香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

COMP 5212
Machine Learning
Lecture 5

Support Vector Machine

Junxian He
Sep 23, 2024

Recap: Kernel Trick

$$\theta = \sum_{i=1}^{n} \beta_i \phi(x^{(i)}) \qquad \beta_i \in R$$

feature map

Recap: Kernel Trick

$$\theta = \sum_{i=1}^{n} \beta_i \phi(x^{(i)}) \qquad \beta_i \in R$$

$$\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j \phi(x^{(j)})^T \phi(x^{(i)}) \right)$$

Recap: Kernel Trick

$$\theta = \sum_{i=1}^{n} \beta_i \phi(x^{(i)}) \qquad \beta_i \in R$$

$$\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j \phi(x^{(j)})^T \phi(x^{(i)}) \right)$$

Kernel $K(x, z)$ $\quad \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ $\quad \mathcal{X}$ is the space of the input

$$K(x, z) \triangleq \langle \phi(x), \phi(z) \rangle$$

Recap: Kernel Trick

Recap: Kernel Trick

- Compute $K(\phi(x^{(i)}), \phi(x^{(j)})) = < \phi(x^{(i)}), \phi(x^{(j)}) >$ for all i, j

n x n

Recap: Kernel Trick

- Compute $K(\phi(x^{(i)}), \phi(x^{(j)})) = < \phi(x^{(i)}), \phi(x^{(j)}) >$ for all i, j
- Loop $$\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j K(x^{(i)}, x^{(j)}) \right) \quad \forall i \in \{1, \ldots, n\}$$

Recap: Kernel Trick

- Compute $K(\phi(x^{(i)}), \phi(x^{(j)})) = < \phi(x^{(i)}), \phi(x^{(j)}) >$ for all i, j
- Loop $\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j K(x^{(i)}, x^{(j)}) \right) \quad \forall i \in \{1, \ldots, n\}$

Recall that n is the number of data samples

Recap: Kernel Trick

- Compute $K(\phi(x^{(i)}), \phi(x^{(j)})) = < \phi(x^{(i)}), \phi(x^{(j)}) >$ for all i, j
- Loop $\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j K(x^{(i)}, x^{(j)}) \right) \quad \forall i \in \{1, \ldots, n\}$

Recall that n is the number of data samples

- Inference: $\theta^T \phi(x) = \sum_{i=1}^{n} \beta_i \phi(x^{(i)})^T \phi(x) = \sum_{i=1}^{n} \beta_i K(x^{(i)}, x)$

Recap: Kernel Trick

- Compute $K(\phi(x^{(i)}), \phi(x^{(j)})) = < \phi(x^{(i)}), \phi(x^{(j)}) >$ for all i, j
- Loop $$\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j K(x^{(i)}, x^{(j)}) \right) \quad \forall i \in \{1, \ldots, n\}$$

Recall that n is the number of data samples

- Inference: $$\theta^T \phi(x) = \sum_{i=1}^{n} \beta_i \phi(x^{(i)})^T \phi(x) = \sum_{i=1}^{n} \beta_i K(x^{(i)}, x)$$

The Kernel function is all we need for training and inference!

Recap: Implicit Feature Map

- Explicit Feature Map: first define feature map $\phi(x)$, then compute the Kernel according to $\phi(x)$

- Implicit Feature Map: first define the Kernel Function K(), without knowing what the feature map is

Recap: Implicit Feature Map (Example)

$$K(x, z) = (x^T z)^2 \qquad x, z \in \mathbb{R}^d$$

Recap: Implicit Feature Map (Example)

$$K(x, z) = (x^T z)^2 \qquad x, z \in \mathbb{R}^d$$

What is the feature map to make K a valid kernel function?

Recap: Implicit Feature Map (Example)

$$K(x,z) = (x^T z)^2 \qquad x, z \in \mathbb{R}^d$$

What is the feature map to make K a valid kernel function?

$$\begin{aligned} K(x,z) &= \left(\sum_{i=1}^{d} x_i z_i\right)\left(\sum_{j=1}^{d} x_j z_j\right) \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} x_i x_j z_i z_j \\ &= \sum_{i,j=1}^{d} (x_i x_j)(z_i z_j) \end{aligned}$$

Recap: Implicit Feature Map (Example)

$$K(x,z) = (x^T z)^2 \qquad x, z \in \mathbb{R}^d$$

What is the feature map to make K a valid kernel function?

$$\begin{aligned} K(x,z) &= \left(\sum_{i=1}^{d} x_i z_i\right)\left(\sum_{j=1}^{d} x_j z_j\right) \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} x_i x_j z_i z_j \\ &= \sum_{i,j=1}^{d} (x_i x_j)(z_i z_j) \end{aligned}$$

$$\phi(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix}$$

$\phi(x)^T \phi(z)$

Recap: Implicit Feature Map (Example)

$$K(x,z) = (x^T z)^2 \qquad x, z \in \mathbb{R}^d$$

O(d)

What is the feature map to make K a valid kernel function?

$$\begin{aligned} K(x,z) &= \left(\sum_{i=1}^{d} x_i z_i\right)\left(\sum_{j=1}^{d} x_j z_j\right) \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} x_i x_j z_i z_j \\ &= \sum_{i,j=1}^{d} (x_i x_j)(z_i z_j) \end{aligned}$$

$$\phi(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix}$$

Requires O(d^2) compute for feature mapping

Recap: Implicit Feature Map (Example)

$$K(x,z) = (x^T z)^2 \qquad x, z \in \mathbb{R}^d$$

What is the feature map to make K a valid kernel function?

$$\begin{aligned} K(x,z) &= \left(\sum_{i=1}^{d} x_i z_i\right)\left(\sum_{j=1}^{d} x_j z_j\right) \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} x_i x_j z_i z_j \\ &= \sum_{i,j=1}^{d} (x_i x_j)(z_i z_j) \end{aligned}$$

$$\phi(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix}$$

Requires O(d^2) compute for feature mapping

Requires O(d) compute for Kernel function

Recap: What Makes a Valid Kernel Function: Necessary Condition

Recap: What Makes a Valid Kernel Function: Necessary Condition

- Kernel Matrix $K_{ij} = K(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^T \phi(x^{(j)})$

Recap: What Makes a Valid Kernel Function: Necessary Condition

- Kernel Matrix $K_{ij} = K(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^T \phi(x^{(j)})$

- K is symmetric

Recap: What Makes a Valid Kernel Function: Necessary Condition

- Kernel Matrix $K_{ij} = K(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^T \phi(x^{(j)})$

- K is symmetric

$$
\begin{aligned}
z^T K z &= \sum_i \sum_j z_i K_{ij} z_j \\
&= \sum_i \sum_j z_i \phi(x^{(i)})^T \phi(x^{(j)}) z_j \\
&= \sum_i \sum_j z_i \sum_k \phi_k(x^{(i)}) \phi_k(x^{(j)}) z_j \\
&= \sum_k \sum_i \sum_j z_i \phi_k(x^{(i)}) \phi_k(x^{(j)}) z_j \\
&= \sum_k \left(\sum_i z_i \phi_k(x^{(i)}) \right)^2 \\
&\geq 0.
\end{aligned}
$$

Recap: What Makes a Valid Kernel Function: Necessary Condition

- Kernel Matrix $K_{ij} = K(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^T \phi(x^{(j)})$

