SVM and Kernels

Alexey Zaytsev Slides by Evgeny Burnaev Errors on slides by me

Skoltech, Moscow, Russia

SKOITECH Skilone Institute at Science and Technology

1/48

Outline

- 1 Convex optimization and Duality: Basics
- Support Vector Machine
- SVMs with kernels
- Support Vector Regression



- Convex optimization and Duality: Basics
- 2 Support Vector Machine

- SVMs with kernels
- 4 Support Vector Regression

Standard form problem (not necessarily convex)

$$\begin{aligned} & \text{minimize}_{\mathbf{x}} \quad f_0(\mathbf{x}) \\ & \text{subject to} \quad f_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m \\ & \quad h_i(\mathbf{x}) = 0, \quad i = 1, \dots, p \end{aligned}$$

variable $\mathbf{x} \in X \subseteq \mathbb{R}^d$, optimal value f^*

• Lagrangian: $L: \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$, with $dom(L) = X \times \mathbb{R}^m \times \mathbb{R}^p$

$$L(\mathbf{x}, \lambda, \nu) = f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x})$$

- weighted sum of objective and constraint functions
- $-\lambda_i$ is Lagrange multiplier associated with $f_i(\mathbf{x}) \leq 0$
- $-\nu_i$ is Lagrange multiplier associate with $h_i(\mathbf{x})=0$

4□ > 4回 > 4 = > 4 = > = のQ⊙

4/48

Standard form problem (not necessarily convex)

minimize_x
$$f_0(\mathbf{x})$$

subject to $f_i(\mathbf{x}) \leq 0, \ i = 1, \dots, m$
 $h_i(\mathbf{x}) = 0, \ i = 1, \dots, p$

variable $\mathbf{x} \in X \subseteq \mathbb{R}^d$, optimal value f^*

• Lagrangian: $L: \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$, with $\mathrm{dom}(L) = X \times \mathbb{R}^m \times \mathbb{R}^p$

$$L(\mathbf{x}, \lambda, \nu) = f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x})$$

- weighted sum of objective and constraint functions
- λ_i is Lagrange multiplier associated with $f_i(\mathbf{x}) \leq 0$
- $-\nu_i$ is Lagrange multiplier associate with $h_i(\mathbf{x}) = 0$

A.Zaytsev, ML

4/48

• Lagrange dual function $g: \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$

$$g(\lambda, \nu) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \nu)$$

= $\inf_{\mathbf{x} \in X} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i h_i(\mathbf{x}) \right),$

g is concave, can be $-\infty$ for some λ, ν

- Lower bound property: if $\lambda \geq 0$, then $g(\lambda, \nu) \leq f^*$
- proof: if $\widetilde{\mathbf{x}}$ is feasible and $\lambda \geq 0$, then

$$f_0(\widetilde{\mathbf{x}}) \ge L(\widetilde{\mathbf{x}}, \lambda, \nu) \ge \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \nu) = g(\lambda, \nu)$$

minimizing over all feasible $\widetilde{\mathbf{x}}$ gives $f^* \geq g(\lambda, \nu)$

4 D > 4 B > 4 E > 4 E > 9 Q O

5/48

• Lagrange dual function $g: \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$

$$g(\lambda, \nu) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \nu)$$

= $\inf_{\mathbf{x} \in X} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i h_i(\mathbf{x}) \right),$

g is concave, can be $-\infty$ for some λ, ν

- Lower bound property: if $\lambda \geq 0$, then $g(\lambda, \nu) \leq f^*$
- proof: if $\widetilde{\mathbf{x}}$ is feasible and $\lambda \geq 0$, then

$$f_0(\widetilde{\mathbf{x}}) \ge L(\widetilde{\mathbf{x}}, \lambda, \nu) \ge \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \nu) = g(\lambda, \nu)$$

minimizing over all feasible $\widetilde{\mathbf{x}}$ gives $f^* \geq g(\lambda, \nu)$

• Lagrange dual function $g: \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$

$$g(\lambda, \nu) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \nu)$$

= $\inf_{\mathbf{x} \in X} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i h_i(\mathbf{x}) \right),$

g is concave, can be $-\infty$ for some λ, ν

- Lower bound property: if $\lambda \geq 0$, then $g(\lambda, \nu) \leq f^*$
- proof: if $\widetilde{\mathbf{x}}$ is feasible and $\lambda \geq 0$, then

$$f_0(\widetilde{\mathbf{x}}) \ge L(\widetilde{\mathbf{x}}, \lambda, \nu) \ge \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \nu) = g(\lambda, \nu)$$

minimizing over all feasible $\widetilde{\mathbf{x}}$ gives $f^* \geq g(\lambda, \nu)$

4 D > 4 B > 4 B > B 9 9 9

5/48

The dual problem

Lower bound property: if $\lambda \geq 0$, then $g(\lambda, \nu) \leq f^*$

- ullet finds best lower bound on f^* , obtained from Lagrange dual function
- a convex optimization problem; optimal value denoted g'
- λ, ν are dual feasible if $\lambda \geq 0$, $(\lambda, \nu) \in \text{dom}(g)$

4□ > 4回 > 4 直 > 4 直 > 直 の 9 ○ ○

6/48

Lagrange dual problem

$$\begin{aligned} & \text{maximize}_{\lambda,\nu} & & g(\lambda,\nu) \\ & \text{subject to} & & \lambda \geq 0 \end{aligned}$$

- ullet finds best lower bound on f^* , obtained from Lagrange dual function
- ullet a convex optimization problem; optimal value denoted g^*
- λ, ν are dual feasible if $\lambda \geq 0$, $(\lambda, \nu) \in \text{dom}(g)$

A.Zaytsev, ML Global retrain flower self-through 6/48

Lagrange dual problem

$$\begin{aligned} & \text{maximize}_{\lambda,\nu} & & g(\lambda,\nu) \\ & \text{subject to} & & \lambda \geq 0 \end{aligned}$$

- ullet finds best lower bound on f^* , obtained from Lagrange dual function
- ullet a convex optimization problem; optimal value denoted g^*
- λ, ν are dual feasible if $\lambda \geq 0$, $(\lambda, \nu) \in \text{dom}(g)$

