

SVM & Kernels

Xuan Zhang
Sep. 23, 2022

HW1. Q1.

$\hat{f}(x^{(0)})$ is relevant to T

$f(x^{(0)})$ is irrelevant to T

$$y = f(x) + \epsilon$$

$$\mathbb{E}[\epsilon] = 0$$

$$\text{Var}[\epsilon] = \sigma_\epsilon^2$$

$$\mathbb{E}[x^2] = \sigma^2 + \mu^2$$

$$\mathbb{E}_{y^{(0)}|x^{(0)}, T} [(y^{(0)} - \hat{f}(x^{(0)}))^2]$$

$$= \mathbb{E}_{y^{(0)}|x^{(0)}, T} [(y^{(0)} - f(x^{(0)}) + f(x^{(0)}) - \hat{f}(x^{(0)}))^2]$$

$$= \underbrace{\mathbb{E}_{y^{(0)}|x^{(0)}, T} [(y^{(0)} - f(x^{(0)}))^2]}_{\textcircled{1}} + \underbrace{\mathbb{E}_{y^{(0)}|x^{(0)}, T} [f(x^{(0)}) - \hat{f}(x^{(0)})]^2]}_{\textcircled{2}} + 2 \mathbb{E}_{y^{(0)}|x^{(0)}, T} [(y^{(0)} - f(x^{(0)})) (f(x^{(0)}) - \hat{f}(x^{(0)}))]_{\textcircled{3}}$$

$$\textcircled{1} = \mathbb{E}_{y^{(0)}|x^{(0)}, T} [\epsilon^2] = \sigma_\epsilon^2 + (\mathbb{E}[\epsilon])^2 = \sigma_\epsilon^2$$

$$\textcircled{3} = \mathbb{E}_T [(f(x^{(0)}) - \hat{f}(x^{(0)})) \mathbb{E}_{y^{(0)}|x^{(0)}} [y^{(0)} - f(x^{(0)})]]$$

$$= \mathbb{E}_T [(f(x^{(0)}) - \hat{f}(x^{(0)})) \mathbb{E}_{y^{(0)}|x^{(0)}} [\epsilon]] = 0$$

$$\textcircled{2} = \mathbb{E}_{y^{(0)}|x^{(0)}, T} [(\hat{f}(x^{(0)}) - \mathbb{E}[\hat{f}(x^{(0)})] + \mathbb{E}[\hat{f}(x^{(0)})] - f(x^{(0)}))^2]$$

$$= \underbrace{\mathbb{E}_{y^{(0)}|x^{(0)}, T} [(\hat{f}(x^{(0)}) - \mathbb{E}[\hat{f}(x^{(0)})])^2]}_{\textcircled{1}} + \underbrace{\mathbb{E}_{y^{(0)}|x^{(0)}, T} [(\mathbb{E}[\hat{f}(x^{(0)})] - f(x^{(0)}))^2]}_{\textcircled{2}} + 2 \mathbb{E}_{y^{(0)}|x^{(0)}, T} [(\hat{f}(x^{(0)}) - \mathbb{E}[\hat{f}(x^{(0)})]) (\mathbb{E}[\hat{f}(x^{(0)})] - f(x^{(0)}))]_{\textcircled{3}}$$

$$\textcircled{1} = \mathbb{E}_T [(\hat{f}(x^{(0)}) - \mathbb{E}[\hat{f}(x^{(0)})])^2] = \text{Variance}$$

$$\textcircled{2} = (\mathbb{E}[\hat{f}(x^{(0)})] - f(x^{(0)}))^2 = \text{bias}^2$$

$$\textcircled{3} = (\mathbb{E}[\hat{f}(x^{(0)})] - f(x^{(0)})) \mathbb{E}_T [\hat{f}(x^{(0)}) - \mathbb{E}[\hat{f}(x^{(0)})]] = (\mathbb{E}[\hat{f}(x^{(0)})] - f(x^{(0)})) (\mathbb{E}[\hat{f}(x^{(0)})] - \mathbb{E}[\hat{f}(x^{(0)})]) = 0$$

Bias - Variance Decomposition

expected squared loss:

$$\mathbb{E}[L] = \mathbb{E}[(f(x) - y)^2]$$

$$\begin{aligned} &\text{error from} \quad \text{model} \quad \text{data} \\ &= \mathbb{E}[(f(x) - h(x))^2] + \mathbb{E}[(h(x) - y)^2] \\ &\quad \text{depends on the choice for } f(x). \\ &\quad \text{we want to find a } f(x) \\ &\quad \text{to minimize this term.} \quad \text{arises from the intrinsic} \\ &\quad \text{noise on the data} \end{aligned}$$

first item depends on dataset D :

$$\begin{aligned} \mathbb{E}_D[(f(x; D) - h(x))^2] &= (\mathbb{E}_D[f(x; D)] - h(x))^2 + \mathbb{E}_D[(f(x; D) - \mathbb{E}_D[f(x; D)])^2] \\ &\quad \text{model risk} \quad \quad \quad \text{(bias)}^2 \quad \quad \quad \text{variance} \end{aligned}$$

$$\text{expected squared loss} = \underbrace{(\text{bias})^2}_{\text{model risk}} + \text{variance} + \text{noise}$$