- K is symmetric

- K is positive semidefinite

$$
\begin{aligned}
z^T K z &= \sum_i \sum_j z_i K_{ij} z_j \\
&= \sum_i \sum_j z_i \phi(x^{(i)})^T \phi(x^{(j)}) z_j \\
&= \sum_i \sum_j z_i \sum_k \phi_k(x^{(i)}) \phi_k(x^{(j)}) z_j \\
&= \sum_k \sum_i \sum_j z_i \phi_k(x^{(i)}) \phi_k(x^{(j)}) z_j \\
&= \sum_k \left(\sum_i z_i \phi_k(x^{(i)}) \right)^2 \\
&\geq 0.
\end{aligned}
$$

What Makes a Valid Kernel Function: Necessary and Sufficient Condition

Theorem (Mercer). Let $K : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ be given. Then for K to be a valid (Mercer) kernel, it is necessary and sufficient that for any $\{x^{(1)}, \ldots, x^{(n)}\}$, $(n < \infty)$, the corresponding kernel matrix is symmetric positive semi-definite.

Recap: Application of Kernel Methods

Recap: Application of Kernel Methods

- In generalized linear models (which we have shown)

Recap: Application of Kernel Methods

- In generalized linear models (which we have shown)
- In support vector machines (which we will show next)

Recap: Application of Kernel Methods

- In generalized linear models (which we have shown)
- In support vector machines (which we will show next)
- Any learning algorithm that you can write in terms of only <x, z>

Recap: Application of Kernel Methods

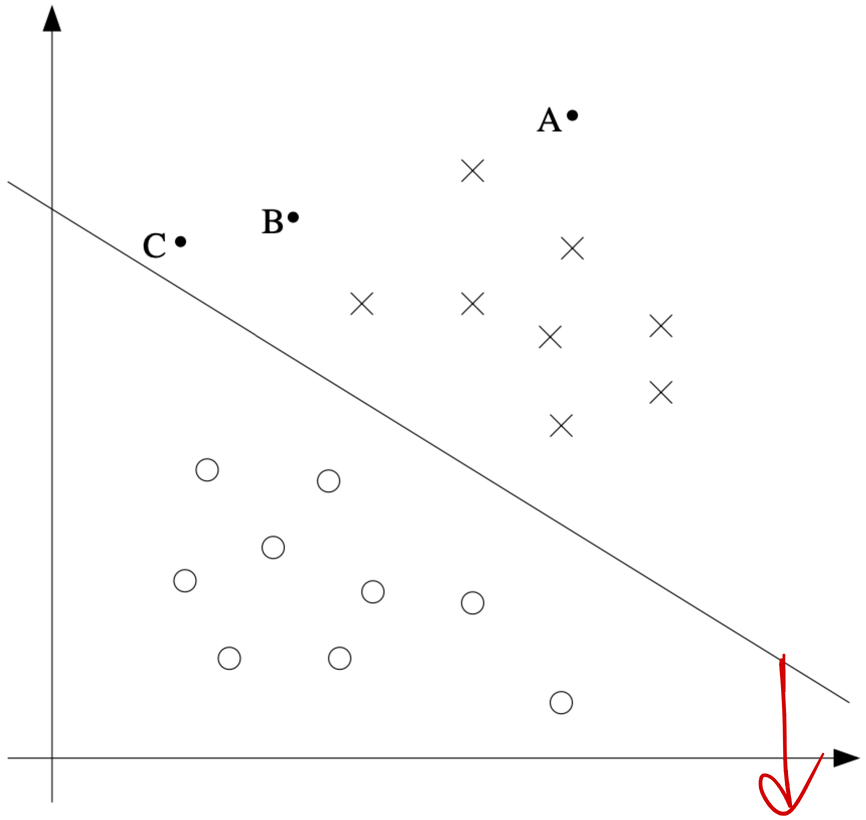
- In generalized linear models (which we have shown)
- In support vector machines (which we will show next)
- Any learning algorithm that you can write in terms of only <x, z>

Just replace <x, z> with K(x, z), you magically transform the algorithm to work efficiently in the *implicit* high dimensional feature space