Lagrange dual problem

$$\begin{aligned} & \text{maximize}_{\lambda,\nu} & & g(\lambda,\nu) \\ & \text{subject to} & & \lambda \geq 0 \end{aligned}$$

- ullet finds best lower bound on f^* , obtained from Lagrange dual function
- ullet a convex optimization problem; optimal value denoted g^*
- λ, ν are dual feasible if $\lambda \geq 0$, $(\lambda, \nu) \in \text{dom}(g)$

A.Zaytsev, ML

6/48

Lagrange dual problem

$$\begin{aligned} & \text{maximize}_{\lambda,\nu} & & g(\lambda,\nu) \\ & \text{subject to} & & \lambda \geq 0 \end{aligned}$$

- ullet finds best lower bound on f^* , obtained from Lagrange dual function
- ullet a convex optimization problem; optimal value denoted g^*
- λ, ν are dual feasible if $\lambda \geq 0$, $(\lambda, \nu) \in \text{dom}(g)$

6/48

Weak and strong duality

weak duality: $g^* \leq f^*$

- always holds (for convex and nonconvex problems)
- can be used to find nontrivial lower bounds for difficult problems

```
strong duality: g^* = f^*
```

- does not hold in general
- (usually) holds for convex problems
- conditions that guarantee strong duality in convex problems are called constraint qualifications

7/48

Weak and strong duality

weak duality: $g^* \leq f^*$

- always holds (for convex and nonconvex problems)
- can be used to find nontrivial lower bounds for difficult problems

strong duality: $g^* = f^*$

- does not hold in general
- (usually) holds for convex problems
- conditions that guarantee strong duality in convex problems are called constraint qualifications

- Slater's condition (or Slater condition) is a sufficient condition for strong duality to hold for a convex optimization problem
- Strong duality holds for a convex problem

$$\begin{aligned} & \text{minimize}_{\mathbf{x}} & & f_0(\mathbf{x}) \\ & \text{subject to} & f_i(\mathbf{x}) \leq 0, & i = 1, \dots, m \\ & & A\mathbf{x} = b \end{aligned}$$

$$\exists \mathbf{x} \in \mathbf{int}(X) : f_i(\mathbf{x}) < 0, i = 1, \dots, m, A\mathbf{x} = b$$

ullet also guarantees that the dual optimum is attained (if $f^* > -\infty$)

(ロ) (部) (注) (注) 注 り(C)

8/48

A.Zaytsev, ML Subservice of Green

- Slater's condition (or Slater condition) is a sufficient condition for strong duality to hold for a convex optimization problem
- Strong duality holds for a convex problem

$$\begin{aligned} & \text{minimize}_{\mathbf{x}} \quad f_0(\mathbf{x}) \\ & \text{subject to} \quad f_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m \\ & A\mathbf{x} = b \end{aligned}$$

$$\exists \mathbf{x} \in \mathbf{int}(X) : f_i(\mathbf{x}) < 0, i = 1, \dots, m, A\mathbf{x} = b$$

ullet also guarantees that the dual optimum is attained (if $f^* > -\infty$)

(ロ) (部) (注) (注) 注 り(())

8/48

A.Zaytsev, ML Submitted to the Control of the Contr

- Slater's condition (or Slater condition) is a sufficient condition for strong duality to hold for a convex optimization problem
- Strong duality holds for a convex problem

$$\begin{aligned} & \text{minimize}_{\mathbf{x}} \quad f_0(\mathbf{x}) \\ & \text{subject to} \quad f_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m \\ & A\mathbf{x} = b \end{aligned}$$

$$\exists \mathbf{x} \in \mathbf{int}(X) : f_i(\mathbf{x}) < 0, \ i = 1, \dots, m, \ A\mathbf{x} = b$$

ullet also guarantees that the dual optimum is attained (if $f^* > -\infty$)

◆ロ > ◆部 > ◆注 > ◆注 > 注 の < @ >

8/48

- Slater's condition (or Slater condition) is a sufficient condition for strong duality to hold for a convex optimization problem
- Strong duality holds for a convex problem

$$\begin{aligned} & \text{minimize}_{\mathbf{x}} \quad f_0(\mathbf{x}) \\ & \text{subject to} \quad f_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m \\ & A\mathbf{x} = b \end{aligned}$$

$$\exists \mathbf{x} \in \mathbf{int}(X) : f_i(\mathbf{x}) < 0, i = 1, \dots, m, A\mathbf{x} = b$$

ullet also guarantees that the dual optimum is attained (if $f^* > -\infty$)

A.Zaytsev, ML

8/48

Assume strong duality holds, \mathbf{x}^* is primal optimal, (λ^*, ν^*) is dual optimal

$$f_0(\mathbf{x}^*) = g(\lambda^*, \nu^*) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \nu^*) = \inf_{\mathbf{x}} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}) \right)$$

$$\leq f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}^*)$$

$$\leq f_0(\mathbf{x}^*)$$

hence, the two inequlities hold with equality

- \mathbf{x}^* minimizes $L(\mathbf{x}, \lambda^*, \nu^*)$
- $\lambda_i^* f_i(\mathbf{x}^*) = 0$ for i = 1, 2, ..., m (known as complementary slackness):

$$\lambda_i^* > 0 \Rightarrow f_i(\mathbf{x}^*) = 0, \quad f_i(\mathbf{x}^*) < 0 \Rightarrow \lambda_i^* = 0$$

Assume strong duality holds, \mathbf{x}^* is primal optimal, (λ^*, ν^*) is dual optimal

$$f_0(\mathbf{x}^*) = g(\lambda^*, \nu^*) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \nu^*) = \inf_{\mathbf{x}} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}) \right)$$

$$\leq f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}^*)$$

$$\leq f_0(\mathbf{x}^*)$$

hence, the two inequlities hold with equality

- \mathbf{x}^* minimizes $L(\mathbf{x}, \lambda^*, \nu^*)$
- $\lambda_i^* f_i(\mathbf{x}^*) = 0$ for i = 1, 2, ..., m (known as complementary slackness):

$$\lambda_i^* > 0 \Rightarrow f_i(\mathbf{x}^*) = 0, \ f_i(\mathbf{x}^*) < 0 \Rightarrow \lambda_i^* = 0$$