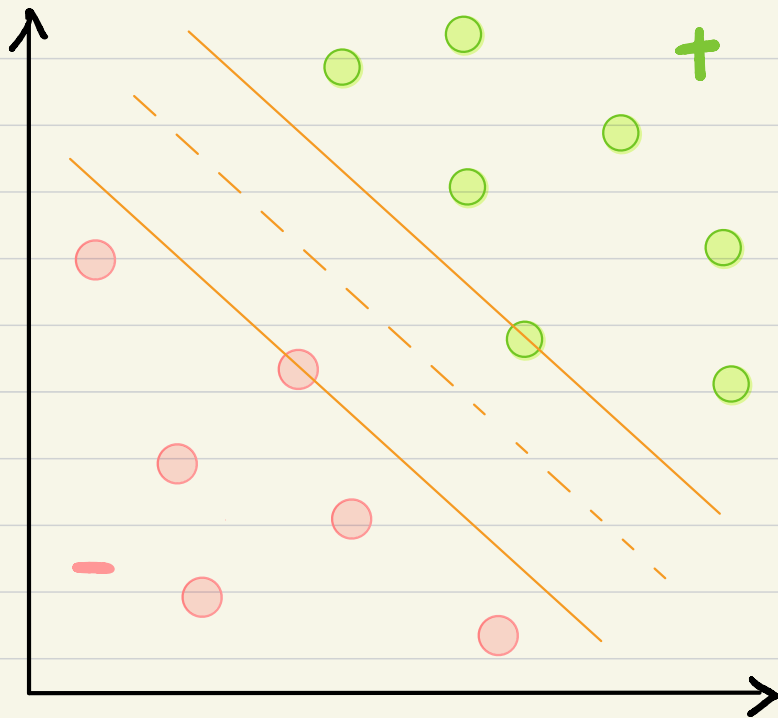
I. SVM

- Geometric representation 1
- Geometric representation 2 *
- dual formation *
- Soft margin SVM

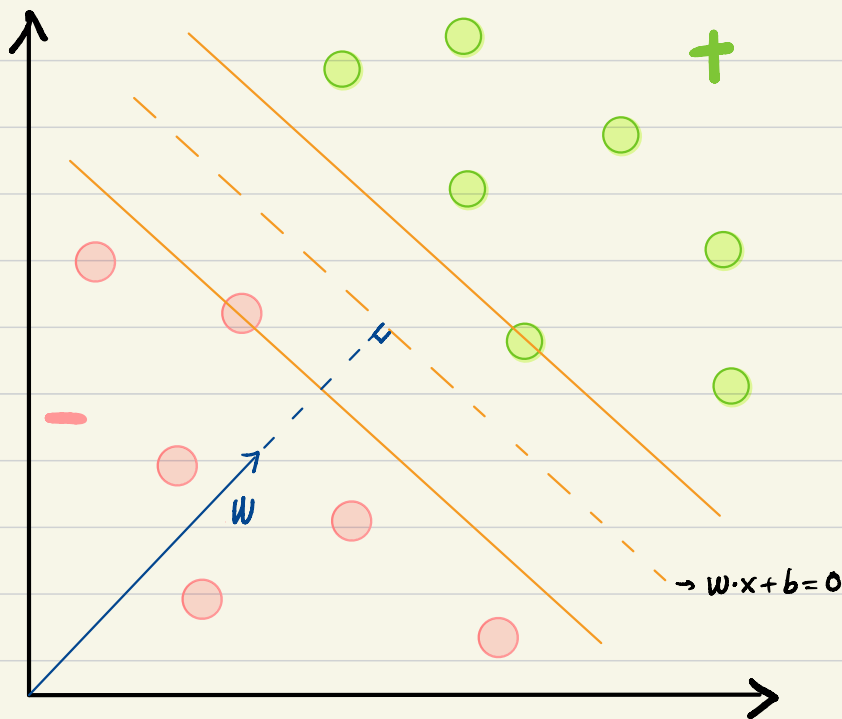
* Slides are adapted from MIT Artificial Intelligence, Fall 2010, Lesson 16 by Patrick Winston.
Lecture recording is available on YouTube.



Geometric Representation 1

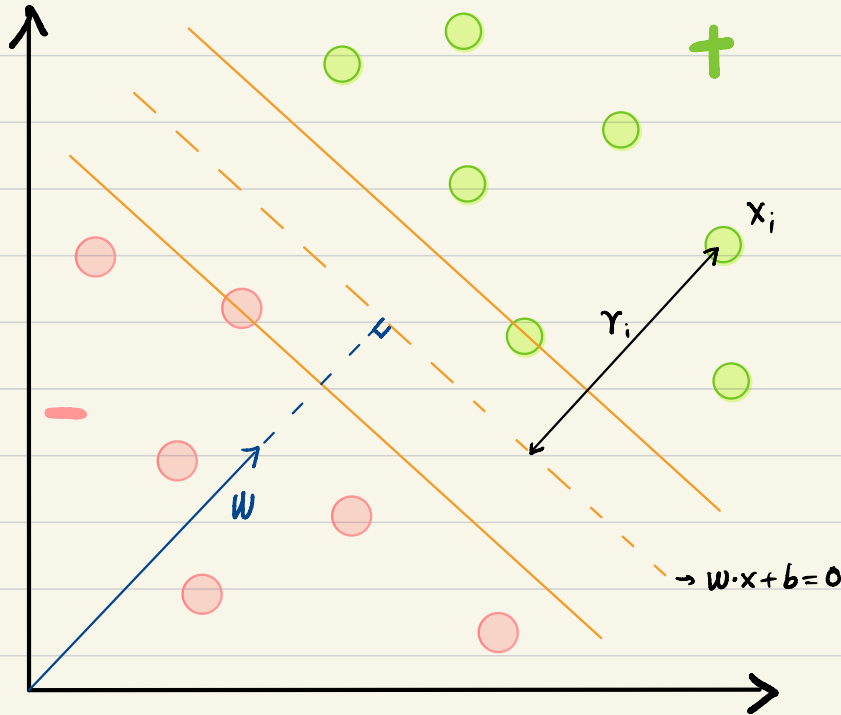


Geometric Representation I



Let's define a vector $\vec{w}$ that is perpendicular to the decision boundary $\vec{w} \cdot \vec{x} + b = 0$.

