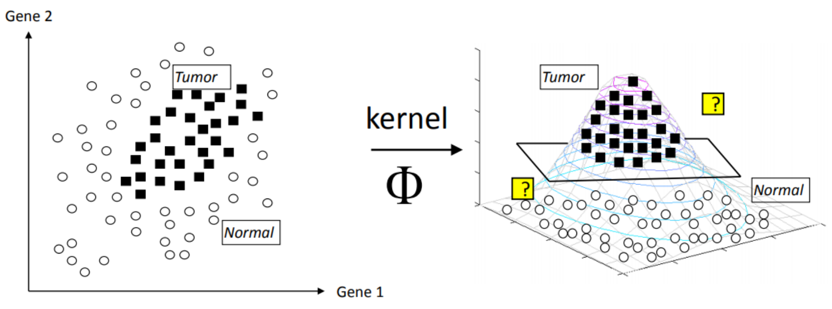


Non-Linearly Separable Data, Kernel Trick



Here data is not linearly separable in the input space

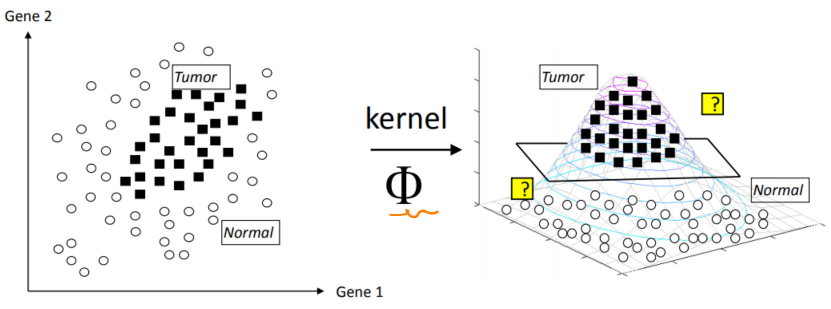
Non-Linearly Separable Data, Kernel Trick



Here data is not linearly separable in the input space

After mapping to a new feature space, data is now linearly separable!

Non-Linearly Separable Data, Kernel Trick



Here data is not linearly separable in the input space

After mapping to a new feature space, data is now linearly separable!

Quiz: How to find such map $\Phi : \mathbb{R}^N \to H$

Kernel Trick

For data in input space

$$f(x) = \text{sign}(w \cdot x + b)$$

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

Kernel Trick

$x \longrightarrow \phi(x)$

For data in input space

$$f(x) = \text{sign}(w \cdot x + b)$$

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

Data in feature space $\Phi(x)$

$$f(x) = \text{sign}(w \cdot \Phi(x) + b)$$

$$w = \sum_{i=1}^{m} \alpha_i y_i \Phi(x_i)$$

- Solve the dual problem for α_i

Kernel Trick

For data in input space

$$f(x) = \text{sign}(w \cdot x + b)$$

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

Data in feature space $\Phi(x)$

$$f(x) = \text{sign}(w \cdot \Phi(x) + b)$$

$$w = \sum_{i=1}^{m} \alpha_i y_i \Phi(x_i)$$

- Solve the dual problem for α_i
- Compute w, b from above

Kernel Trick

For data in input space

$$f(x) = \text{sign}(w \cdot x + b)$$

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

Data in feature space $\Phi(x)$

$$f(x) = \text{sign}(w \cdot \Phi(x) + b)$$

$$w = \sum_{i=1}^{m} \alpha_i y_i \Phi(x_i)$$

- Solve the dual problem for α_i
- Compute w, b from above
- Predict using $f(x)$ above

Kernel Trick

For data in input space

$$f(x) = \text{sign}(w \cdot x + b)$$

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

- Solve the dual problem for α_i
- Compute w, b from above
- Predict using $f(x)$ above

Data in feature space $\Phi(x)$

$$f(x) = \text{sign}(w \cdot \Phi(x) + b)$$

$$w = \sum_{i=1}^{m} \alpha_i y_i \Phi(x_i)$$

- Solve for α_i using dual

Kernel Trick

For data in input space

$$f(x) = \text{sign}(w \cdot x + b)$$

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

- Solve the dual problem for α_i
- Compute w, b from above
- Predict using $f(x)$ above

Data in feature space $\Phi(x)$

$$f(x) = \text{sign}(w \cdot \Phi(x) + b)$$

$$w = \sum_{i=1}^{m} \alpha_i y_i \Phi(x_i)$$

- Solve for α_i using dual
- Compute w, b (x_i by $\Phi(x_i)$)

Kernel Trick

For data in input space

$$f(x) = \text{sign}(w \cdot x + b)$$

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

- Solve the dual problem for α_i
- Compute w, b from above
- Predict using $f(x)$ above

Data in feature space $\Phi(x)$

$$f(x) = \text{sign}(w \cdot \Phi(x) + b)$$

$$w = \sum_{i=1}^{m} \alpha_i y_i \Phi(x_i)$$

- Solve for α_i using dual
- Compute w, b (x_i by $\Phi(x_i)$)
- Predict with $f(x)$. So what? Trick: substitute w in $f(x)$ above!

Kernel Trick:

$$f(x) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i y_i \, \Phi(x_i) \cdot \Phi(x) + b\right) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i y_i \, K(x_i, x) + b\right)$$

Kernel Trick

$x \longrightarrow \underline{\underline{\Phi(x)}}$

For data in input space

$$f(x) = \text{sign}(w \cdot x + b)$$

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

- Solve the dual problem for α_i
- Compute w, b from above
- Predict using $f(x)$ above

Data in feature space $\Phi(x)$

$$f(x) = \text{sign}(w \cdot \Phi(x) + b)$$

$$w = \sum_{i=1}^{m} \alpha_i y_i \Phi(x_i)$$

- Solve for α_i using dual
- Compute w, b (x_i by $\Phi(x_i)$)
- Predict with $f(x)$. So what? Trick: substitute w in $f(x)$ above!

Kernel Trick:

$$f(x) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i y_i \, \underline{\Phi}(x_i) \cdot \underline{\Phi}(x) + b\right) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i y_i \, K(x_i, x) + b\right)$$

Dont need to know Φ; Only a function $K(\cdot,\cdot) : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$ will do!