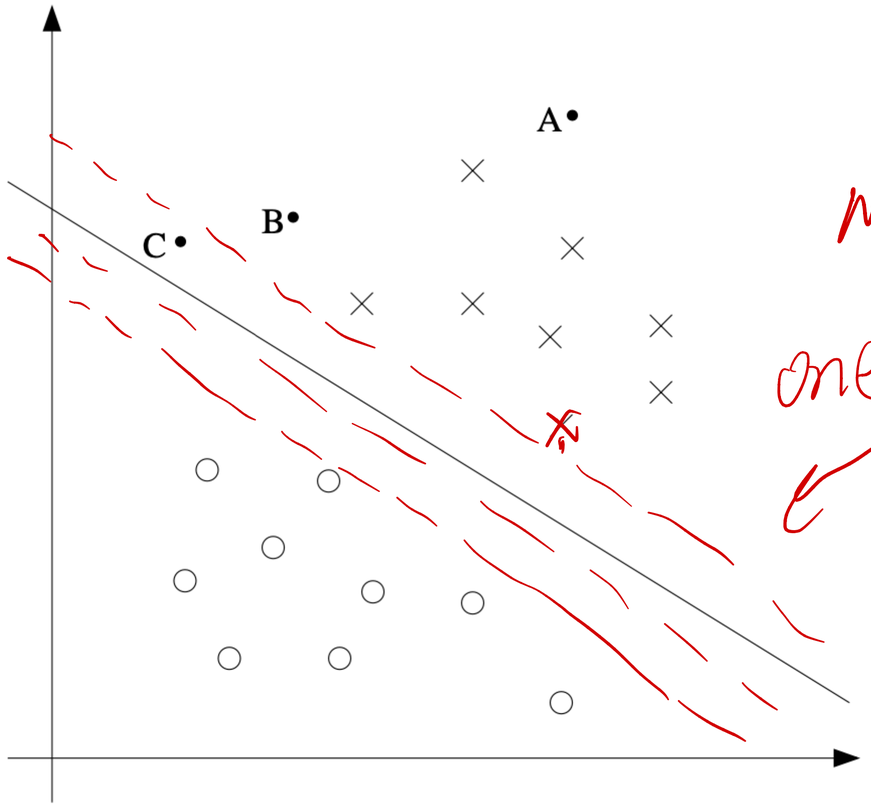
Support Vector Machines

Confidence in Logistic Regression

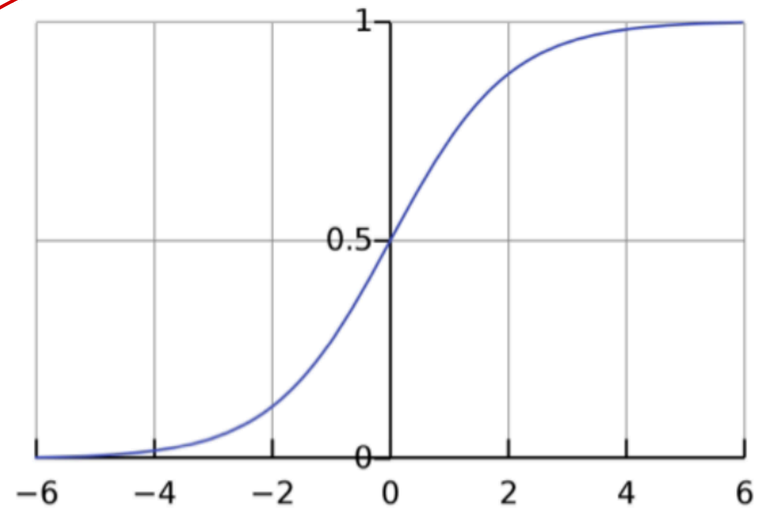
Confidence in Logistic Regression



Confidence in Logistic Regression



$$p(y) = \frac{1}{1 + e^{-\theta^T x}}$$



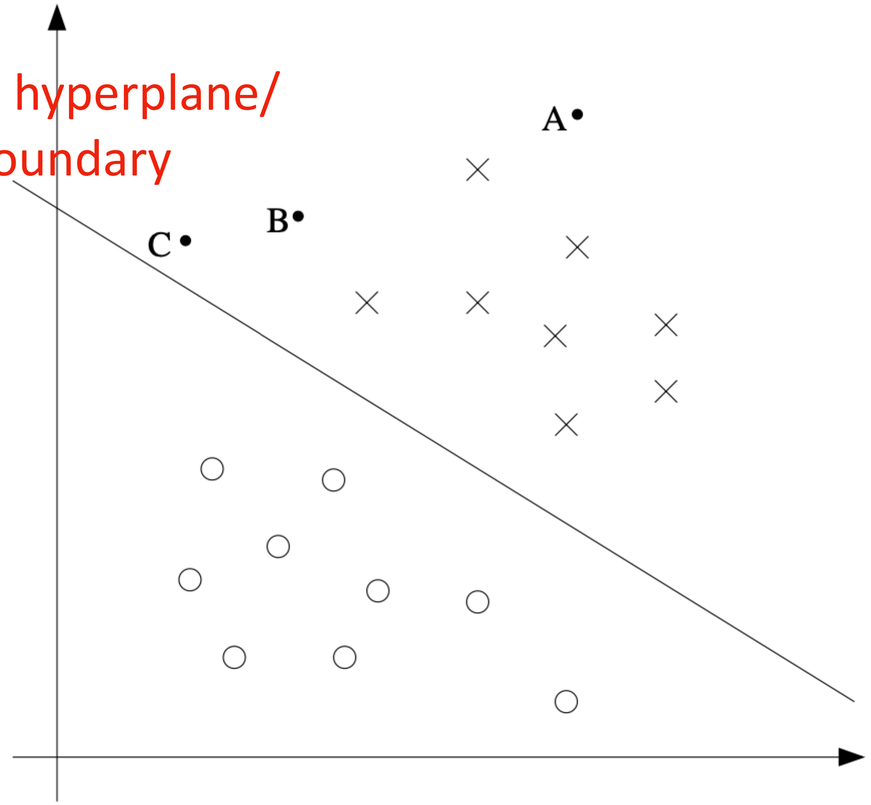
MLE

one

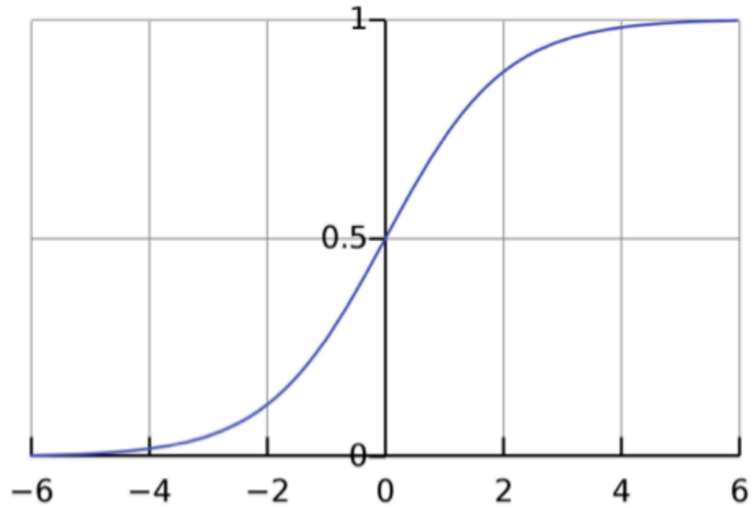
robust

Confidence in Logistic Regression

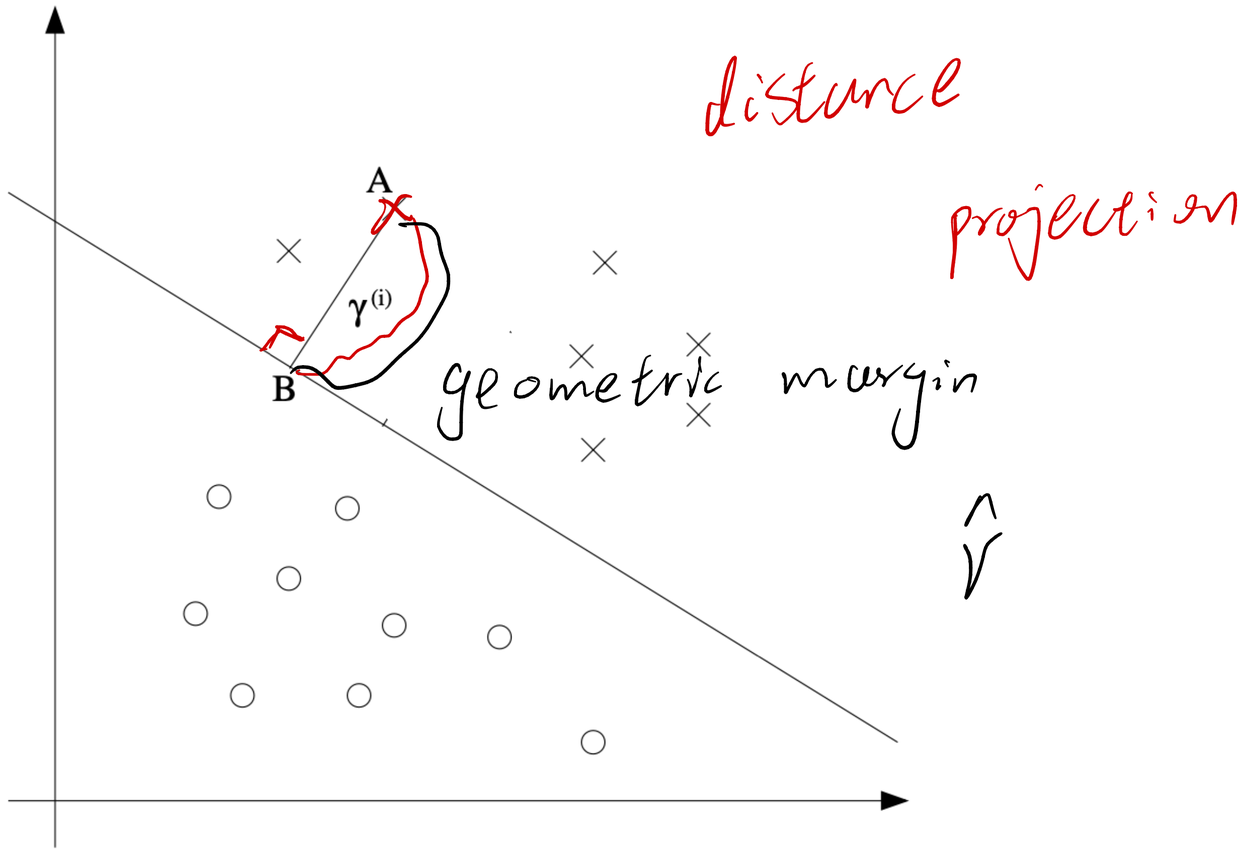
Separating hyperplane/
decision boundary