Assume strong duality holds, \mathbf{x}^* is primal optimal, (λ^*, ν^*) is dual optimal

$$f_0(\mathbf{x}^*) = g(\lambda^*, \nu^*) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \nu^*) = \inf_{\mathbf{x}} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}) \right)$$

$$\leq f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}^*)$$

$$\leq f_0(\mathbf{x}^*)$$

hence, the two inequlities hold with equality

- \mathbf{x}^* minimizes $L(\mathbf{x}, \lambda^*, \nu^*)$
- $\lambda_i^* f_i(\mathbf{x}^*) = 0$ for i = 1, 2, ..., m (known as complementary slackness):

$$\lambda_i^* > 0 \Rightarrow f_i(\mathbf{x}^*) = 0, \ f_i(\mathbf{x}^*) < 0 \Rightarrow \lambda_i^* = 0$$

Assume strong duality holds, \mathbf{x}^* is primal optimal, (λ^*, ν^*) is dual optimal

$$f_0(\mathbf{x}^*) = g(\lambda^*, \nu^*) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \nu^*) =$$

$$\inf_{\mathbf{x}} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}) \right)$$

$$\leq f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}^*)$$

$$\leq f_0(\mathbf{x}^*)$$

hence, the two inequlities hold with equality

- \mathbf{x}^* minimizes $L(\mathbf{x}, \lambda^*, \nu^*)$
- $\lambda_i^* f_i(\mathbf{x}^*) = 0$ for i = 1, 2, ..., m (known as complementary slackness):

$$\lambda_i^* > 0 \Rightarrow f_i(\mathbf{x}^*) = 0, \ f_i(\mathbf{x}^*) < 0 \Rightarrow \lambda_i^* = 0$$

Assume strong duality holds, \mathbf{x}^* is primal optimal, (λ^*, ν^*) is dual optimal

$$f_0(\mathbf{x}^*) = g(\lambda^*, \nu^*) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \nu^*) =$$

$$\inf_{\mathbf{x}} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}) \right)$$

$$\leq f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}^*)$$

$$\leq f_0(\mathbf{x}^*)$$

hence, the two inequlities hold with equality

- \mathbf{x}^* minimizes $L(\mathbf{x}, \lambda^*, \nu^*)$
- $\lambda_i^* f_i(\mathbf{x}^*) = 0$ for i = 1, 2, ..., m (known as complementary slackness):

$$\lambda_i^* > 0 \Rightarrow f_i(\mathbf{x}^*) = 0, \ f_i(\mathbf{x}^*) < 0 \Rightarrow \lambda_i^* = 0$$

Assume strong duality holds, \mathbf{x}^* is primal optimal, (λ^*, ν^*) is dual optimal

$$f_0(\mathbf{x}^*) = g(\lambda^*, \nu^*) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \nu^*) = \inf_{\mathbf{x}} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}) \right)$$

$$\leq f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}^*)$$

$$\leq f_0(\mathbf{x}^*)$$

hence, the two inequlities hold with equality

- \mathbf{x}^* minimizes $L(\mathbf{x}, \lambda^*, \nu^*)$
- $\lambda_i^* f_i(\mathbf{x}^*) = 0$ for i = 1, 2, ..., m (known as complementary slackness):

$$\lambda_i^* > 0 \Rightarrow f_i(\mathbf{x}^*) = 0, \quad f_i(\mathbf{x}^*) < 0 \Rightarrow \lambda_i^* = 0$$

- 1. Primal constraints: $f_i(\mathbf{x}) \leq 0$, i = 1, 2, ..., m, $h_i(\mathbf{x}) = 0$, i = 1, 2, ..., p
- 2. Dual constraints: $\lambda \geq 0$
- 3. Complementary slackness: $\lambda_i f_i(\mathbf{x}) = 0, i = 1, 2, \dots, m$
- 4. Gradient of Lagrangian with respect to x vanishes:

$$\nabla f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i \nabla f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i \nabla h_i(\mathbf{x}) = 0$$

• If strong duality holds and x, λ, ν are optimal, then they must satisfy the KKT conditions

10/48

A.Zaytsev, ML Substitute of Consultation

- 1. Primal constraints: $f_i(\mathbf{x}) \leq 0$, i = 1, 2, ..., m, $h_i(\mathbf{x}) = 0$, i = 1, 2, ..., p
- 2. Dual constraints: $\lambda \geq 0$
- 3. Complementary slackness: $\lambda_i f_i(\mathbf{x}) = 0, i = 1, 2, \dots, m$
- 4. Gradient of Lagrangian with respect to x vanishes:

$$\nabla f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i \nabla f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i \nabla h_i(\mathbf{x}) = 0$$

• If strong duality holds and x, λ, ν are optimal, then they must satisfy the KKT conditions

◆ロ > ◆ 個 > ◆ 達 > ◆ 達 ・ 釣 へ ②

10/48

A.Zaytsev, ML Subunivation of Concession

- 1. Primal constraints: $f_i(\mathbf{x}) \leq 0$, i = 1, 2, ..., m, $h_i(\mathbf{x}) = 0$, i = 1, 2, ..., p
- 2. Dual constraints: $\lambda \geq 0$
- 3. Complementary slackness: $\lambda_i f_i(\mathbf{x}) = 0$, $i = 1, 2, \dots, m$
- 4. Gradient of Lagrangian with respect to x vanishes:

$$\nabla f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i \nabla f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i \nabla h_i(\mathbf{x}) = 0$$

• If strong duality holds and x, λ, ν are optimal, then they must satisfy the KKT conditions

4□ > <</p>
4□ >
4 = >
5
9
0

10/48

A. Zaytsev, ML Supposed the Control of the Control

- 1. Primal constraints: $f_i(\mathbf{x}) \leq 0$, i = 1, 2, ..., m, $h_i(\mathbf{x}) = 0$, i = 1, 2, ..., p
- 2. Dual constraints: $\lambda \geq 0$
- 3. Complementary slackness: $\lambda_i f_i(\mathbf{x}) = 0$, $i = 1, 2, \dots, m$
- 4. Gradient of Lagrangian with respect to \mathbf{x} vanishes:

$$\nabla f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i \nabla f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i \nabla h_i(\mathbf{x}) = 0$$