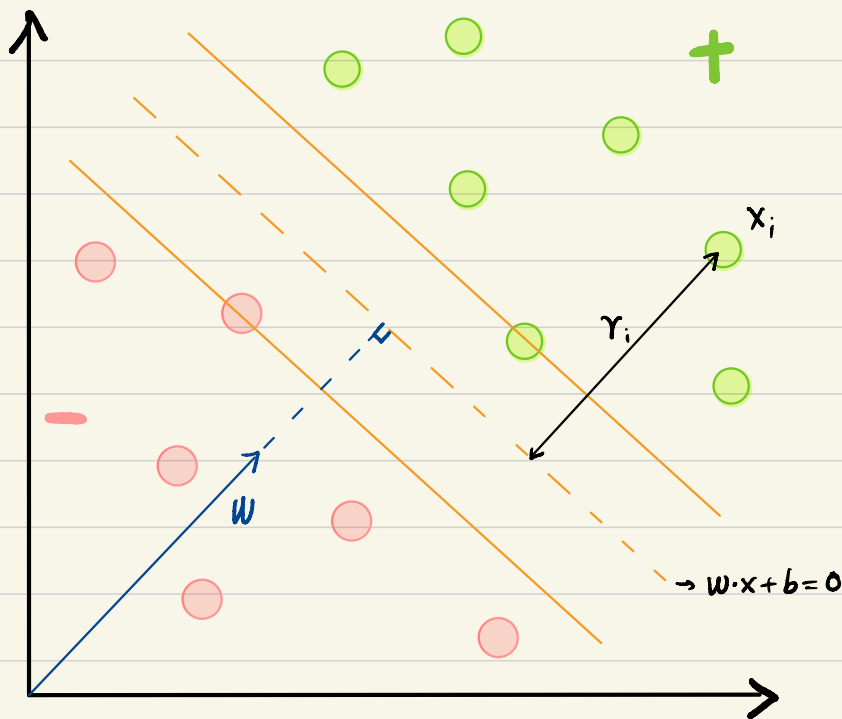
Geometric Representation I



Let's define a vector $\vec{w}$ that is perpendicular to the decision boundary $\vec{w} \cdot \vec{x} + b = 0$.

For a random point x_i , its distance to the decision boundary $\gamma_i = \frac{|\vec{w} \cdot \vec{x}_i + b|}{\|\vec{w}\|}$.

Geometric Representation I



For a random point x_i ,
it's distance to the decision boundary
$$\gamma_i = \frac{|\vec{w} \cdot \vec{x}_i + b|}{\|\vec{w}\|}$$

We introduce y :

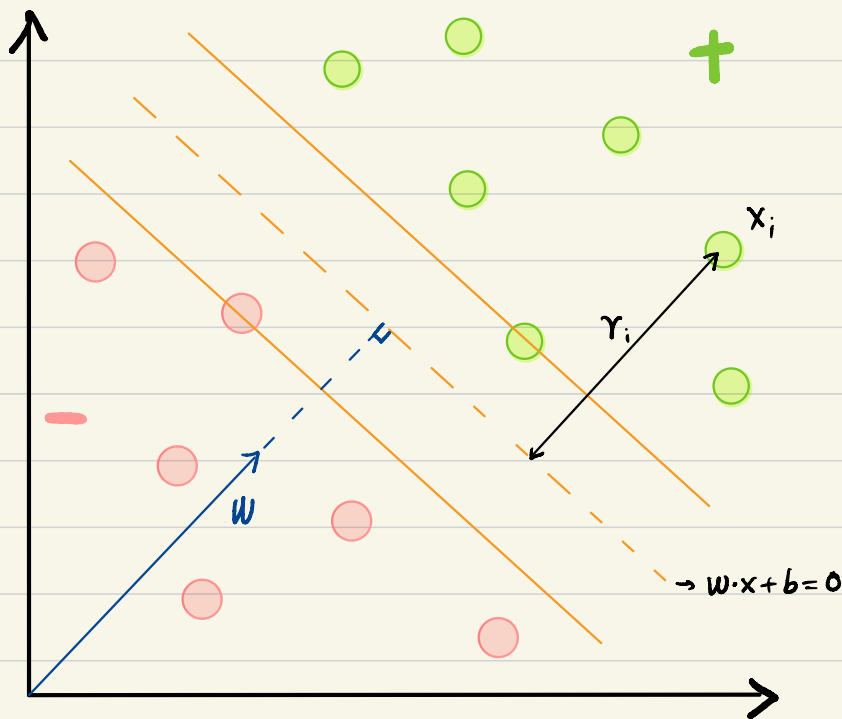
$y = +1$ for $+$ samples

$y = -1$ for $-$ samples

$$\gamma_i = y_i \left(\frac{\vec{w} \cdot \vec{x}_i + b}{\|\vec{w}\|} \right)$$

geometric margin

Geometric Representation I



We introduce y :

$y = +1$ for $+$ samples

$y = -1$ for $-$ samples

$$r_i = y_i \left(\frac{\vec{w} \cdot \vec{x}_i + b}{\|\vec{w}\|} \right)$$

geometric margin

$$\hat{\gamma}_i = \underbrace{y_i}_{\text{label}} \underbrace{(\vec{w} \cdot \vec{x}_i + b)}_{\text{prediction}}$$

functional margin

Maximum Geometric Margin

$$r_i = y_i \left(\frac{\vec{w} \cdot \vec{x}_i + b}{\|\vec{w}\|} \right)$$

geometric margin

Let's say each sample has distance at least r ,
i.e. we constrain $r_i \geq r \forall i$.