Kernel Trick

$$\boxed{x \in \mathbb{R}^2} \rightarrow \boxed{\phi(x) \in \mathbb{R}^3}$$

$$\underbrace{\phi(x_i)}_{\in \mathbb{R}^3} \cdot \underbrace{\phi(x)}_{\mathbb{R}^3}$$

For data in input space

$$f(x) = \text{sign}(w \cdot x + b)$$
$$w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

- Solve the dual problem for α_i
- Compute w, b from above
- Predict using $f(x)$ above

Data in feature space $\Phi(x)$

$$f(x) = \text{sign}(w \cdot \Phi(x) + b)$$
$$w = \sum_{i=1}^{m} \alpha_i y_i \Phi(x_i)$$

- Solve for α_i using dual
- Compute w, b (x_i by $\Phi(x_i)$)
- Predict with $f(x)$. So what? Trick: substitute w in $f(x)$ above!

Kernel Trick:

$$f(x) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i y_i \, \Phi(x_i) \cdot \Phi(x) + b\right) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i y_i \boxed{K(x_i, x)} + b\right)$$

Dont need to know Φ; Only a function $K(\cdot,\cdot) : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$ will do!

→ Avoid explicit computation of features or infinite length vectors

Popular Kernels

Kernel: A kernel is a dot product in some feature space

$$\underline{K(x_i, x_j)} = \Phi(x_i) \cdot \Phi(x_j)$$

Popular Kernels

Kernel: A kernel is a dot product in some feature space

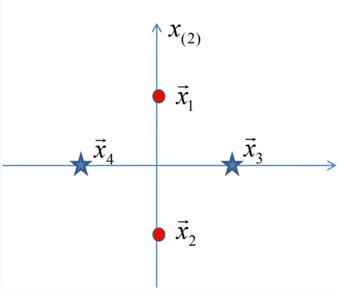
$$K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$$

$$e^{-\gamma \underbrace{(x_i - x_j)^T (x_i - x_j)}}$$

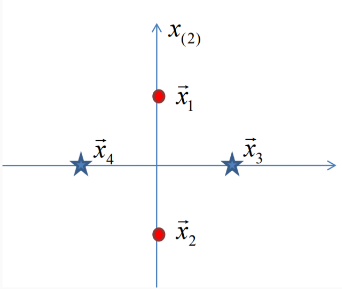
✓ $K(x_i, x_j) = \underline{x_i} \cdot \underline{x_j}$	Linear kernel
✓ $K(x_i, x_j) = \exp(-\gamma \lVert \underline{x_i - x_j} \rVert^2)$	Gaussian kernel
✓ $K(x_i, x_j) = \exp(-\gamma \lVert x_i - x_j \rVert)$	Exponential kernel
✓ $K(x_i, x_j) = (p + \underline{x_i \cdot x_j})^q$	Polynomial kernel
✓ $\underline{K(x_j, x_j)} = (p + \underline{x_i \cdot x_j})^q \exp(-\gamma \lVert x_i - x_j \rVert^2)$	Hybrid kernel
$K(x_i, x_j) = \tanh(kx_i \cdot x_j - \delta)$	Sigmoidal

Polynomial Kernel

Polynomial Kernel

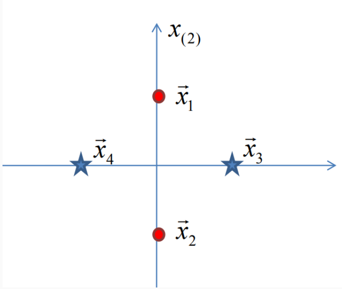


Polynomial Kernel



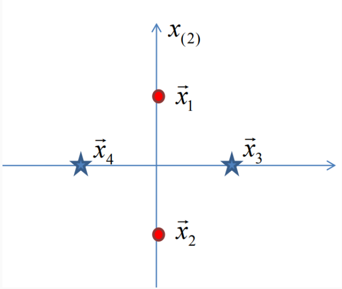
- Data is not linearly separable

Polynomial Kernel



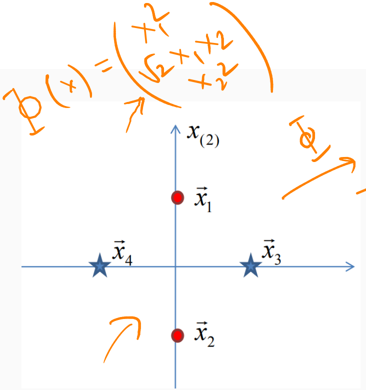
- Data is not linearly separable
- Apply Kernel $\underline{K(x, z)} = (x \cdot z)^2$ to map data to higher dimension, to make it linearly separable

Polynomial Kernel



- Data is not linearly separable
- Apply Kernel $K(x, z) = (x \cdot z)^2$ to map data to higher dimension, to make it linearly separable
- $x_1 = [x_{(1)}, x_{(2)}]^T$, $z_1 = [z_{(1)}, z_{(2)}]^T$

Polynomial Kernel



- Data is not linearly separable
- Apply Kernel $K(x,z) = (x\cdot z)^2$ to map data to higher dimension, to make it linearly separable
- $x_1 = [x_{(1)}, x_{(2)}]^T, \ z_1 = [z_{(1)}, z_{(2)}]^T$

much cheaper than

Cost=4

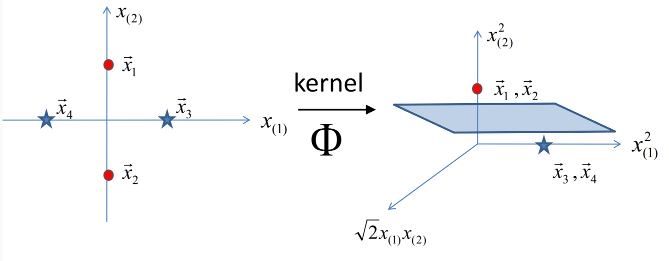
Cost=11

We have

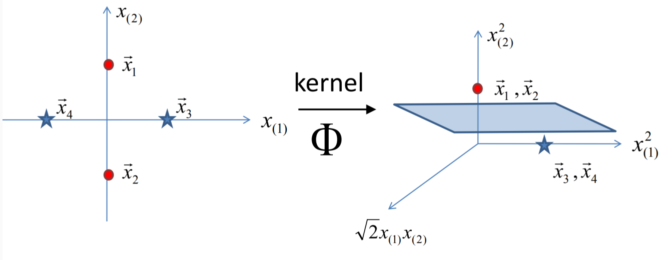
$$K(x\cdot z) = (x\cdot z)^2 = \left(\begin{bmatrix}x_{(1)}\\x_{(2)}\end{bmatrix}\cdot\begin{bmatrix}z_{(1)}\\z_{(2)}\end{bmatrix}\right)^2 = (x_{(1)}z_{(1)} + x_{(2)}z_{(2)})^2$$

$$= x_{(1)}^2 z_{(1)}^2 + 2x_{(1)}z_{(1)}x_{(2)}z_{(2)} + x_{(2)}^2 z_{(2)}^2 = \begin{bmatrix}x_{(1)}^2\\ \sqrt{2}x_{(1)}x_{(2)}\\ x_{(2)}^2\end{bmatrix}\cdot\begin{bmatrix}z_{(1)}^2\\ \sqrt{2}z_{(1)}z_{(2)}\\ z_{(2)}^2\end{bmatrix} = \Phi(x)\cdot\Phi(z)$$

Separation Happens After Using Polynomial Kernel!