• If strong duality holds and x, λ, ν are optimal, then they must satisfy the KKT conditions

◆ロ > ◆ 個 > ◆ 達 > ◆ 達 ・ 釣 へ ②

10/48

Karush-Kuhn-Tucker (KKT) conditions

The following four conditions for \mathbf{x}, λ, ν are called KKT conditions (for a problem with differentiable f_i, h_i)

- 1. Primal constraints: $f_i(\mathbf{x}) \leq 0$, i = 1, 2, ..., m, $h_i(\mathbf{x}) = 0$, i = 1, 2, ..., p
- 2. Dual constraints: $\lambda \geq 0$
- 3. Complementary slackness: $\lambda_i f_i(\mathbf{x}) = 0$, $i = 1, 2, \dots, m$
- 4. Gradient of Lagrangian with respect to x vanishes:

$$\nabla f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i \nabla f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i \nabla h_i(\mathbf{x}) = 0$$

• If strong duality holds and \mathbf{x}, λ, ν are optimal, then they must satisfy the KKT conditions

◆ロト ◆部 → ◆差 > ・差 ● からぐ

10/48

KKT conditions for convex problem

If $\widetilde{\mathbf{x}},\widetilde{\lambda},\widetilde{\nu}$ satisfy KKT for a convex problem, then they are optimal:

- from slackness: $f_0(\widetilde{\mathbf{x}}) = L(\widetilde{\mathbf{x}}, \widetilde{\lambda}, \widetilde{\nu})$
- $\bullet \ \ \text{from 4th condition (and convexity)} \colon \ g(\widetilde{\lambda},\widetilde{\nu}) = L(\widetilde{\mathbf{x}},\widetilde{\lambda},\widetilde{\nu})$

hence,
$$f_0(\widetilde{\mathbf{x}}) = g(\widetilde{\lambda}, \widetilde{\nu})$$

If Slater's condition is satisfied:

- f x is optimal if and only if there exist $\lambda,
 u$ that satisfy KKT conditions
- recall that Slater implies strong duality, and dual optimum is attained
- generalizes optimality condition $\nabla f_0(\mathbf{x}) = 0$ for unconstrained problem

A.Zaytsev, ML Substitute of Location Horizon 11/48

- Convex optimization and Duality: Basic
- Support Vector Machine

SVMs with kernels

Support Vector Regression

A.Zaytsev, ML

• Training data: sample drawn i.i.d. w.r.t. D on $X \subseteq \mathbb{R}^d$

$$S_m = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\} \in \{X \times \{-1, +1\}\}^m$$

- **Problem**: find hypothesis $h: X \to \{-1, +1\}$ in H (classifier) with small generalization error R(h)
- ullet First we consider linear classification (hyperplanes) if dimension d is not too large

◆ロ > ◆ 個 > ◆ 重 > ◆ 重 ・ の へ ②

Guidan variant (from east Torkning) 13/48

• Training data: sample drawn i.i.d. w.r.t. D on $X \subseteq \mathbb{R}^d$

$$S_m = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\} \in \{X \times \{-1, +1\}\}^m$$

- **Problem**: find hypothesis $h: X \to \{-1, +1\}$ in H (classifier) with small generalization error R(h)
- ullet First we consider linear classification (hyperplanes) if dimension d is not too large

4 D > 4 B > 4 E > 4 E > 9 Q C

A.Zaytsev, ML

13/48

• Training data: sample drawn i.i.d. w.r.t. D on $X \subseteq \mathbb{R}^d$

$$S_m = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\} \in \{X \times \{-1, +1\}\}^m$$

- **Problem**: find hypothesis $h: X \to \{-1, +1\}$ in H (classifier) with small generalization error R(h)
- ullet First we consider linear classification (hyperplanes) if dimension d is not too large

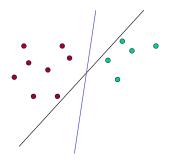
(ロ) (画) (差) (差) (差) のQで

A.Zaytsev, ML

13/48

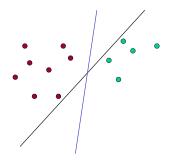
Support Vectors

- Support vectors are the data points that lie closest to the decision surface (or hyperplane)
- Support vectors are the elements of the training set that would change the position of the dividing hyperplane if removed
- They are the data points most difficult to classify
- In general, lots of possible solutions for a hyperplane



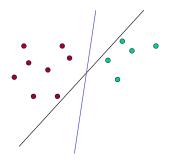
Support Vectors

- Support vectors are the data points that lie closest to the decision surface (or hyperplane)
- Support vectors are the elements of the training set that would change the position of the dividing hyperplane if removed
- They are the data points most difficult to classify
- In general, lots of possible solutions for a hyperplane



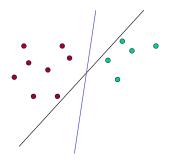
Support Vectors

- Support vectors are the data points that lie closest to the decision surface (or hyperplane)
- Support vectors are the elements of the training set that would change the position of the dividing hyperplane if removed
- They are the data points most difficult to classify
- In general, lots of possible solutions for a hyperplane

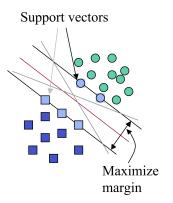


Support Vectors

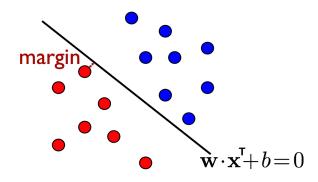
- Support vectors are the data points that lie closest to the decision surface (or hyperplane)
- Support vectors are the elements of the training set that would change the position of the dividing hyperplane if removed
- They are the data points most difficult to classify
- In general, lots of possible solutions for a hyperplane



Support Vector Machine



- Support Vector Machine (SVM) finds an optimal solution
- SVMs maximize the margin (the "street") around the separating hyperplane
- The decision function is fully specified by a (usually very small) subset of training samples, the support vectors

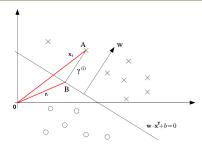


• classifiers: $H = \{\mathbf{x} \to \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x}^\top + b), \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\}$