And we want to maximize r , i.e. we want points closest to the decision boundary to have large margin.

Objective:

$$r = \frac{\hat{r}}{\|\vec{w}\|}$$

$$\begin{aligned} \max \quad & r \\ \text{s.t.} \quad & \frac{y_i (\vec{w} \cdot \vec{x}_i + b)}{\|\vec{w}\|} \geq r \end{aligned}$$

$$\hat{r}_i = \underbrace{y_i}_{\text{label}} \underbrace{(\vec{w} \cdot \vec{x}_i + b)}_{\text{prediction}}$$

functional margin

Objective:

$$\begin{aligned} \max \quad & \frac{\hat{r}}{\|\vec{w}\|} \\ \text{s.t.} \quad & y_i (\vec{w} \cdot \vec{x}_i + b) \geq \hat{r} \end{aligned}$$

Maximum Geometric Margin

$$r_i = y_i \left(\frac{\vec{w} \cdot \vec{x}_i + b}{\|\vec{w}\|} \right)$$

geometric margin

Objective:

$$\begin{aligned} \max \quad & r \\ \text{s.t.} \quad & y_i (\vec{w} \cdot \vec{x}_i + b) \geq r \|\vec{w}\| \end{aligned}$$

$$\hat{r}_i = \underbrace{y_i}_{\text{label}} \underbrace{(\vec{w} \cdot \vec{x}_i + b)}_{\text{prediction}}$$

functional margin

Objective:

$$\begin{aligned} \max \quad & \frac{\hat{r}}{\|\vec{w}\|} \\ \text{s.t.} \quad & y_i (\vec{w} \cdot \vec{x}_i + b) \geq \hat{r} \end{aligned}$$

Notice that the value of $\hat{r}$ doesn't matter: $\vec{w}' = \frac{\vec{w}}{\hat{r}}$. We can set r to 1.

$$\max \quad \frac{1}{\|\vec{w}\|}$$

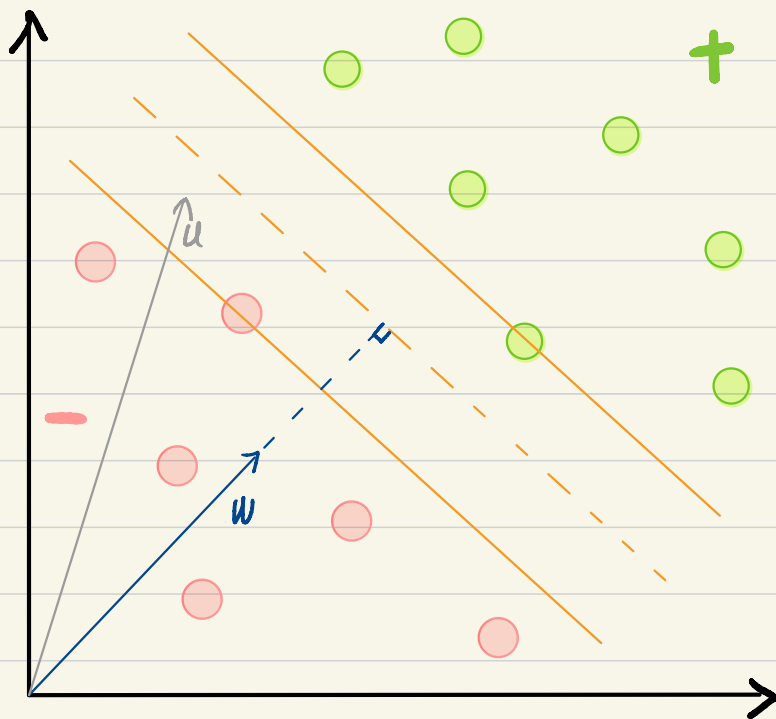
$$\text{s.t.} \quad y_i (\vec{w} \cdot \vec{x}_i + b) \geq 1$$

$\Leftrightarrow$

$$\min \quad \frac{1}{2} \|\vec{w}\|^2$$

$$\text{s.t.} \quad y_i (\vec{w} \cdot \vec{x}_i + b) \geq 1$$

Geometric Representation 2



Let's define a vector $\vec{w}$ that is perpendicular to the decision boundary.

We want to know which side (-/+) an unknown vector $\vec{u}$ is on.

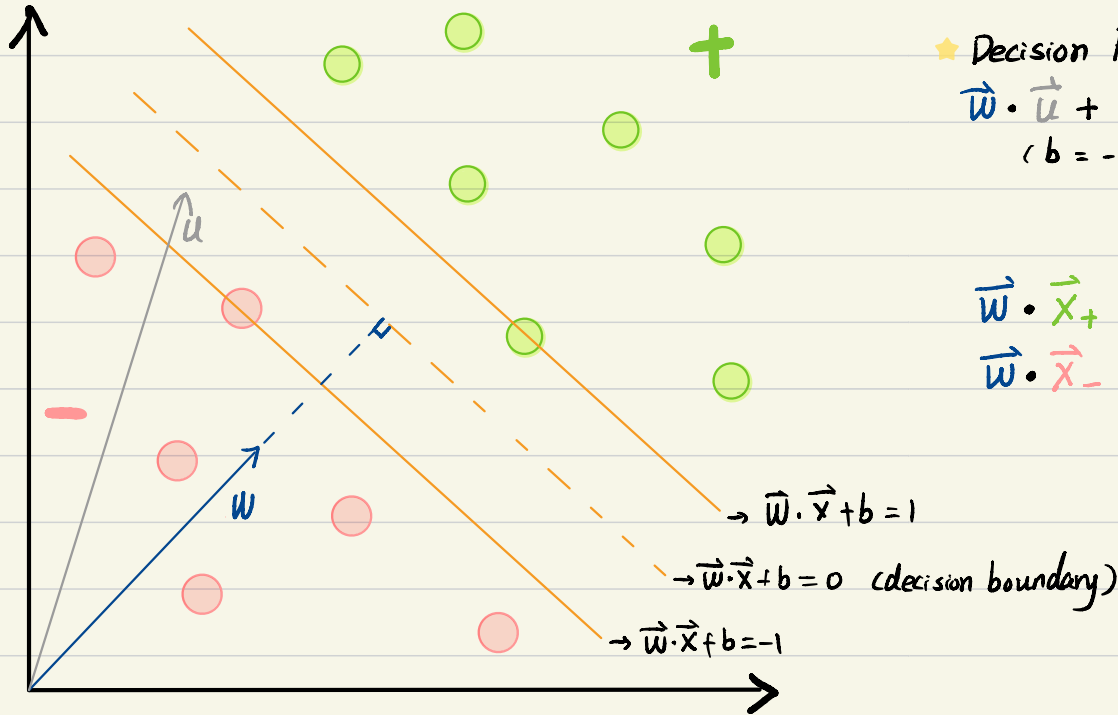
Project $\vec{u}$ to $\vec{w}$:

$\vec{w} \cdot \vec{u} \geq c$ then +

★ Decision Rule :

$\vec{w} \cdot \vec{u} + b \geq 0$ then +
($b = -c$)

Geometric Representation 2



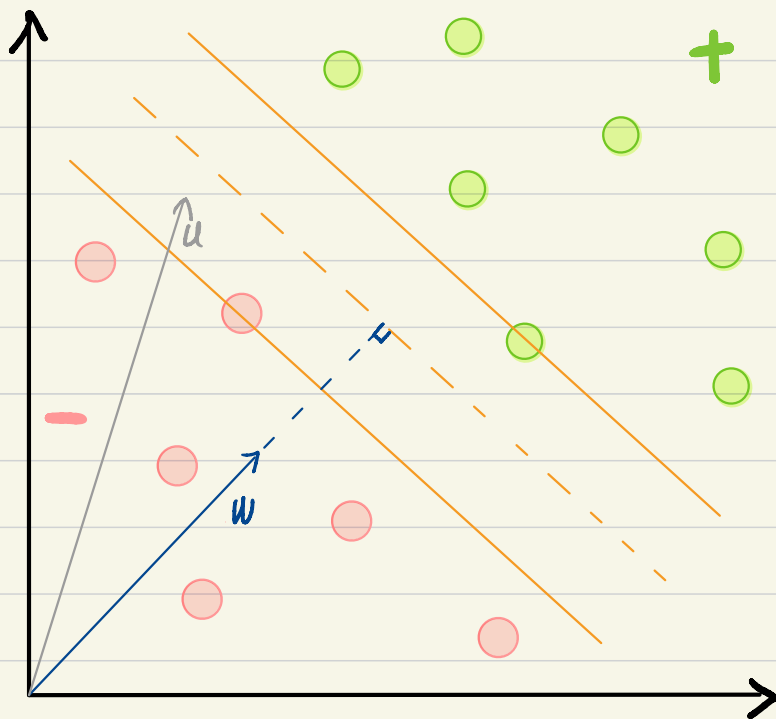
★ Decision Rule :

$$\vec{w} \cdot \vec{u} + b \geq 0 \text{ then } + \\ (b = -c)$$

$$\begin{aligned} \vec{w} \cdot \vec{x}_+ + b &\geq 1 \\ \vec{w} \cdot \vec{x}_- + b &\leq -1 \end{aligned}$$

Arbitrary numbers
we choose

Geometric Representation 2



★ Decision Rule :

$$\vec{w} \cdot \vec{u} + b \geq 0 \text{ then } +$$

($b = -c$)

$$\begin{aligned} \vec{w} \cdot \vec{x}_+ + b &\geq 1 \\ \vec{w} \cdot \vec{x}_- + b &\leq -1 \end{aligned}$$

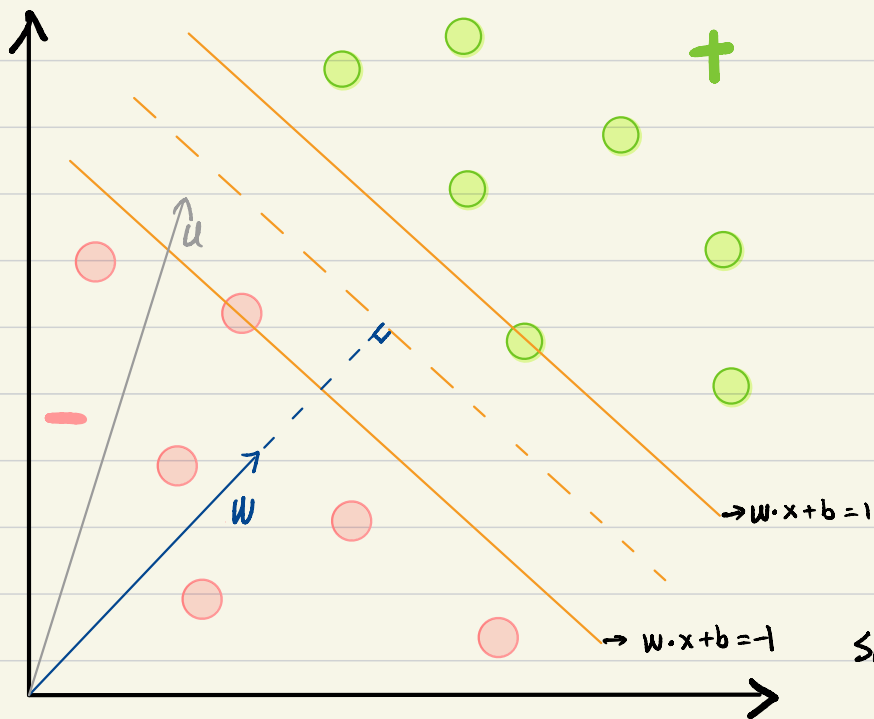
Arbitrary numbers
we choose

We introduce y :