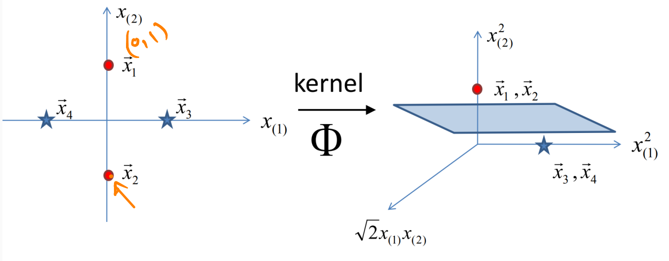


Separation Happens After Using Polynomial Kernel!



- In last slide: $\Phi(x) = \begin{bmatrix} x_{(1)}^2 \\ \sqrt{2}x_{(1)}x_{(2)} \\ x_{(2)}^2 \end{bmatrix}$

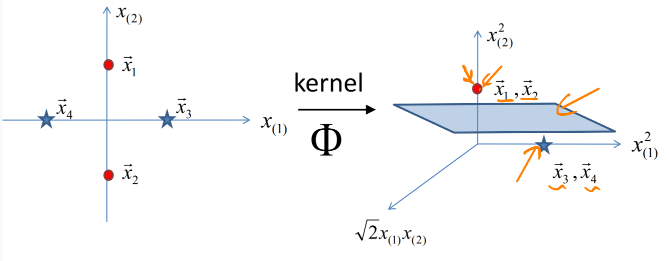
Separation Happens After Using Polynomial Kernel!



- In last slide: $\Phi(x) = \begin{bmatrix} x_{(1)}^2 \\ \sqrt{2}x_{(1)}x_{(2)} \\ x_{(2)}^2 \end{bmatrix}$
- The points in figure above are

$$x_1 = (0,1), \quad x_2 = (0,-1)$$
$$x_3 = (1,0), \quad x_4 = (-1,0)$$

Separation Happens After Using Polynomial Kernel!



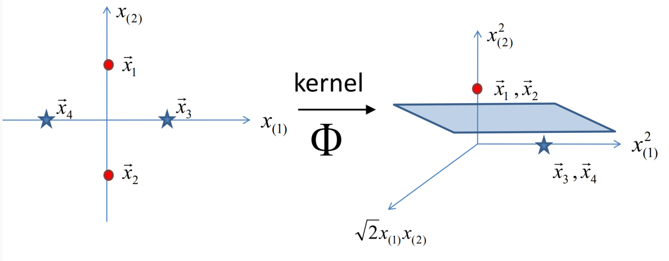
- In last slide: $\Phi(x) = \begin{bmatrix} x_{(1)}^2 \\ \sqrt{2}x_{(1)}x_{(2)} \\ x_{(2)}^2 \end{bmatrix} \begin{matrix} 0 \\ 0 \\ 1 \end{matrix}$
- The points in figure above are

$$x_1 = (0, 1), \quad x_2 = (0, -1)$$
$$x_3 = (1, 0), \quad x_4 = (-1, 0)$$

- We have

$$\Phi(x_1) = [0, 0, 1]^T$$
$$\Phi(x_2) = [0, 0, 1]^T$$
$$\Phi(x_3) = [1, 0, 0]^T$$
$$\Phi(x_4) = [1, 0, 0]^T$$

Separation Happens After Using Polynomial Kernel!



- In last slide: $\Phi(x) = \begin{bmatrix} x_{(1)}^2 \\ \sqrt{2}x_{(1)}x_{(2)} \\ x_{(2)}^2 \end{bmatrix}$
- The points in figure above are

$$x_1 = (0, 1), \quad x_2 = (0, -1)$$
$$x_3 = (1, 0), \quad x_4 = (-1, 0)$$

- We have

$$\Phi(x_1) = [0, 0, 1]^T$$
$$\Phi(x_2) = [0, 0, 1]^T$$
$$\Phi(x_3) = [1, 0, 0]^T$$
$$\Phi(x_4) = [1, 0, 0]^T$$

- After Φ transformation, points are now linearly separable!

Same Kernel Function but Different Features

Note: Feature space might not be unique for the same kernel function.

Same Kernel Function but Different Features

Note: Feature space might not be unique for the same kernel function.

Example-1: Consider $\Phi : \mathbb{R}^2 \to \mathbb{R}^3$

$$\begin{aligned}
\Phi(x_1, x_2) &= (x_1^2, x_2^2, \sqrt{2}x_2x_2) \\
\Phi(x) \cdot \Phi(z) &= (x_1^2, x_2^2, \sqrt{2}x_1x_2) \cdot (z_1^2, z_2^2, \sqrt{2}z_1z_2) \\
&= (x_1z_1 + x_2z_2)^2 = (x \cdot z)^2 = \underline{K(x, z)}.
\end{aligned}$$

Same Kernel Function but Different Features

Note: Feature space might not be unique for the same kernel function.