◆ロ → ← 部 → ← き → を ● り へ で

16/48

A. Zaytsev, ML Sudan retract of Construction and Text



- $\gamma^{(i)}$ is a distance from \mathbf{x}_i to the hyperplane $\mathbf{w} \cdot \mathbf{x}^\top + b = 0$
- ullet $\mathbf{w}/\|\mathbf{w}\|$ is a unit perpendicular to the hyperplane
- ullet Vector ${f r}_i$ of a point B is equal to

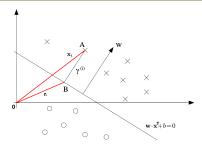
$$\mathbf{r}_i = \mathbf{x}_i - \gamma^{(i)} \mathbf{w} / ||\mathbf{w}||$$

ullet Since point B belongs to the hyperplane: $\mathbf{w}\cdot\mathbf{r}_i^{ op}+b=0$, i.e

 \bullet Thus we get that $\gamma^{(i)} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \mathbf{x}_i^\top + \frac{b}{\|\mathbf{w}\|}$



17/48



- $\gamma^{(i)}$ is a distance from \mathbf{x}_i to the hyperplane $\mathbf{w} \cdot \mathbf{x}^\top + b = 0$
- ullet $\mathbf{w}/\|\mathbf{w}\|$ is a unit perpendicular to the hyperplane
- ullet Vector ${f r}_i$ of a point B is equal to

$$\mathbf{r}_i = \mathbf{x}_i - \gamma^{(i)} \mathbf{w} / \|\mathbf{w}\|$$

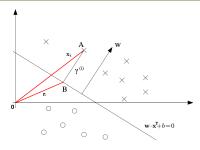
• Since point B belongs to the hyperplane: $\mathbf{w} \cdot \mathbf{r}_i^\top + b = 0$, i.e

 \bullet Thus we get that $\gamma^{(i)} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \mathbf{x}_i^\top + \frac{b}{\|\mathbf{w}\|}$



17/48

A. Zaytsev, ML Substitute of the control of the con



- $oldsymbol{\circ} \gamma^{(i)}$ is a distance from \mathbf{x}_i to the hyperplane $\mathbf{w} \cdot \mathbf{x}^{\top} + b = 0$
- ullet $\mathbf{w}/\|\mathbf{w}\|$ is a unit perpendicular to the hyperplane
- ullet Vector ${f r}_i$ of a point B is equal to

$$\mathbf{r}_i = \mathbf{x}_i - \gamma^{(i)} \mathbf{w} / \|\mathbf{w}\|$$

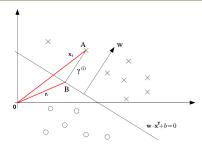
• Since point B belongs to the hyperplane: $\mathbf{w} \cdot \mathbf{r}_i^\top + b = 0$, i.e

 \bullet Thus we get that $\gamma^{(i)} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \mathbf{x}_i^\top + \frac{b}{\|\mathbf{w}\|}$

◆ロト ◆部 ◆ をき ◆ を を め の

17/48

A. Zaytsev, ML Subscienta of Core and Technique



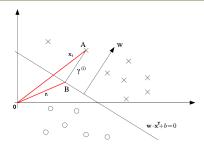
- $oldsymbol{\circ} \gamma^{(i)}$ is a distance from \mathbf{x}_i to the hyperplane $\mathbf{w} \cdot \mathbf{x}^{\top} + b = 0$
- ullet $\mathbf{w}/\|\mathbf{w}\|$ is a unit perpendicular to the hyperplane
- ullet Vector ${f r}_i$ of a point B is equal to

$$\mathbf{r}_i = \mathbf{x}_i - \gamma^{(i)} \mathbf{w} / \|\mathbf{w}\|$$

 \bullet Since point B belongs to the hyperplane: $\mathbf{w}\cdot\mathbf{r}_i^\top+b=0$, i.e.

 \bullet Thus we get that $\gamma^{(i)} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \mathbf{x}_i^\top + \frac{b}{\|\mathbf{w}\|}$

17/48



- $oldsymbol{\circ} \gamma^{(i)}$ is a distance from \mathbf{x}_i to the hyperplane $\mathbf{w} \cdot \mathbf{x}^\top + b = 0$
- $\bullet \ \mathbf{w}/\|\mathbf{w}\|$ is a unit perpendicular to the hyperplane
- ullet Vector ${f r}_i$ of a point B is equal to

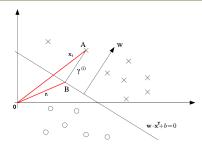
$$\mathbf{r}_i = \mathbf{x}_i - \gamma^{(i)} \mathbf{w} / \|\mathbf{w}\|$$

• Since point B belongs to the hyperplane: $\mathbf{w} \cdot \mathbf{r}_i^\top + b = 0$, i.e.

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

 \bullet Thus we get that $\gamma^{(i)} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \mathbf{x}_i^\top + \frac{b}{\|\mathbf{w}\|}$

A.Zaytsev, ML Skoltech



- $oldsymbol{\circ} \gamma^{(i)}$ is a distance from \mathbf{x}_i to the hyperplane $\mathbf{w} \cdot \mathbf{x}^{\top} + b = 0$
- $\bullet \ \mathbf{w}/\|\mathbf{w}\|$ is a unit perpendicular to the hyperplane
- ullet Vector ${f r}_i$ of a point B is equal to

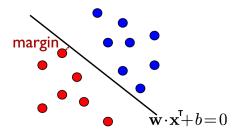
$$\mathbf{r}_i = \mathbf{x}_i - \gamma^{(i)} \mathbf{w} / \|\mathbf{w}\|$$

 \bullet Since point B belongs to the hyperplane: $\mathbf{w}\cdot\mathbf{r}_i^\top+b=0$, i.e.