$$p(y) = \frac{1}{1 + e^{-\theta^T x}}$$



Margin



distance

projection

geometric margin

$\hat{\gamma}$

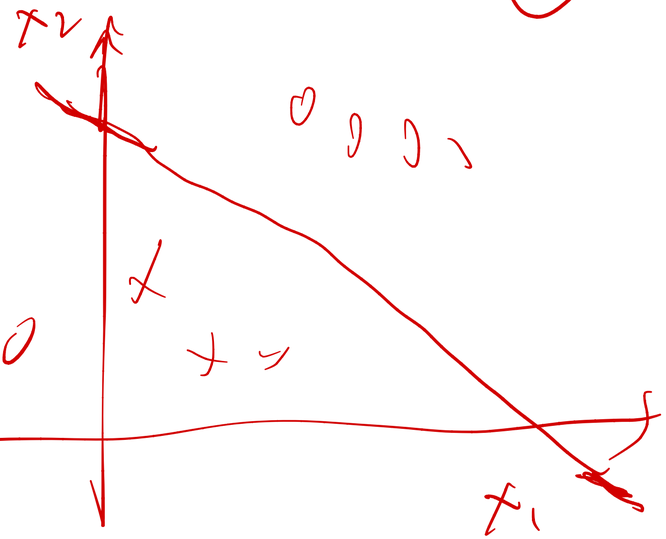
New Notations

Consider a binary classification problem, with the input feature x and $y \in \{-1,1\}$ (instead of $\{0,1\}$), the classifier is:

$$h_{w,b}(x) = g(w^T x + b)$$

$g(z) = 1$ if $z \geq 0$, and $g(z) = -1$

$$g(z) = \begin{cases} 1 & z \geq 0 \\ -1 & z < 0 \end{cases}$$



$w_1 x_1 + w_2 x_2 + b = 0$

$z = w^T x + b$

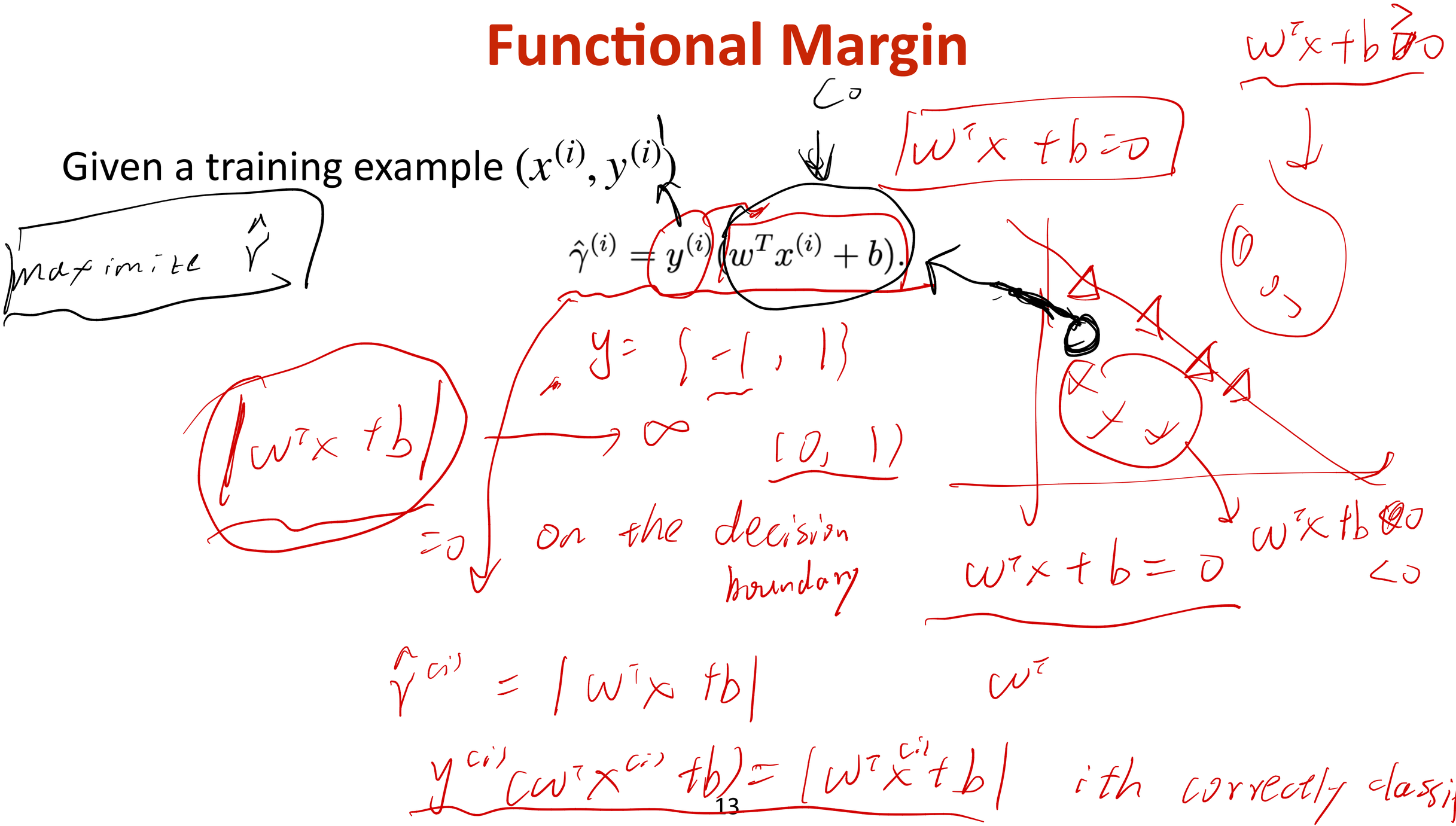
$w^T \vec{x} + b = 0$

$\vec{w}$, b

Functional Margin

Given a training example $(x^{(i)}, y^{(i)})$

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b).$$



Functional Margin

Given a training example $(x^{(i)}, y^{(i)})$

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b).$$

Given a training set $S = \{(x^{(i)}, y^{(i)}); i = 1,\ldots,n\}$

$$\hat{\gamma} = \min_{i=1,\ldots,n} \hat{\gamma}^{(i)}$$

Functional Margin

Given a training example $(x^{(i)}, y^{(i)})$

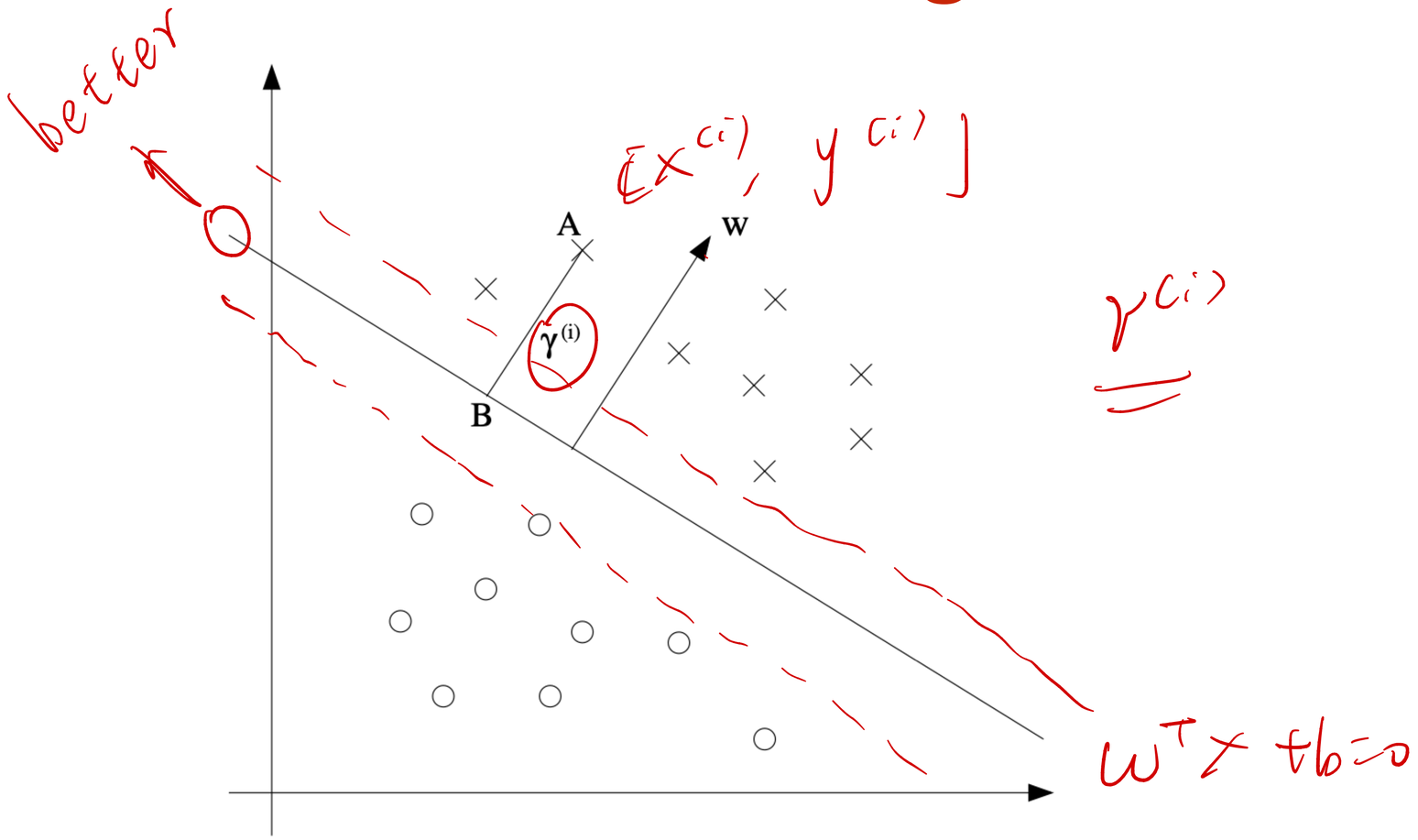
$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b).$$

Given a training set $S = \{(x^{(i)}, y^{(i)}); i = 1,\ldots,n\}$

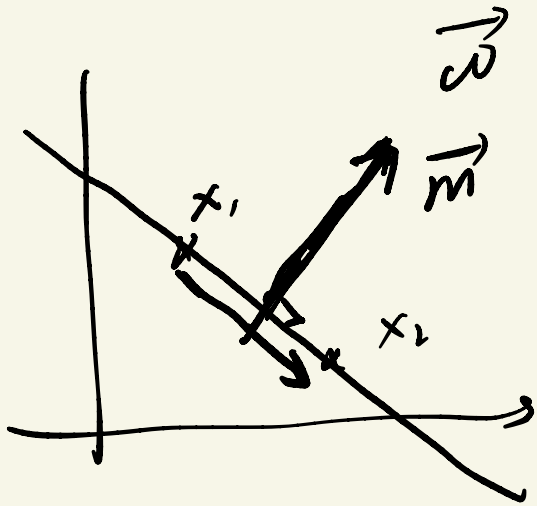
$$\hat{\gamma} = \min_{i=1,\ldots,n} \hat{\gamma}^{(i)}$$

Functional margin changes when rescaling parameters, making it a bad objective, e.g. when w->2w, b->2b, the functional margin changes while the separating plane does not really change

Geometric Margin



What is the geometric margin?