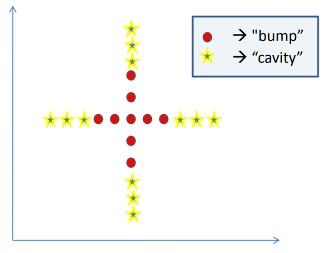
Example-1: Consider $\Phi : \mathbb{R}^2 \to \mathbb{R}^3$

$$\begin{aligned}\Phi(x_1, x_2) &= (x_1^2, x_2^2, \sqrt{2}x_2x_2)\\ \Phi(x) \cdot \Phi(z) &= (x_1^2, x_2^2, \sqrt{2}x_1x_2) \cdot (z_1^2, z_2^2, \sqrt{2}z_1z_2)\\ &= (x_1z_1 + x_2z_2)^2 = (x \cdot z)^2 = K(x, z).\end{aligned}$$

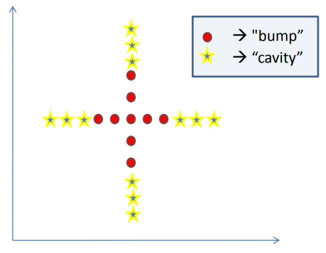
Example-2: Consider $\Phi : \mathbb{R}^2 \to \mathbb{R}^4$

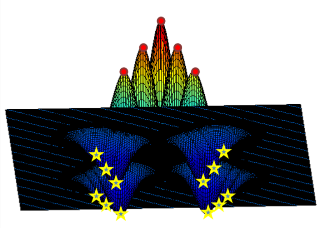
$$\begin{aligned}\Phi(x_1, x_2) &= (x_1^2, x_2^2, x_1x_2, x_2x_1)\\ \Phi(x) \cdot \Phi(z) &= (x_1^2, x_2^2, x_1x_2, x_2x_1) \cdot (z_1^2, z_2^2, z_1z_2, z_2z_1)\\ &= (x \cdot z)^2 = \underline{K(x, z)}.\end{aligned}$$

RBF or Gaussian Kernel

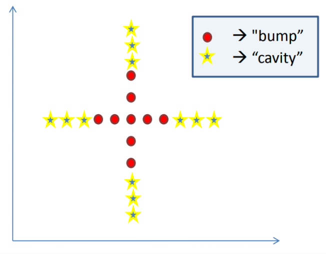


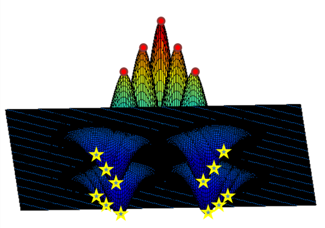
RBF or Gaussian Kernel

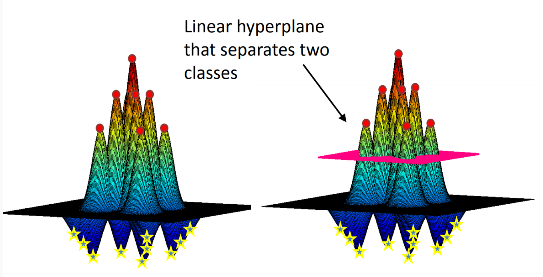




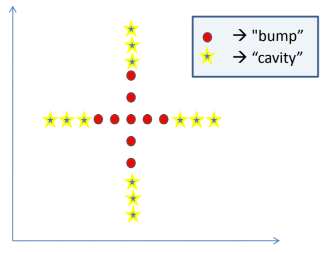
RBF or Gaussian Kernel

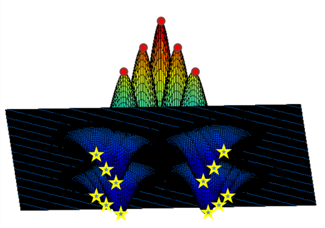


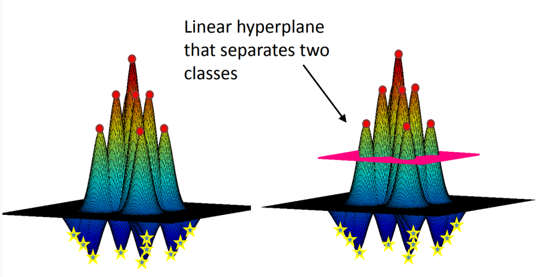



Linear hyperplane that separates two classes

RBF or Gaussian Kernel

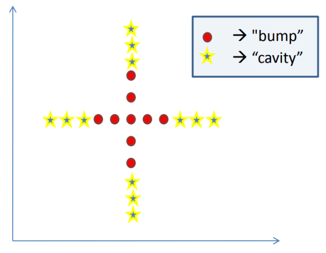


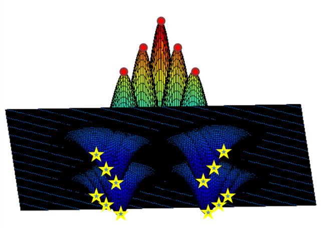


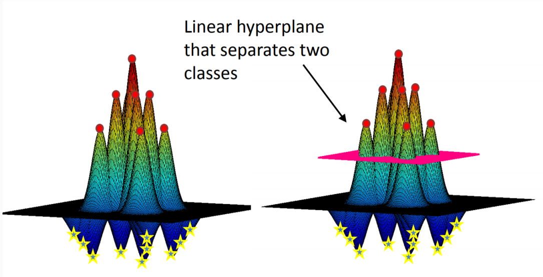


After applying the kernel, we observe:

RBF or Gaussian Kernel



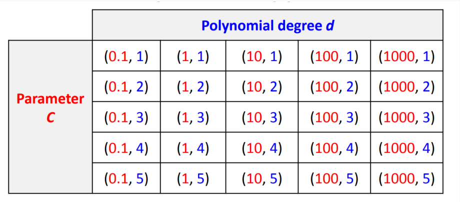



Linear hyperplane that separates two classes

After applying the kernel, we observe:

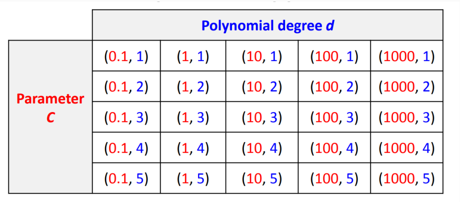
- red data points are bumped up
- yellow data points are pushed to cavity

Model Selection: Which Kernels with which parameter?



	Polynomial degree *d*				
Parameter *C*	(0.1, 1)	(1, 1)	(10, 1)	(100, 1)	(1000, 1)
	(0.1, 2)	(1, 2)	(10, 2)	(100, 2)	(1000, 2)
	(0.1, 3)	(1, 3)	(10, 3)	(100, 3)	(1000, 3)
	(0.1, 4)	(1, 4)	(10, 4)	(100, 4)	(1000, 4)
	(0.1, 5)	(1, 5)	(10, 5)	(100, 5)	(1000, 5)

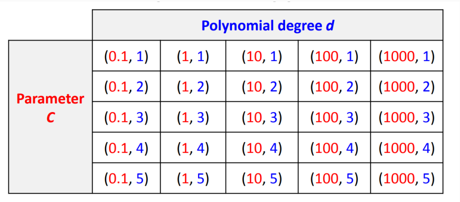
Model Selection: Which Kernels with which parameter?