$y = +1$ for $+$ samples

$y = -1$ for $-$ samples

Geometric Representation 2



$$\vec{w} \cdot \vec{x}_+ + b \geq 1$$

$$\vec{w} \cdot \vec{x}_- + b \leq -1$$

We introduce y :

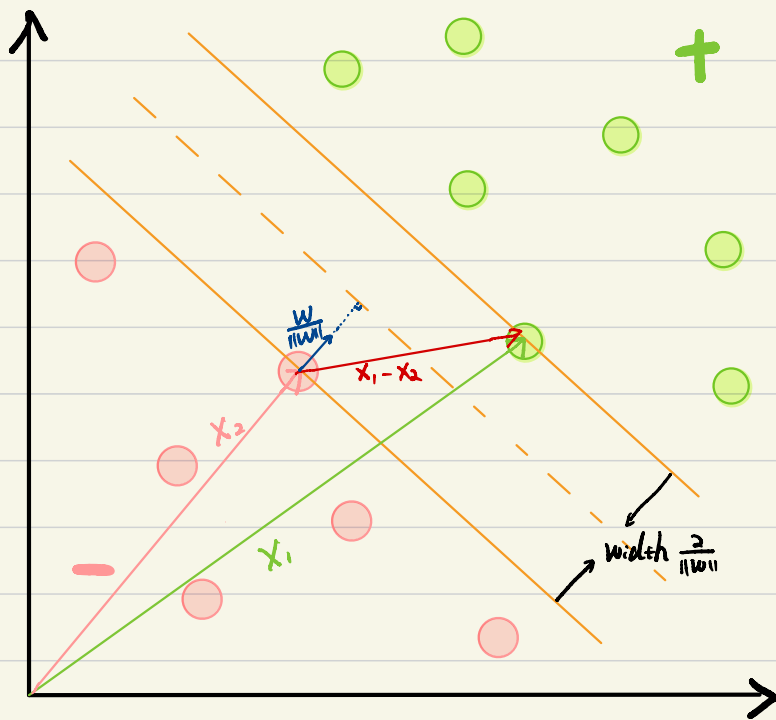
$y = +1$ for $+$ samples

$y = -1$ for $-$ samples

We want $y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 \quad \forall i$

support vectors $\vec{x} : y(\vec{w} \cdot \vec{x} + b) = 1$

Geometric Representation 2



$$y(\vec{w} \cdot \vec{x} + b) = 1 \quad \text{for } \vec{x}_1, \vec{x}_2$$

$$\text{width} = (\vec{x}_1 - \vec{x}_2) \cdot \frac{\vec{w}}{\|\vec{w}\|}$$

$$= \frac{1}{\|\vec{w}\|} (\vec{x}_1 \cdot \vec{w} - \vec{x}_2 \cdot \vec{w})$$

$$= \frac{1}{\|\vec{w}\|} (1 - b - (-1 - b))$$

$$= \frac{2}{\|\vec{w}\|} \quad \text{distance between two parallel lines}$$

$$\begin{aligned} \text{goal: } \max \frac{2}{\|\vec{w}\|} &\Leftrightarrow \min \|\vec{w}\| \\ &\Leftrightarrow \min \frac{1}{2} \|\vec{w}\|^2 \end{aligned}$$

$$\text{constraint: } y(\vec{w} \cdot \vec{x} + b) \geq 1$$

Geometric Representation

Geometric Representation 1:

We maximize the distance / margin of the samples that are the closest to the decision boundary.

Geometric Representation 2:

We maximize the width of the "street".

support vectors: samples lie on the boundary of the "street".

Dual Formation

primal problem:

$$\begin{aligned} \min_{\vec{w}, b} \quad & \frac{1}{2} \|\vec{w}\|^2 \\ \text{s.t.} \quad & y(\vec{w} \cdot \vec{x} + b) - 1 \geq 0 \end{aligned}$$

dual problem: (Lagrange function)

$$L(\vec{w}, b, \alpha) = \frac{1}{2} \|\vec{w}\|^2 - \sum_i^N \alpha_i [y_i(\vec{w} \cdot \vec{x}_i + b) - 1]$$

objective:

$$\max_{\alpha} \min_{\vec{w}, b} L(\vec{w}, b, \alpha) \quad \text{s.t.} \quad \alpha_i \geq 0 \quad \forall i$$

$$\frac{\partial L}{\partial \vec{w}} = \vec{w} - \sum_i^N \alpha_i y_i \vec{x}_i = 0 \quad \Rightarrow \quad \vec{w} = \sum_i^N \alpha_i y_i \vec{x}_i$$

$$\frac{\partial L}{\partial b} = -\sum_i^N \alpha_i y_i = 0 \quad \Rightarrow \quad \sum_i^N \alpha_i y_i = 0$$

* $\frac{\partial \|\vec{w}\|^2}{\partial \vec{w}} = 2\vec{w}$ << The Matrix Cookbook >> page 14