$\vec{m} \cdot (\vec{x_2} - \vec{x_1}) = 0$

$w^T \vec{x_1} + b = 0$

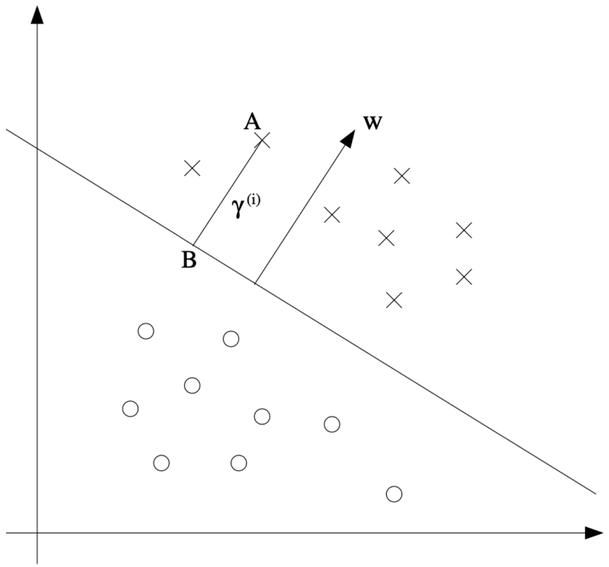
$w^T \vec{x_2} + b = 0$

$w^T (\vec{x_1} - \vec{x_2}) = 0$

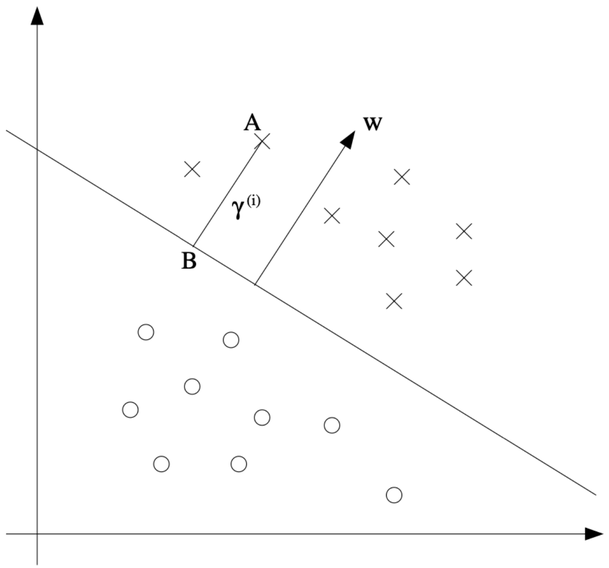
$\vec{m} = \vec{w}$

Geometric Margin

Geometric Margin



Geometric Margin

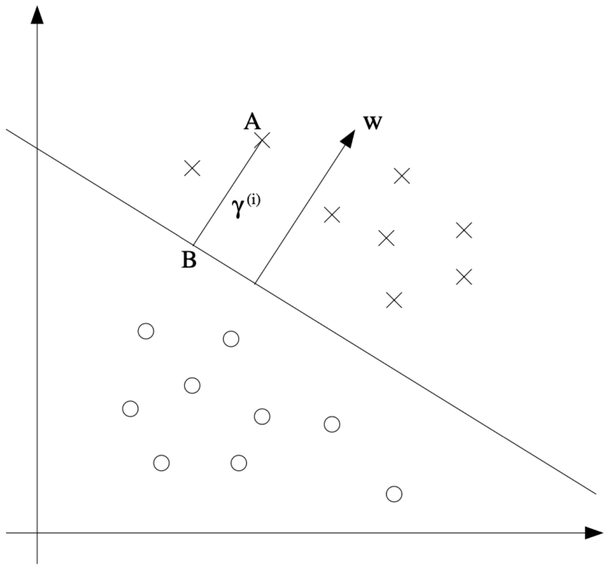


$$\vec{B} = \vec{A} - \vec{BA} = x^{(i)} - \gamma^{(i)} \cdot \frac{w}{\|w\|}$$

$$w^T \left(x^{(i)} - \gamma^{(i)} \frac{w}{||w||} \right) + b = 0.$$

$$\vec{w} \cdot \vec{B} + b = 0$$

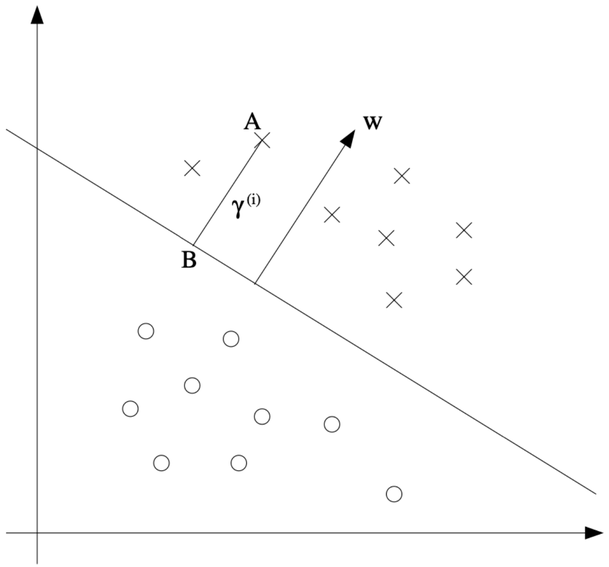
Geometric Margin



$$w^T\left(x^{(i)} - \gamma^{(i)}\frac{w}{||w||}\right) + b = 0.$$

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

Geometric Margin

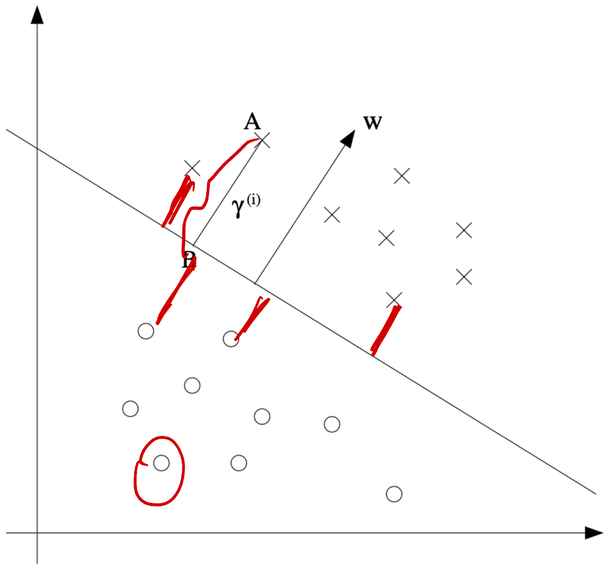


$$w^T \left(x^{(i)} - \gamma^{(i)} \frac{w}{||w||} \right) + b = 0.$$

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

Generally

Geometric Margin



$$w^T\left(x^{(i)} - \gamma^{(i)}\frac{w}{||w||}\right) + b = 0.$$

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

Generally

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

$\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b)$

Geometric Margin

Given a training set $S = \{(x^{(i)}, y^{(i)}); i = 1,\ldots,n\}$

$$\gamma = \min_{i=1,\ldots,n} \gamma^{(i)}$$

The Optimization Problem

$$\max_{w,b} \min_{i=1,\ldots,n} \gamma^{(i)}$$

max min

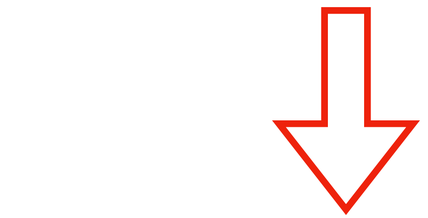
$\gamma = \min \gamma^{(i)}$

$\max_{w,b} \gamma$

$\gamma \leq \gamma^{(i)}$ for any i

The Optimization Problem

$$\max_{w,b} \min_{i=1,\ldots,n} \gamma^{(i)}$$



$$\max_{\hat{\gamma},w,b} \frac{\hat{\gamma}}{||w||}$$

$\hat{\gamma}$ functional margin

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, n$$

$\hat{\gamma} \to -\infty$

The Optimization Problem

$$\max_{w,b} \min_{i=1,\ldots,n} \gamma^{(i)}$$

geometric

$$\max_{\hat{\gamma},w,b} \frac{\hat{\gamma}}{||w||}$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, n$$

$w \to 2w$

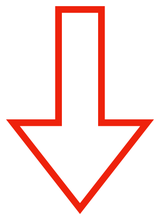
$b \to 2b.$

$\hat{\gamma} \to 2\hat{\gamma}$

Infinite solutions, as $\hat{\gamma}$ can be at any scale without changing the classifier