	Polynomial degree *d*				
Parameter *C*	(0.1, 1)	(1, 1)	(10, 1)	(100, 1)	(1000, 1)
	(0.1, 2)	(1, 2)	(10, 2)	(100, 2)	(1000, 2)
	(0.1, 3)	(1, 3)	(10, 3)	(100, 3)	(1000, 3)
	(0.1, 4)	(1, 4)	(10, 4)	(100, 4)	(1000, 4)
	(0.1, 5)	(1, 5)	(10, 5)	(100, 5)	(1000, 5)

- Consider polynomial kernel.

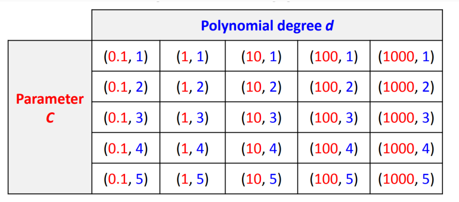
Model Selection: Which Kernels with which parameter?



	Polynomial degree *d*				
Parameter *C*	(0.1, 1)	(1, 1)	(10, 1)	(100, 1)	(1000, 1)
	(0.1, 2)	(1, 2)	(10, 2)	(100, 2)	(1000, 2)
	(0.1, 3)	(1, 3)	(10, 3)	(100, 3)	(1000, 3)
	(0.1, 4)	(1, 4)	(10, 4)	(100, 4)	(1000, 4)
	(0.1, 5)	(1, 5)	(10, 5)	(100, 5)	(1000, 5)

- Consider polynomial kernel.
- There are parameters: C, d.

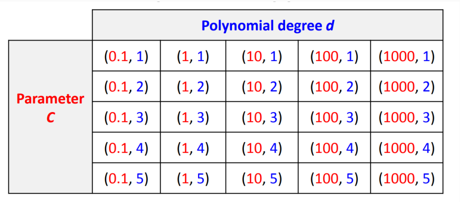
Model Selection: Which Kernels with which parameter?



	Polynomial degree *d*				
Parameter *C*	(0.1, 1)	(1, 1)	(10, 1)	(100, 1)	(1000, 1)
	(0.1, 2)	(1, 2)	(10, 2)	(100, 2)	(1000, 2)
	(0.1, 3)	(1, 3)	(10, 3)	(100, 3)	(1000, 3)
	(0.1, 4)	(1, 4)	(10, 4)	(100, 4)	(1000, 4)
	(0.1, 5)	(1, 5)	(10, 5)	(100, 5)	(1000, 5)

- Consider polynomial kernel.
- There are parameters: C, d.
- We can consider possible values.

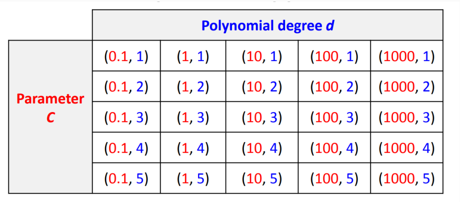
Model Selection: Which Kernels with which parameter?



	Polynomial degree *d*				
Parameter *C*	(0.1, 1)	(1, 1)	(10, 1)	(100, 1)	(1000, 1)
	(0.1, 2)	(1, 2)	(10, 2)	(100, 2)	(1000, 2)
	(0.1, 3)	(1, 3)	(10, 3)	(100, 3)	(1000, 3)
	(0.1, 4)	(1, 4)	(10, 4)	(100, 4)	(1000, 4)
	(0.1, 5)	(1, 5)	(10, 5)	(100, 5)	(1000, 5)

- Consider polynomial kernel.
- There are parameters: C, d.
- We can consider possible values.
- Check our choice by doing cross-validations.

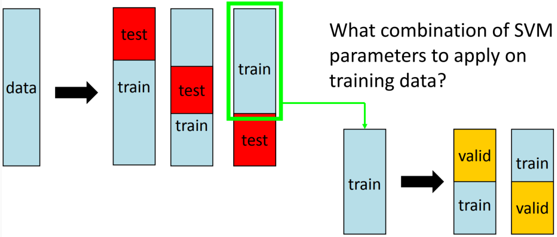
Model Selection: Which Kernels with which parameter?



	Polynomial degree *d*				
Parameter *C*	(0.1, 1)	(1, 1)	(10, 1)	(100, 1)	(1000, 1)
	(0.1, 2)	(1, 2)	(10, 2)	(100, 2)	(1000, 2)
	(0.1, 3)	(1, 3)	(10, 3)	(100, 3)	(1000, 3)
	(0.1, 4)	(1, 4)	(10, 4)	(100, 4)	(1000, 4)
	(0.1, 5)	(1, 5)	(10, 5)	(100, 5)	(1000, 5)

- Consider polynomial kernel.
- There are parameters: C, d.
- We can consider possible values.
- Check our choice by doing cross-validations.

Recall the main idea of cross-validation:



What combination of SVM parameters to apply on training data?

Perform “grid search” using another nested loop of cross-validation.

SVM in Unconstrained Form: loss + penalty form

Primal Soft-Margin SVM:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{subject to} \quad & y_i(w \cdot x_i + b) \geq 1 - \xi_i, \\ & i = 1, \ldots, m. \end{aligned}$$

SVM in Unconstrained Form: loss + penalty form

Primal Soft-Margin SVM:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{subject to} \quad & y_i(w \cdot x_i + b) \geq 1 - \xi_i, \\ & i = 1, \ldots, m. \end{aligned}$$

1. Write constraint as

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i$$

$$\underbrace{\xi_i \geq -(w \cdot x_i}_{} + b)y_i + 1$$

SVM in Unconstrained Form: loss + penalty form

Primal Soft-Margin SVM:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{subject to} \quad & y_i(w\cdot x_i + b) \geq 1 - \xi_i, \\ & i = 1, \ldots, m. \end{aligned}$$

1. Write constraint as

$$\begin{aligned} y_i(w\cdot x_i + b) &\geq 1 - \xi_i \\ \xi_i &\geq -\underbrace{(w\cdot x_i + b)}_{f(x_i)}y_i + 1 \end{aligned}$$

and $\xi_i \geq 0$.

2. Write constraints on ξ as function?

$$\xi_i = \max(0, 1 - y_i f(x_i))$$

max is not diff.

SVM in Unconstrained Form: loss + penalty form

Primal Soft-Margin SVM:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{subject to} \quad & y_i(w\cdot x_i + b) \geq 1 - \xi_i, \\ & i = 1, \ldots, m. \end{aligned}$$

1. Write constraint as

$$\begin{aligned} y_i(w\cdot x_i + b) &\geq 1 - \xi_i \\ \xi_i &\geq -(w\cdot x_i + b)y_i + 1 \end{aligned}$$

and $\xi_i \geq 0$.