Dual Formation

dual problem:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \|\bar{\mathbf{w}}\|^2 - \sum_i \alpha_i [y_i (\bar{\mathbf{w}} \cdot \bar{\mathbf{x}}_i + b) - 1]$$

objective:

$$\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha})$$

$$\text{s.t. } \alpha_i \geq 0$$

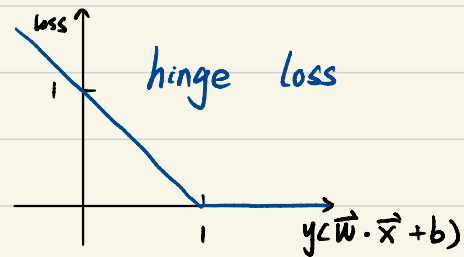
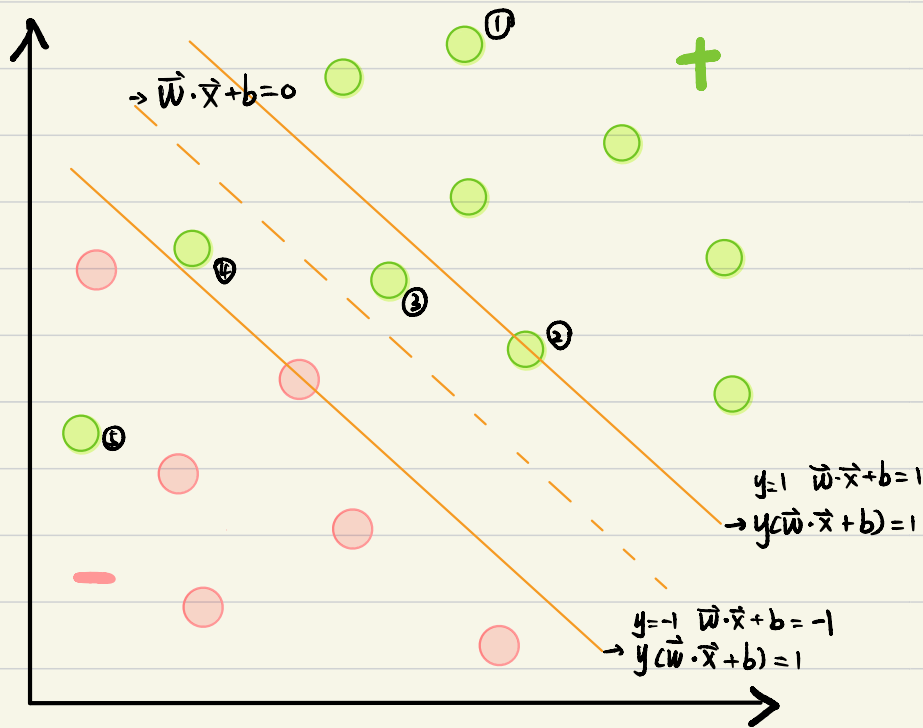
$$\sum_i \alpha_i y_i = 0$$

$$\begin{cases} \bar{\mathbf{w}} = \sum_i \alpha_i y_i \bar{\mathbf{x}}_i \\ \sum_i \alpha_i y_i = 0 \end{cases}$$

notice: for non-support vectors, $\alpha_i = 0$

$$\begin{aligned} L(\boldsymbol{\alpha}) &= \frac{1}{2} \left(\sum_i \alpha_i y_i \bar{\mathbf{x}}_i \right) \cdot \left(\sum_i \alpha_i y_i \bar{\mathbf{x}}_i \right) - \sum_i \alpha_i y_i \bar{\mathbf{x}}_i \cdot \left(\sum_j \alpha_j y_j \bar{\mathbf{x}}_j \right) - \underbrace{\sum_i \alpha_i y_i b}_{=0} + \sum_i \alpha_i \\ &= \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \bar{\mathbf{x}}_i \cdot \bar{\mathbf{x}}_j \end{aligned}$$

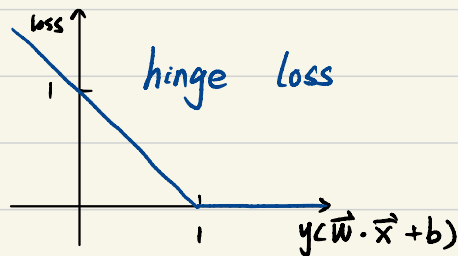
Soft margin SVM



$$\text{hinge loss} = \max(0, 1 - y(\vec{w} \cdot \vec{x} + b))$$

- ① : $\max(0, 1 - 1 \times \text{greater than } 1) = 0$
- ② : $\max(0, 1 - 1 \times 1) = 0$
- ③ : $\max(0, 1 - 1 \times \text{between } 0 \& 1) = \text{between } 0 \& 1$
- ④ : $\max(0, 1 - 1 \times \text{between } -1 \& 0) = \text{between } 1 \& 2$
- ⑤ : $\max(0, 1 - 1 \times \text{smaller than } -1) = \text{greater than } 2$

Soft margin SVM



$$\text{hinge loss} = \max(0, 1 - y(\vec{w} \cdot \vec{x} + b))$$

minimize empirical loss with regularization.

$$\min \frac{1}{N} \sum_{i=1}^N (1 - y_i(\vec{w} \cdot \vec{x}_i + b))^+ + \lambda \frac{1}{2} \|\vec{w}\|^2$$

slack variables:

$$\xi_i \geq 0$$

soft margin constraints:

$$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 - \xi_i$$

[Recall hard margin constraints:]

$$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1$$

Objective:

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

$$\text{s.t. } y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 - \xi_i \quad \forall i$$

$$\xi_i \geq 0 \quad \forall i$$

II. Kernels

- kernel tricks
- valid kernels
- RBF kernel

kernel tricks

prediction for a new example $\vec{x}$ given SVM solution $\vec{w}^*, b^*, \alpha^*$:

$$\begin{aligned} y &= \vec{w}^* \cdot \vec{x} + b^* \\ &= \left(\sum_i^N \alpha_i^* y_i \vec{x}_i \right) \cdot \vec{x} + b^* \\ &= \sum_i^N \alpha_i^* y_i \boxed{\vec{x}_i \cdot \vec{x}} + b^* \end{aligned}$$

kernel tricks

prediction for a new example $\vec{x}$

given SVM solution $\vec{w}^*, b^*, \alpha^*$

$$y = \vec{w}^* \cdot \vec{x} + b^*$$

$$= \left(\sum_i \alpha_i^* y_i \vec{x}_i \right) \cdot \vec{x} + b^*$$

$$= \sum_i \alpha_i^* y_i \boxed{\vec{x}_i \cdot \vec{x}} + b^*$$

$$k(x_i, x)$$

We only need to compute the inner product between support vectors and $\vec{x}$.