The Optimization Problem

$$\max_{w,b} \min_{i=1,\ldots,n} \gamma^{(i)}$$



$$\|w\| = \sqrt{w^T w}$$

$$\max_{\hat{\gamma},w,b} \frac{\hat{\gamma}}{||w||}$$

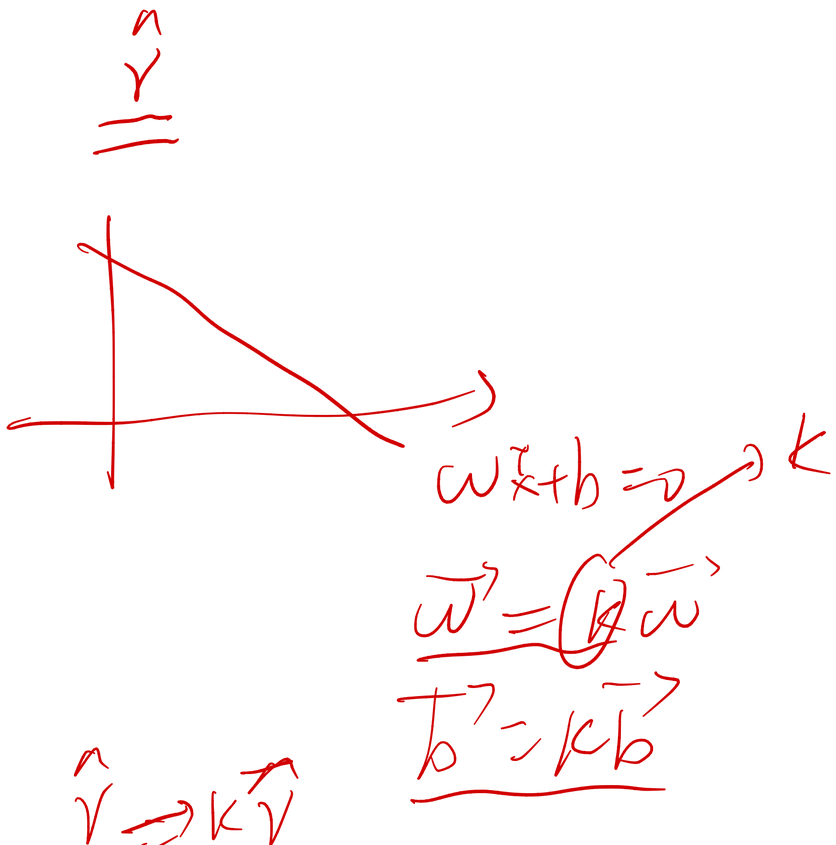
$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, n$$

Infinite solutions, as $\hat{\gamma}$ can be at any scale without changing the classifier

||w|| is not easy to deal with, non-convex objective

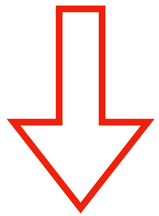
The Optimization Problem

$$\max_{\hat{\gamma},w,b} \quad \frac{\hat{\gamma}}{||w||}$$
$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, n$$



The Optimization Problem

$$\max_{\hat{\gamma},w,b} \quad \frac{\hat{\gamma}}{||w||}$$
$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, n$$



Add constraint $\hat{\gamma} = 1$

2. 3

$\arg\max_{w,b} \frac{1}{||w||}$, s.t. $y^{(i)}(w^T x^{(i)} + b) \geq 1$

$\arg\min_{w,b} ||w||^2$

The Optimization Problem

$$\max_{\hat{\gamma},w,b} \quad \frac{\hat{\gamma}}{||w||}$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, n$$

$w^T x + b = 0$

$w_{optimal}$ $b_{optimal}$

Add constraint $\hat{\gamma} = 1$

$\hat{\gamma} = 3$

$3w_{optimal}$ $3b_{optimal}$

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

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

$\min_{w,b} \frac{1}{2}\|3w\|^2$

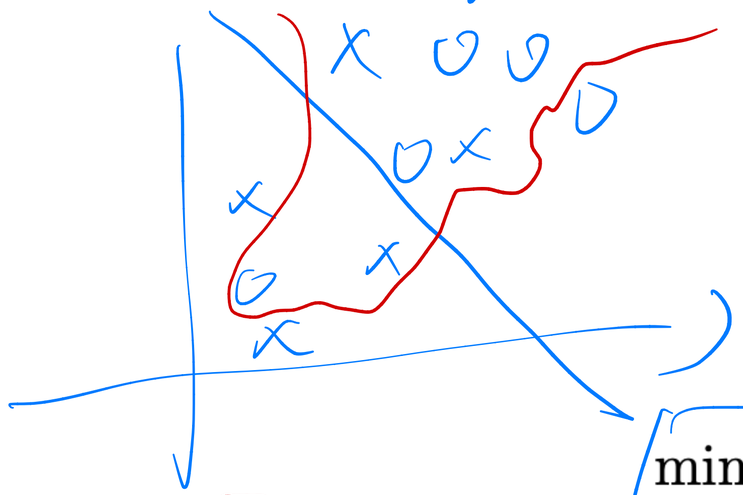
s.t. $y^{(i)}(w^T x^{(i)} + b) \geq 3$.

The Optimization Problem

quadratic constrained opt program

$$\max_{\hat{\gamma},w,b} \quad \frac{\hat{\gamma}}{||w||}$$
$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, n$$

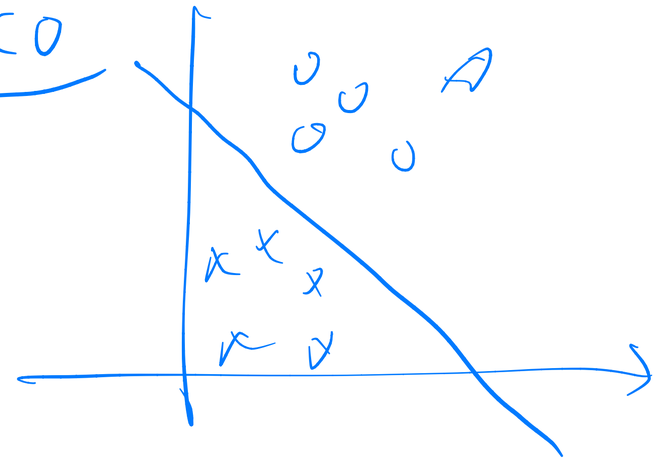
not linearly separable:



Add constraint $\hat{\gamma} = 1$

$y^{(i)}(w^T x^{(i)} + b) < 0$

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



$\langle x^{(i)}, x^{(j)} \rangle$

Assumption: the training dataset is linearly separable

$w^T \phi(x^{(i)})$

Lagrange Duality — Lagrange Multiplier

Lagrange Duality — Lagrange Multiplier

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

Lagrange Duality — Lagrange Multiplier

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

$$\mathcal{L}(w, \beta) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

Lagrange Duality — Lagrange Multiplier

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

$$\mathcal{L}(w, \beta) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

Solve w, β

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0; \quad \frac{\partial \mathcal{L}}{\partial \beta_i} = 0,$$

Lagrange Multiplier: Example

$$\min_{x,y} 5x - 3y$$

$$\text{s.t.} \quad x^2 + y^2 = 136$$

Generalized Lagrangian

Generalized Lagrangian

Primal optimization problem

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \leq 0, \quad i = 1, \ldots, k \\ & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

Generalized Lagrangian

Primal optimization problem

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \leq 0, \quad i = 1, \ldots, k \\ & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

Generalized Lagrangian

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

Generalized Lagrangian

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

Generalized Lagrangian

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\theta_{\mathcal{P}}(w) = \max_{\alpha, \beta \,:\, \alpha_i \geq 0} f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

Generalized Lagrangian

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\theta_{\mathcal{P}}(w) = \max_{\alpha, \beta \,:\, \alpha_i \geq 0} f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\theta_{\mathcal{P}}(w) = \begin{cases} f(w) & \text{if } w \text{ satisfies primal constraints} \\ \infty & \text{otherwise.} \end{cases}$$

Generalized Lagrangian

Consider this optimization problem

$$\min_{w} \theta_{\mathcal{P}}(w) = \min_{w} \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$$

Generalized Lagrangian

Consider this optimization problem

$$\min_{w} \theta_{\mathcal{P}}(w) = \min_{w} \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$$