2. Write constraints on ξ as function?

$$\xi_i = \max(0, 1 - y_i f(x_i))$$

3. Substituting ξ_i in objective

$$\begin{aligned} &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\max(0, 1 - y_i f(x_i)) \\ &= \frac{1}{2C}\|w\|^2 + \sum_{i=1}^{m}\max(0, 1 - y_i f(x_i)) \\ &= \lambda\|w\|^2 + \sum_{i=1}^{m}\max(0, 1 - y_i f(x_i)) \\ &= \lambda\|w\|^2 + \sum_{i=1}^{m}[0, 1 - y_i f(x_i)]_+ \end{aligned}$$

$\lambda\left\{\frac{1}{2C}\|w\|^2 + \sum\right.$

$\boxed{\lambda = \frac{1}{2C}}$

Unconstrained SVM: Find w, b s.t. minimize $\sum_{i=1}^{m}[1 - y_i f(x_i)]_+ + \lambda\|w\|_2^2$

w b

SVM in Unconstrained Form: loss + penalty form

Primal Soft-Margin SVM:

$$\begin{aligned}\text{minimize} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{subject to} \quad & y_i(w\cdot x_i + b) \geq 1 - \xi_i, \\ & i = 1,\ldots,m.\end{aligned}$$

1. Write constraint as

$$\begin{aligned} y_i(w\cdot x_i + b) &\geq 1 - \xi_i \\ \xi_i &\geq -(w\cdot x_i + b)y_i + 1 \end{aligned}$$

and $\xi_i \geq 0$.

2. Write constraints on ξ as function?

$$\xi_i = \max(0, 1 - y_i f(x_i))$$

3. Substituting ξ_i in objective

$$\begin{aligned} &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\max(0, 1 - y_i f(x_i)) \\ &= \frac{1}{2C}\|w\|^2 + \sum_{i=1}^{m}\max(0, 1 - y_i f(x_i)) \\ &= \lambda\|w\|^2 + \sum_{i=1}^{m}\max(0, 1 - y_i f(x_i)) \\ &= \lambda\|w\|^2 + \sum_{i=1}^{m}[0, 1 - y_i f(x_i)]_+ \end{aligned}$$

Unconstrained SVM: Find w, b s.t. minimize $\sum_{i=1}^{m}[1 - y_i f(x_i)]_+ + \lambda\|w\|_2^2$

Penalty: $\lambda\|w\|_2^2$.

Loss (Hinge Loss): $\sum_{i=1}^{m}[1 - y_i f(x_i)]_+$.

Variety of Loss+Penalty Formulations for SVM

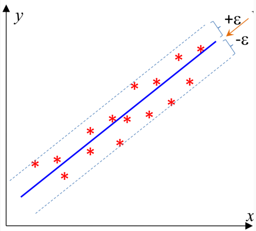
Loss	Penalty function	Resulting algorithm
Hinge Loss: $\sum_{i=1}^{m}[1 - y_i f(x_i)]_+$	$\lambda \lVert w \rVert_2^2$	SVM
Mean Squared Error $\sum_{i=1}^{m}(y_i - f(x_i))^2$	$\lambda \lVert w \rVert_2^2$	Ridge Regression
Mean Squared Error $\sum_{i=1}^{m}(y_i - f(x_i))^2$	$\lambda \lVert w \rVert_1$	Lasso
Mean Squared Error $\sum_{i=1}^{m}(y_i - f(x_i))^2$	$\lambda_1 \lVert w \rVert_1 + \lambda_2 \lVert w \rVert_2^2$	Elastic Net
Hinge Loss: $\sum_{i=1}^{m}[1 - y_i f(x_i)]_+$	$\lambda \lVert w \rVert_1$	1-Norm SVM

Variety of Loss+Penalty Formulations for SVM

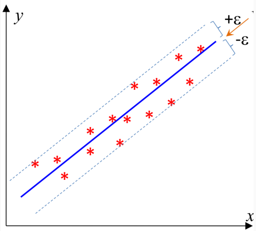
Loss	Penalty function	Resulting algorithm
Hinge Loss: $\sum_{i=1}^{m}[1 - y_i f(x_i)]_+$	$\lambda\|w\|^2$	SVM
Mean Squared Error $\sum_{i=1}^{m}(y_i - f(x_i))^2$	$\lambda\|w\|_2^2$	Ridge Regression
Mean Squared Error $\sum_{i=1}^{m}(y_i - f(x_i))^2$	$\lambda\|w\|_1$	Lasso
Mean Squared Error $\sum_{i=1}^{m}(y_i - f(x_i))^2$	$\lambda_1\|w\|_1 + \lambda_2\|w\|_2^2$	Elastic Net
Hinge Loss: $\sum_{i=1}^{m}[1 - y_i f(x_i)]_+$	$\lambda\|w\|_1$	1-Norm SVM

Algorithm	Loss	Penalty
SVM	convex, non-differentiable	convex, differentiable
Ridge Regression	convex, differentiable	convex, differentiable
Lasso	convex, differentiable	convex, non-differentiable
Elastic Net	convex, differentiable	convex, non-differentiable
Hinge Loss	convex, non-differentaible	convex, non-differentiable

Hard Margin SVM Regression: ϵ-SVR

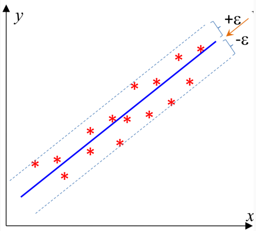


Hard Margin SVM Regression: ϵ-SVR



Goal: Find a linear function that fits the red data points.

Hard Margin SVM Regression: ϵ-SVR

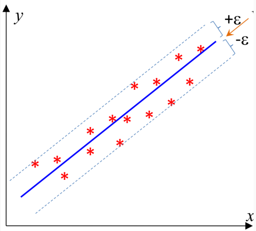


Goal: Find a linear function that fits the red data points.

Optimization Model:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\left\|w\right\|^2 \\ \text{subject to:} \quad & y_i(w \cdot x_i + b) \leq \epsilon, \\ & y_i(w \cdot x_i + b) \geq -\epsilon, \\ & i = 1, \dots, m. \end{aligned}$$

Hard Margin SVM Regression: ϵ-SVR



Goal: Find a linear function that fits the red data points.