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

• Thus we get that $\gamma^{(i)} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \mathbf{x}_i^\top + \frac{b}{\|\mathbf{w}\|}$

17/48



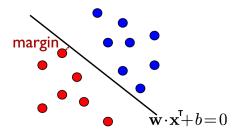
In general case

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

The margin is

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1, m]} \gamma^{(i)} \right]$$

18/48



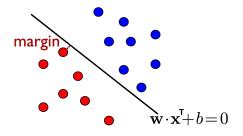
In general case

$$\gamma^{(i)} = \left| \frac{\mathbf{w}}{\|\mathbf{w}\|} \mathbf{x}_i^\top + \frac{b}{\|\mathbf{w}\|} \right| = \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$$

The margin is

$$\rho = \max_{\mathbf{w}, b: \, y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1, m]} \gamma^{(i)} \right.$$

18/48



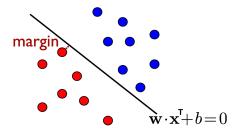
In general case

$$\gamma^{(i)} = \left|\frac{\mathbf{w}}{\|\mathbf{w}\|}\mathbf{x}_i^\top + \frac{b}{\|\mathbf{w}\|}\right| = \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \to \min_{i \in [1,m]} \text{ (worst case!)}$$

The margin is

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0} \left[\min_{i \in [1, m]} \gamma^{(i)} \right]$$

18/48



In general case

$$\gamma^{(i)} = \left|\frac{\mathbf{w}}{\|\mathbf{w}\|}\mathbf{x}_i^\top + \frac{b}{\|\mathbf{w}\|}\right| = \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \to \min_{i \in [1,m]} \text{ (worst case!)}$$

The margin is

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0} \left[\min_{i \in [1, m]} \gamma^{(i)} \right]$$

Skoltech

Optimization problem

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

• Target $\min_{i \in [1,m]} rac{|\mathbf{w} \cdot \mathbf{x}_i^+ + b|}{\|\mathbf{w}\|}$ is scale-invariant, i.e.

- Inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0 \Leftrightarrow y_i(\widetilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \widetilde{b}) \ge 0$
- We can always re-normalize w and b such that
 - it holds: $\min_{i \in [1.m]} |\mathbf{w} \cdot \mathbf{x}_i^{ op} + b| = 1$
 - inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^{\perp} + b) \geq 0$ does not change
 - target function $\min_{i \in [1,m]} rac{|\mathbf{w} \cdot \mathbf{x}_i^*| + b|}{\|\mathbf{w}\|}$ does not change

←□ → ←□ → ← = → = =

Optimization problem

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

• Target $\min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ is scale-invariant, i.e.

Inequality
$$y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0 \Leftrightarrow y_i(\widetilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \widetilde{b}) \ge 0$$

- ullet We can always re-normalize ${f w}$ and b such tha
 - it holds: $\min_{i \in [1.m]} |\mathbf{w} \cdot \mathbf{x}_i^{ op} + b| = 1$
 - inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^{\perp} + b) \geq 0$ does not change
 - target function $\min_{i \in [1,m]} rac{|\mathbf{w} \cdot \mathbf{x}_i| + \sigma_i}{\|\mathbf{w}\|}$ does not change

Optimization problem

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

• Target $\min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^{\top} + b|}{\|\mathbf{w}\|}$ is scale-invariant, i.e.

$$\min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} = \frac{\min_{i \in [1,m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| \cdot \text{const}}{\|\mathbf{w}\| \cdot \text{const}}$$

- Inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0 \Leftrightarrow y_i(\widetilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \widetilde{b}) \ge 0$
- ullet We can always re-normalize ${f w}$ and b such that
 - it holds: $\min_{i \in [1,m]} |\mathbf{w} \cdot \mathbf{x}_i| + b| = 1$
 - inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^{\top} + b) \geq 0$ does not change
 - target function $\min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i| + |\mathbf{v}|}{\|\mathbf{w}\|}$ does not change

Optimization problem

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

• Target $\min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^{\top} + b|}{\|\mathbf{w}\|}$ is scale-invariant, i.e.

$$\begin{split} & \min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} = \frac{\min_{i \in [1,m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| \cdot \text{const}}{\|\mathbf{w}\| \cdot \text{const}} = \\ & = \frac{\min_{i \in [1,m]} |\widetilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \widetilde{b}|}{\|\widetilde{\mathbf{w}}\|} \left(\widetilde{\mathbf{w}} = \mathbf{w} \cdot \text{const}, \, \widetilde{b} = b \cdot \text{const}\right) \end{split}$$

- Inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0 \Leftrightarrow y_i(\widetilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \widetilde{b}) \ge 0$
- ullet We can always re-normalize ${f w}$ and b such tha
 - It holds: $\min_{i \in [1,m]} |\mathbf{w} \cdot \mathbf{x}_i^+ + b| = 1$
 - inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^{\top} + b) \geq 0$ does not change
 - target function $\min_{i \in [1,m]} rac{|\mathbf{w} \cdot \mathbf{x}_i| + o|}{\|\mathbf{w}\|}$ does not change

Optimization problem

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

• Target $\min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^{\top} + b|}{\|\mathbf{w}\|}$ is scale-invariant, i.e.

$$\begin{split} & \min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} = \frac{\min_{i \in [1,m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| \cdot \text{const}}{\|\mathbf{w}\| \cdot \text{const}} = \\ & = \frac{\min_{i \in [1,m]} |\widetilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \widetilde{b}|}{\|\widetilde{\mathbf{w}}\|} \left(\widetilde{\mathbf{w}} = \mathbf{w} \cdot \text{const}, \, \widetilde{b} = b \cdot \text{const}\right) \end{split}$$

- Inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0 \Leftrightarrow y_i(\widetilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \widetilde{b}) \ge 0$
- ullet We can always re-normalize ${f w}$ and b such that
 - It holds: $\min_{i \in [1,m]} |\mathbf{w} \cdot \mathbf{x}_i^+ + b| = 1$
 - inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^{\top} + b) \geq 0$ does not change
 - target function $\min_{i \in [1,m]} rac{|\mathbf{w} \cdot \mathbf{x}_i| + |\mathbf{v}|}{\|\mathbf{w}\|}$ does not change

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

• Target $\min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^{\top} + b|}{\|\mathbf{w}\|}$ is scale-invariant, i.e.

$$\begin{split} & \min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} = \frac{\min_{i \in [1,m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| \cdot \text{const}}{\|\mathbf{w}\| \cdot \text{const}} = \\ & = \frac{\min_{i \in [1,m]} |\widetilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \widetilde{b}|}{\|\widetilde{\mathbf{w}}\|} \left(\widetilde{\mathbf{w}} = \mathbf{w} \cdot \text{const}, \, \widetilde{b} = b \cdot \text{const}\right) \end{split}$$

- Inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0 \Leftrightarrow y_i(\widetilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \widetilde{b}) \ge 0$
- ullet We can always re-normalize ${f w}$ and b such that
 - it holds: $\min_{i \in [1,m]} |\mathbf{w} \cdot \mathbf{x}_i^{\top} + b| = 1$
 - inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^{\top} + b) \geq 0$ does not change
 - target function $\min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^*| + b|}{\|\mathbf{w}\|}$ does not change

4 □ > <□ > < □ > < □ > < □ > < □ > < □ </p>

Optimization problem

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

• Target $\min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^{\top} + b|}{\|\mathbf{w}\|}$ is scale-invariant, i.e.

$$\begin{split} & \min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} = \frac{\min_{i \in [1,m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| \cdot \text{const}}{\|\mathbf{w}\| \cdot \text{const}} = \\ & = \frac{\min_{i \in [1,m]} |\widetilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \widetilde{b}|}{\|\widetilde{\mathbf{w}}\|} \left(\widetilde{\mathbf{w}} = \mathbf{w} \cdot \text{const}, \, \widetilde{b} = b \cdot \text{const}\right) \end{split}$$

- Inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0 \Leftrightarrow y_i(\widetilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \widetilde{b}) \ge 0$
- ullet We can always re-normalize ${f w}$ and b such that
 - it holds: $\min_{i \in [1,m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1$
 - inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^{\top} + b) \geq 0$ does not change
 - target function $\min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^*| + b|}{\|\mathbf{w}\|}$ does not change

Optimization problem

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

• Target $\min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^{\top} + b|}{\|\mathbf{w}\|}$ is scale-invariant, i.e.

$$\begin{split} & \min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} = \frac{\min_{i \in [1,m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| \cdot \text{const}}{\|\mathbf{w}\| \cdot \text{const}} = \\ & = \frac{\min_{i \in [1,m]} |\widetilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \widetilde{b}|}{\|\widetilde{\mathbf{w}}\|} \; (\widetilde{\mathbf{w}} = \mathbf{w} \cdot \text{const}, \, \widetilde{b} = b \cdot \text{const}) \end{split}$$

- Inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0 \Leftrightarrow y_i(\widetilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \widetilde{b}) \ge 0$
- ullet We can always re-normalize ${f w}$ and b such that
 - it holds: $\min_{i \in [1,m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1$
 - inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^{\top} + b) \geq 0$ does not change
 - target function $\min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i| + b|}{\|\mathbf{w}\|}$ does not change

Optimization problem

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

• Target $\min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^{\top} + b|}{\|\mathbf{w}\|}$ is scale-invariant, i.e.

$$\begin{split} & \min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} = \frac{\min_{i \in [1,m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| \cdot \text{const}}{\|\mathbf{w}\| \cdot \text{const}} = \\ & = \frac{\min_{i \in [1,m]} |\widetilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \widetilde{b}|}{\|\widetilde{\mathbf{w}}\|} \left(\widetilde{\mathbf{w}} = \mathbf{w} \cdot \text{const}, \, \widetilde{b} = b \cdot \text{const}\right) \end{split}$$

- Inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0 \Leftrightarrow y_i(\widetilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \widetilde{b}) \ge 0$
- ullet We can always re-normalize ${f w}$ and b such that
 - it holds: $\min_{i \in [1,m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1$
 - inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^{\top} + b) \geq 0$ does not change
 - target function $\min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{||\mathbf{w}||}$ does not change

A.Zaytsev, ML Sub-ridge of Group through

$$\rho = \max_{\mathbf{w},b:\ y_i(\mathbf{w}\cdot\mathbf{x}_i^\top + b) \ge 0} \left[\min_{i \in [1,m]} \frac{|\mathbf{w}\cdot\mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

$$= \max_{\mathbf{w},b:\ y_i(\mathbf{w}\cdot\mathbf{x}_i^\top + b) \ge 0} \left[\min_{i \in [1,m]} \frac{|\mathbf{w}\cdot\mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right] \text{ (scale-invar.)}$$

$$= \max_{\mathbf{w},b:\ y_i(\mathbf{w}\cdot\mathbf{x}_i^\top + b) \ge 0} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

$$= \max_{\mathbf{w},b:\ y_i(\mathbf{w}\cdot\mathbf{x}_i^\top + b) \ge 0} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

$$= \max_{\mathbf{w},b:\ y_i(\mathbf{w}\cdot\mathbf{x}_i^\top + b) \ge 1} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

Example Science and Technology 20/48

$$\rho = \max_{\mathbf{w},b:\,y_i(\mathbf{w}\cdot\mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1,m]} \frac{|\mathbf{w}\cdot\mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

$$= \max_{\mathbf{w},b:\,y_i(\mathbf{w}\cdot\mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1,m]} \frac{|\mathbf{w}\cdot\mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right] \text{ (scale-invar.)}$$

$$= \max_{\mathbf{w},b:\,y_i(\mathbf{w}\cdot\mathbf{x}_i^\top + b) \geq 0} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

$$= \max_{\mathbf{w},b:\,y_i(\mathbf{w}\cdot\mathbf{x}_i^\top + b) \geq 0} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

$$= \max_{\mathbf{w},b:\,y_i(\mathbf{w}\cdot\mathbf{x}_i^\top + b) \geq 1} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

Institute of Science and Technology 20/48

$$\rho = \max_{\mathbf{w},b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0} \left[\min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

$$= \max_{\mathbf{w},b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0} \left[\min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right] \text{ (scale-invar.)}$$

$$= \max_{\mathbf{w},b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 0} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

$$\min_{i \in [1,m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| \ge 1$$

$$= \max_{\mathbf{w},b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 1} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