KKT Conditions *

$$\sum_i \alpha_i y_i \vec{x}_i = \vec{w}$$

$$\sum_i \alpha_i y_i = 0$$

$$\alpha_i \geq 0$$

$$y_i (\vec{w} \cdot \vec{x}_i + b) - 1 \geq 0$$

$$\alpha_i [y_i (\vec{w} \cdot \vec{x}_i + b) - 1] = 0$$



α_i^* is zero for all non-support vectors.

* Under KKT conditions, the solution for primal and dual is the same.

kernel tricks

$$K(x, y) = \Phi(x) \cdot \Phi(y)$$

Φ : feature space mapping

$\Phi(x)$: high dimensional feature vector ; example: $x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$, $\Phi(x) = \begin{pmatrix} x_1^2 \\ x_1 x_2 \\ x_1^2 \\ x_2^3 \\ x_3 \\ x_3^2 \end{pmatrix}$

A kernel is an inner product between two vectors.

It measures the similarity between two vectors.

Valid kernels

- It can be written as an inner product of two vectors.

- It satisfies Mercer's theorem.

For Gram matrix $K_{ij} = K(\vec{x}_i, \vec{x}_j)$, $\mathbf{z}^T \mathbf{K} \mathbf{z} \geq 0$ i.e. $\sum_i \sum_j z_i z_j K(\vec{x}_i, \vec{x}_j) \geq 0$

- It can be constructed from other kernels.

$$K(x, x') = cK_1(x, x')$$

$$K(x, x') = f(x)K_1(x, x')f(x')$$

$$K(x, x') = \exp(K_1(x, x'))$$

$$K(x, x') = K_1(x, x') + K_2(x, x')$$

$$K(x, x') = K_1(x, x')K_2(x, x')$$

Example 1

Show $k(\vec{x}, \vec{y}) = (1 + \vec{x} \cdot \vec{y})^2$ is a valid kernel.

- We need to rewrite $k(\vec{x}, \vec{y})$ as $\Phi(\vec{x}) \cdot \Phi(\vec{y})$.

$$\text{Suppose } \vec{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \vec{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

$$k(\vec{x}, \vec{y}) = (1 + \vec{x} \cdot \vec{y})^2$$

$$= (1 + x_1 y_1 + x_2 y_2)^2$$

$$= 1 + x_1^2 y_1^2 + x_2^2 y_2^2 + 2x_1 y_1 + 2x_2 y_2 + 2x_1 y_1 x_2 y_2$$

$$= (1, x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2) \cdot (1, y_1^2, y_2^2, \sqrt{2}y_1, \sqrt{2}y_2, \sqrt{2}y_1 y_2)$$

$$= \Phi(\vec{x}) \cdot \Phi(\vec{y})$$

Example 2

If $k_1(\vec{x}, \vec{x}')$ is valid kernel, show $k_c(\vec{x}, \vec{x}') = c k_1(\vec{x}, \vec{x}')$, $c \geq 0$ is also a valid kernel.

The Gram matrix: $\mathbf{K} = c \mathbf{K}_1$,

which implies $\forall \mathbf{z} \in \mathbb{R}^n$, $\mathbf{z}^T \mathbf{K} \mathbf{z} = c \mathbf{z}^T \mathbf{K}_1 \mathbf{z} \geq 0$.

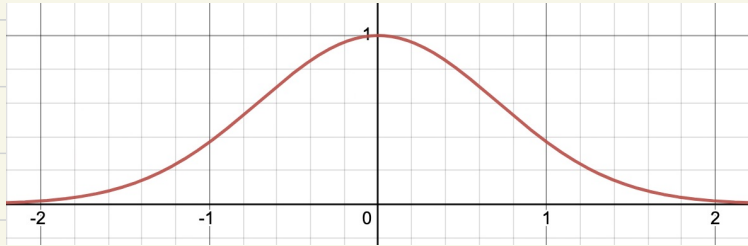
RBF Kernel

$$k(\vec{x}, \vec{x}') = \exp(-\gamma \|\vec{x} - \vec{x}'\|^2)$$

γ sets the width of the bell shaped curve.

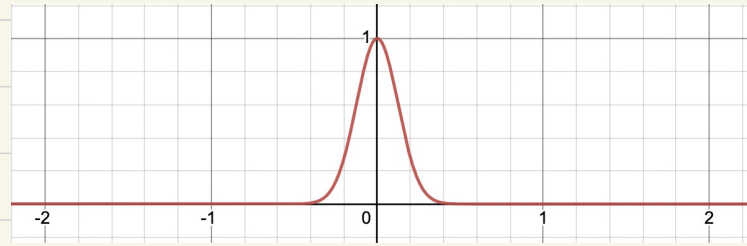
Which value
of γ tends to
overfit?

$\gamma = 1$



region of similarity

$\gamma = 32$



region of
dissimilarity

region of
similarity

region of
dissimilarity



Questions?