Optimization Model:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\,\|w\|^2 \\ \text{subject to:} \quad & y_i(w \cdot x_i + b) \le \epsilon, \\ & y_i(w \cdot x_i + b) \ge -\epsilon, \\ & i = 1, \dots, m. \end{aligned}$$

Remark: Hence, the model suggests that the difference between y_i and fitted function should be smaller than ϵ (hyperparameter) and larger than $-\epsilon$. That is

$$|y_i - (w \cdot x_i + b)| \le \epsilon.$$

Here y_i is the height of the samples x_i, and not $+1$ or -1 labels!

Formlate Dual Margin Kernel SVM Problem for QP

Dual Soft-Margin Kernel SVM:

$$\begin{aligned}
\text{Maximize} \quad & \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j \, y_i y_j \, \Phi(x_i) \cdot \Phi(x_j) \\
\text{subject to} \quad & 0 \leq \lambda_i \leq C, \quad i = 1, \ldots, m, \\
& \sum_{i=1}^{m} \lambda_i y_i = 0.
\end{aligned}$$

Formlate Dual Margin Kernel SVM Problem for QP

Dual Soft-Margin Kernel SVM:

$$\begin{aligned} \text{Maximize} \quad & \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j \, y_i y_j \, \Phi(x_i) \cdot \Phi(x_j) \\ \text{subject to} \quad & 0 \le \lambda_i \le C, \quad i = 1, \dots, m, \\ & \sum_{i=1}^{m} \lambda_i y_i = 0. \end{aligned}$$

Recall QP:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} x^T P x + q^T x, \\ \text{subject to} \quad & Gx \le h, \\ & Ax = b \end{aligned}$$

How to set: x, P, q, G, h, A, b?

Formlate Dual Margin Kernel SVM Problem for QP

Dual Soft-Margin Kernel SVM:

$$\begin{aligned} \text{Maximize} \quad & \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j \, y_i y_j \, \Phi(x_i) \cdot \Phi(x_j) \\ \text{subject to} \quad & 0 \leq \lambda_i \leq C, \quad i = 1, \ldots, m, \\ & \sum_{i=1}^{m} \lambda_i y_i = 0. \end{aligned}$$

Recall QP:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \underline{x}^T P x + q^T x, \\ \text{subject to} \quad & Gx \leq h, \\ & Ax = b \end{aligned}$$

How to set: x, P, q, G, h, A, b?

1. Set $\underline{x} = [\lambda_1, \lambda_2, \ldots, \lambda_m]^T$

Formlate Dual Margin Kernel SVM Problem for QP

Dual Soft-Margin Kernel SVM:

$$\begin{aligned} \text{Maximize} \quad & \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j \, y_i y_j \, \underline{\Phi(x_i) \cdot \Phi(x_j)} \\ \text{subject to} \quad & 0 \leq \lambda_i \leq C, \quad i = 1, \ldots, m, \\ & \sum_{i=1}^{m} \lambda_i y_i = 0. \end{aligned}$$

Recall QP:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} x^T P x + q^T x, \\ \text{subject to} \quad & G x \leq h, \\ & A x = b \end{aligned}$$

How to set: x, P, q, G, h, A, b?

1. Set $x = [\lambda_1, \lambda_2, \ldots, \lambda_m]^T$
2. Let $\underline{K(x_i, x_j)} = \underline{\Phi(x_i) \cdot \Phi(x_j)}$

Formlate Dual Margin Kernel SVM Problem for QP

Dual Soft-Margin Kernel SVM:

$$\begin{aligned} \text{Maximize} \quad & \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j \, y_i y_j \, \underbrace{\Phi(x_i) \cdot \Phi(x_j)}_{K} \\ \text{subject to} \quad & 0 \le \lambda_i \le C, \quad i = 1, \ldots, m, \\ & \sum_{i=1}^{m} \lambda_i y_i = 0. \end{aligned}$$

Recall QP:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} x^T P x + q^T x, \\ \text{subject to} \quad & Gx \le h, \\ & Ax = b \end{aligned}$$

How to set: x, P, q, G, h, A, b?

$$[\lambda_1 \cdot \ \lambda_m] \, P \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_m \end{bmatrix}$$

1. Set $x = [\lambda_1, \lambda_2, \ldots, \lambda_m]^T$
2. Let $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$
3. Set matrix $P_{ij} = y_i y_j K(x_i, x_j)$

Formlate Dual Margin Kernel SVM Problem for QP

Dual Soft-Margin Kernel SVM:

$$\begin{aligned}
\text{Maximize} \quad & \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j \, y_i y_j \, \Phi(x_i) \cdot \Phi(x_j) \\
\text{subject to} \quad & 0 \le \lambda_i \le C, \quad i = 1, \ldots, m, \\
& \sum_{i=1}^{m} \lambda_i y_i = 0.
\end{aligned}$$

$$\min \boxed{-\sum \lambda_i} + \frac{1}{2} \sum \lambda_i \lambda_j y_i y_j \phi(x_i) \phi(x_j)$$

Recall QP:

$$\begin{aligned}
\underline{\text{minimize}} \quad & \frac{1}{2} x^T P x + \underline{q}^T x, \\
\text{subject to} \quad & Gx \le h, \\
& Ax = b
\end{aligned}$$

How to set: x, P, q, G, h, A, b?

1. Set $x = [\lambda_1, \lambda_2, \ldots, \lambda_m]^T$
2. Let $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$
3. Set matrix $P_{ij} = y_i y_j K(x_i, x_j)$
4. q is column vector containing $\underline{\text{-1}}$.

Formlate Dual Margin Kernel SVM Problem for QP

Dual Soft-Margin Kernel SVM:

$$\begin{aligned} \text{Maximize} \quad & \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j \, y_i y_j \, \Phi(x_i) \cdot \Phi(x_j) \\ \text{subject to} \quad & 0 \leq \lambda_i \leq C, \quad i = 1, \ldots, m, \\ & \sum_{i=1}^{m} \underset{x}{\boxed{\lambda_i}} y_i = \boxed{0}. \end{aligned}$$

Recall QP:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} x^T P x + q^T x, \\ \text{subject to} \quad & Gx \leq h, \\ & \underline{Ax} = \underline{b} \end{aligned}$$

How to set: x, P, q, G, h, A, b?