It has exactly the same solution as our original problem

Generalized Lagrangian

Consider this optimization problem

$$\min_{w} \theta_{\mathcal{P}}(w) = \min_{w} \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$$

It has exactly the same solution as our original problem

$$p^* = \min_w \theta_{\mathcal{P}}(w)$$

The Dual Problem in Optimization

original

In optimization, sometimes the primal optimization is hard to solve, then we may find a related alternative optimization problem that can be solved more easily, to solve the orignal problem in an indirect way

The Dual Problem

$$\theta_{\mathcal{D}}(\alpha, \beta) = \min_{w} \mathcal{L}(w, \alpha, \beta)$$

The Dual Problem

$$\theta_{\mathcal{D}}(\alpha, \beta) = \min_{w} \mathcal{L}(w, \alpha, \beta)$$

The dual optimization problem

$$\max_{\alpha,\beta\,:\,\alpha_i \geq 0} \theta_{\mathcal{D}}(\alpha, \beta) = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_{w} \mathcal{L}(w, \alpha, \beta)$$

The Dual Problem

$$\theta_{\mathcal{D}}(\alpha, \beta) = \min_{w} \mathcal{L}(w, \alpha, \beta)$$

The dual optimization problem

$$\max_{\alpha,\beta\,:\,\alpha_i \geq 0} \theta_{\mathcal{D}}(\alpha, \beta) = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_{w} \mathcal{L}(w, \alpha, \beta)$$

The primal optimization problem

$$\min_{w} \theta_{\mathcal{P}}(w) = \min_{w} \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$$

The Dual Problem

$$\theta_{\mathcal{D}}(\alpha, \beta) = \min_{w} \mathcal{L}(w, \alpha, \beta)$$

The dual optimization problem

$$\max_{\alpha,\beta\,:\,\alpha_i \geq 0} \theta_{\mathcal{D}}(\alpha, \beta) = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_{w} \mathcal{L}(w, \alpha, \beta)$$

The primal optimization problem

$$\min_{w} \theta_{\mathcal{P}}(w) = \min_{w} \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$$

What is the relation of the two problems?

The Dual Problem

$$d^* = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_{w} \mathcal{L}(w, \alpha, \beta) \leq \min_{w} \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta) = p^*$$

The Dual Problem

$$d^* = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_{w} \mathcal{L}(w,\alpha,\beta) \leq \min_{w} \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w,\alpha,\beta) = p^*$$

$$\max_{x} \min_{y} f(x,y) \leq \min_{y} \max_{x} f(x,y)$$

The Dual Problem

$$d^* = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_{w} \mathcal{L}(w,\alpha,\beta) \leq \min_{w} \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w,\alpha,\beta) = p^*$$

$$\max_x \min_y f(x,y) \leq \min_y \max_x f(x,y)$$

Under certain conditions: $d^* = p^*$

The Dual Problem

$$d^* = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_{w} \mathcal{L}(w,\alpha,\beta) \leq \min_{w} \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w,\alpha,\beta) = p^*$$

$$\max_{x} \min_{y} f(x,y) \leq \min_{y} \max_{x} f(x,y)$$

Under certain conditions: $d^* = p^*$ Zero-duality Gap

The Dual Problem

$$d^* = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_{w} \mathcal{L}(w,\alpha,\beta) \leq \min_{w} \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w,\alpha,\beta) = p^*$$

$$\max_{x} \min_{y} f(x,y) \leq \min_{y} \max_{x} f(x,y)$$

Under certain conditions: $d^* = p^*$ Zero-duality Gap

What are the conditions?

Slater's Condition

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \leq 0, \quad i = 1, \ldots, k \\ & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

Slater's Condition

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \leq 0, \quad i = 1, \ldots, k \\ & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

- f(w) and g(w) are convex
- $h_i(w)$ is affine (i.e. linear)
- $g_i(w)$ are strictly feasible for all i, which means there exists some w so that $g_i(w) < 0$ for all i

Slater's Condition

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \leq 0, \quad i = 1, \ldots, k \\ & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

- f(w) and g(w) are convex
- $h_i(w)$ is affine (i.e. linear)
- $g_i(w)$ are strictly feasible for all i, which means there exists some w so that $g_i(w) < 0$ for all i

If slater's condition holds, then $d^* = p^*$

Slater's Condition

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \leq 0, \quad i = 1, \ldots, k \\ & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

- f(w) and g(w) are convex
- $h_i(w)$ is affine (i.e. linear)
- $g_i(w)$ are strictly feasible for all i, which means there exists some w so that $g_i(w) < 0$ for all i

If slater's condition holds, then $d^* = p^*$

The primal optimization problem of SVM satisfies the slater's condition

KKT Conditions

KKT Conditions

Zero duality gap is sufficient and necessary (i.e. equivalent) to satisfy KKT Conditions:

KKT Conditions

Zero duality gap is sufficient and necessary (i.e. equivalent) to satisfy KKT Conditions:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

KKT Conditions

Zero duality gap is sufficient and necessary (i.e. equivalent) to satisfy KKT Conditions:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\begin{aligned}
\frac{\partial}{\partial w_i} \mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, d \\
\frac{\partial}{\partial \beta_i} \mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, \ldots, k \\
g_i(w^*) &\leq 0, \quad i = 1, \ldots, k \\
\alpha^* &\geq 0, \quad i = 1, \ldots, k
\end{aligned}$$

KKT Conditions

Zero duality gap is sufficient and necessary (i.e. equivalent) to satisfy KKT Conditions:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\begin{aligned}
\frac{\partial}{\partial w_i} \mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, d \\
\frac{\partial}{\partial \beta_i} \mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, \ldots, k \\
g_i(w^*) &\leq 0, \quad i = 1, \ldots, k \\
\alpha^* &\geq 0, \quad i = 1, \ldots, k
\end{aligned}$$

Normal Lagrange multiplier equations

KKT Conditions

Zero duality gap is sufficient and necessary (i.e. equivalent) to satisfy KKT Conditions:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\begin{aligned}
\frac{\partial}{\partial w_i} \mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, d \\
\frac{\partial}{\partial \beta_i} \mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, \ldots, k \\
g_i(w^*) &\leq 0, \quad i = 1, \ldots, k \\
\alpha^* &\geq 0, \quad i = 1, \ldots, k
\end{aligned}$$

Normal Lagrange multiplier equations

The original constraints

KKT Conditions

Zero duality gap is sufficient and necessary (i.e. equivalent) to satisfy KKT Conditions:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\begin{aligned}
\frac{\partial}{\partial w_i} \mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, d \\
\frac{\partial}{\partial \beta_i} \mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, \ldots, k \\
g_i(w^*) &\leq 0, \quad i = 1, \ldots, k \\
\alpha^* &\geq 0, \quad i = 1, \ldots, k
\end{aligned}$$

KKT Conditions

Zero duality gap is sufficient and necessary (i.e. equivalent) to satisfy KKT Conditions:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, d \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, \ldots, k \\
g_i(w^*) &\leq 0, \quad i = 1, \ldots, k \\
\alpha^* &\geq 0, \quad i = 1, \ldots, k
\end{aligned}$$

If $\alpha_i^* > 0$, then $g_i(w^*) = 0$, the inequality is actually equality

Thank You!
Q & A