1. Set $x = [\lambda_1, \lambda_2, \ldots, \lambda_m]^T$
2. Let $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$
3. Set matrix $P_{ij} = y_i y_j K(x_i, x_j)$
4. q is column vector containing -1.
5. For equality constraint, set
   $A = \underline{[y_1, y_2, \ldots, y_m]}^T, \; b = \underline{0}$

Formlate Dual Margin Kernel SVM Problem for QP

Dual Soft-Margin Kernel SVM:

$$\text{Maximize} \quad \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j \, y_i y_j \, \Phi(x_i) \cdot \Phi(x_j)$$

$$\text{subject to} \quad \underline{0 \le \lambda_i \le C}, \quad i = 1, \ldots, m,$$

$$\sum_{i=1}^{m} \lambda_i y_i = 0.$$

$\boxed{-\lambda_i \le 0}$

$\lambda_i \geq 0$

$\boxed{\lambda_i \le C}$

Recall QP:

$$\text{minimize} \quad \frac{1}{2} x^T P x + q^T x,$$

$$\text{subject to} \quad Gx \le h,$$

$$Ax = b$$

How to set: x, P, q, G, h, A, b?

1. Set $x = [\lambda_1, \lambda_2, \ldots, \lambda_m]^T$
2. Let $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$
3. Set matrix $P_{ij} = y_i y_j K(x_i, x_j)$
4. q is column vector containing -1.
5. For equality constraint, set $A = [y_1, y_2, \ldots, y_m]^T$, $b = 0$
6. Inequalities: $\underline{-\lambda_i \le 0}$, $\underline{\lambda_i \le C}$

Formlate Dual Margin Kernel SVM Problem for QP

Dual Soft-Margin Kernel SVM:

$$\begin{aligned} \text{Maximize} \quad & \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j \, y_i y_j \, \Phi(x_i) \cdot \Phi(x_j) \\ \text{subject to} \quad & 0 \leq \lambda_i \leq C, \quad i = 1, \ldots, m, \\ & \sum_{i=1}^{m} \lambda_i y_i = 0. \end{aligned}$$

Recall QP:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} x^T P x + q^T x, \\ \text{subject to} \quad & Gx \leq h, \\ & Ax = b \end{aligned}$$

How to set: x, P, q, G, h, A, b?

1. Set $x = [\lambda_1, \lambda_2, \ldots, \lambda_m]^T$
2. Let $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$
3. Set matrix $P_{ij} = y_i y_j K(x_i, x_j)$
4. q is column vector containing -1.
5. For equality constraint, set
   $A = [y_1, y_2, \ldots, y_m]^T, \; b = 0$
6. Inequalities: $-\lambda_i \leq 0, \, \lambda_i \leq C$
7. Set G as

Formlate Dual Margin Kernel SVM Problem for QP

Dual Soft-Margin Kernel SVM:

$$\text{Maximize} \quad \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j \, y_i y_j \, \Phi(x_i) \cdot \Phi(x_j)$$

$$\text{subject to} \quad 0 \leq \lambda_i \leq C, \quad i = 1, \ldots, m,$$

$$\sum_{i=1}^{m} \lambda_i y_i = 0.$$

Recall QP:

$$\text{minimize} \quad \frac{1}{2} x^T P x + q^T x,$$

$$\text{subject to} \quad Gx \leq h,$$

$$Ax = b$$

How to set: x, P, q, G, h, A, b?

1. Set $x = [\lambda_1, \lambda_2, \ldots, \lambda_m]^T$
2. Let $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$
3. Set matrix $P_{ij} = y_i y_j K(x_i, x_j)$
4. q is column vector containing -1.
5. For equality constraint, set $A = [y_1, y_2, \ldots, y_m]^T$, $b = 0$
6. Inequalities: $-\lambda_i \leq 0$, $\lambda_i \leq C$

$$h = \begin{bmatrix} C \\ C \\ C \\ \vdots \\ C \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

7. Set G as

$$G = \left[\begin{array}{cccc} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \ddots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \\ \hline -1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ldots & \vdots \\ 0 & 0 & \ldots & -1 \end{array}\right] \quad \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_m \end{bmatrix}$$

x

$-\lambda_1 \leq 0$

Formlate Dual Margin Kernel SVM Problem for QP

Dual Soft-Margin Kernel SVM:

$$\text{Maximize} \quad \sum_{i=1}^{m} \lambda_i - \frac{1}{2}\sum_{i,j=1}^{m} \lambda_i\lambda_j\, y_i y_j\, \Phi(x_i)\cdot\Phi(x_j)$$

$$\text{subject to} \quad 0 \le \lambda_i \le C, \quad i = 1,\dots,m,$$

$$\sum_{i=1}^{m} \lambda_i y_i = 0.$$

Recall QP:

$$\text{minimize} \quad \frac{1}{2}x^T P x + q^T x,$$

$$\text{subject to} \quad Gx \le h,$$

$$Ax = b$$

How to set: x, P, q, G, h, A, b?

1. Set $x = [\lambda_1, \lambda_2, \dots, \lambda_m]^T$
2. Let $K(x_i, x_j) = \Phi(x_i)\cdot\Phi(x_j)$
3. Set matrix $P_{ij} = y_i y_j K(x_i, x_j)$
4. q is column vector containing -1.
5. For equality constraint, set
   $A = [y_1, y_2, \dots, y_m]^T$, $b = 0$
6. Inequalities: $-\lambda_i \le 0$, $\lambda_i \le C$

$\lambda_1 \le C$

$\lambda_2 \le C$

$\lambda_m \le C$

$$I\begin{bmatrix}\lambda_1\\ \lambda_2\\ \lambda_m\end{bmatrix} \le \begin{bmatrix}C\\ C\\ C\end{bmatrix}$$

7. Set G as

$$G = \left[\begin{array}{cccc} 1 & 0 & \dots & 0\\ 0 & 1 & \dots & 0\\ \vdots & \ddots & \dots & \vdots\\ 0 & 0 & \dots & 1\\ \hline -1 & 0 & \dots & 0\\ \vdots & \ddots & \dots & \vdots\\ 0 & 0 & \dots & -1 \end{array}\right]$$

I

$-I$

8. $h = [C, C, \dots, C \,|\, 0, \dots, 0]^T$

Algorithms for Dual Soft Margin Kernel SVM

Algorithm 1 Algorithm for solving Dual Kernel SVM using QP

```
Initialization:
    • Compute $K = XX^T$, if possible using input space. ✓
    • For linear kernel, return $K$, for polynomial of degree $d$, return $\frac{1}{d}K^d$. ✓
    • For RBF kernel, compute $K = \exp(-(x - x')^2/2\sigma^2)$. ✓
Training: Assemble matrices and vectors to solve QP for dual
        $\min_x \quad (x^T P x + q^T x), \quad \text{subject to } Gx \leq h, \; Ax = b.$
    • Define $x, P, q, G, x, h, b$ as described in previous slide.
```

Pass to some solver.