20/48

$$\rho = \max_{\mathbf{w},b:\,y_i(\mathbf{w}\cdot\mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1,m]} \frac{|\mathbf{w}\cdot\mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

$$= \max_{\mathbf{w},b:\,y_i(\mathbf{w}\cdot\mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1,m]} \frac{|\mathbf{w}\cdot\mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right] \text{ (scale-invar.)}$$

$$= \max_{\mathbf{w},b:\,y_i(\mathbf{w}\cdot\mathbf{x}_i^\top + b) \geq 0} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

$$= \max_{\mathbf{w},b:\,y_i(\mathbf{w}\cdot\mathbf{x}_i^\top + b) \geq 0} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

$$= \max_{\mathbf{w},b:\,y_i(\mathbf{w}\cdot\mathbf{x}_i^\top + b) \geq 1} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

(ㅁ▶◀큔▶◀불▶◀불▶ 불 쒸٩♡

A.Zaytsev, ML Subscription (Circum at Thomas 20/48

$$\max_{\mathbf{w},b: y_i(\mathbf{w}\cdot\mathbf{x}_i^\top + b) \ge 1} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

Constrained Optimization

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$
 s.t. $y_i(\mathbf{w}\cdot\mathbf{x}_i^\top+b)\geq 1,\,i\in[1,m]$

- Properties
 - Convex optimization
 - Unique solution for linearly separable case

hose institute at Science and Technology 21/48

$$\max_{\mathbf{w},b: y_i(\mathbf{w}\cdot\mathbf{x}_i^\top + b) \ge 1} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

Constrained Optimization:

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

s.t.
$$y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 1, i \in [1, m]$$

- Properties
 - Convex optimization
 - Unique solution for linearly separable case

(ロ) (部) (注) (注) 注 り(C)

21/48

$$\max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 1} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

Constrained Optimization:

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

s.t.
$$y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 1, i \in [1, m]$$

- Properties:
 - Convex optimization
 - Unique solution for linearly separable case

◆ロ > ◆部 > ◆き > き め Q ©

21/48

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i^{\top} + b) - 1]$$

• KKT conditions:

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i^{\top} = 0 \Leftrightarrow \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$$

otkase at Science and Technology 22/48

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1]$$

KKT conditions:

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i^{\top} = 0 \Leftrightarrow \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$$

at Science and Technology 22/48

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i^{\top} + b) - 1]$$

KKT conditions:

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i^{\top} = 0 \Leftrightarrow \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$$
$$\nabla_b L = -\sum_{i=1}^{m} \alpha_i y_i = 0 \Leftrightarrow \sum_{i=1}^{m} \alpha_i y_i = 0$$

at Science and Technology 22/48

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1]$$

KKT conditions:

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i^{\top} = 0 \Leftrightarrow \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$$
$$\nabla_b L = -\sum_{i=1}^{m} \alpha_i y_i = 0 \Leftrightarrow \sum_{i=1}^{m} \alpha_i y_i = 0$$
$$\forall i \in [1, m], \ \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i^{\top} + b) - 1] = 0$$

A.Zaytsev, ML 22/48

Complementary conditions:

$$\alpha_i[y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1] = 0 \Rightarrow \alpha_i = 0 \text{ or } y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) = 1$$

ullet Support vectors: vectors \mathbf{x}_i such that

$$\alpha_i \neq 0$$
 and $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) = 1$

 Support Vectors: Input vectors that just touch the boundary of the margin (street)

Station instant of Science and Technology

Complementary conditions:

$$\alpha_i[y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1] = 0 \Rightarrow \alpha_i = 0 \text{ or } y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) = 1$$

ullet Support vectors: vectors ${f x}_i$ such that

$$\alpha_i \neq 0$$
 and $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) = 1$

 Support Vectors: Input vectors that just touch the boundary of the margin (street)

SKUILECII

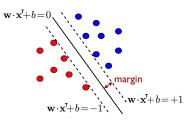
Complementary conditions:

$$\alpha_i[y_i(\mathbf{w}\cdot\mathbf{x}_i^\top+b)-1]=0 \Rightarrow \ \alpha_i=0 \ \text{or} \ y_i(\mathbf{w}\cdot\mathbf{x}_i^\top+b)=1$$

ullet Support vectors: vectors ${f x}_i$ such that

$$\alpha_i \neq 0$$
 and $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) = 1$

 Support Vectors: Input vectors that just touch the boundary of the margin (street)



◆ロト ◆部 → ◆注 → 注 ・ り へ

A.Zaytsev, ML Substitute (Constitution) 23/48

From KKT we get that optimal

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$$

$$L = \underbrace{\frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \right\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}_{-\frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}$$
$$- \underbrace{\sum_{i=1}^{m} \alpha_i y_i}_{-0} b + \sum_{i=1}^{m} \alpha_i$$

$$L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)$$

From KKT we get that optimal

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$$

• Plugging ${\bf w}$ in $L=\frac{1}{2}\|{\bf w}\|^2-\sum_{i=1}^m\alpha_i[y_i({\bf w}\cdot{\bf x}_i^\top+b)-1]$ we get

$$L = \underbrace{\frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \right\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}_{-\frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}$$
$$- \underbrace{\sum_{i=1}^{m} \alpha_i y_i}_{-0} b + \sum_{i=1}^{m} \alpha_i$$

Thus

$$L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)$$

Skoltech

• From KKT we get that optimal

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$$

• Plugging ${\bf w}$ in $L=\frac{1}{2}\|{\bf w}\|^2-\sum_{i=1}^m\alpha_i[y_i({\bf w}\cdot{\bf x}_i^\top+b)-1]$ we get

$$L = \underbrace{\frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \right\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}_{-\frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}$$
$$- \underbrace{\sum_{i=1}^{m} \alpha_i y_i}_{=0} b + \sum_{i=1}^{m} \alpha_i$$

Thus

$$L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)$$

Skoltech

From KKT we get that optimal

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

• Plugging w in $L = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^m \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1]$ we get

$$L = \underbrace{\frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \right\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}_{-\frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}$$
$$- \underbrace{\sum_{i=1}^{m} \alpha_i y_i}_{=0} b + \sum_{i=1}^{m} \alpha_i$$

Thus

$$L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)$$

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top})$$

s.t.
$$\alpha_i \geq 0$$
 and $\sum_{i=1}^n \alpha_i y_i = 0, i \in [1,m]$

Optimal

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$$

• Solution: classifier based on the separating hyperplane $\mathbf{w} \cdot \mathbf{x}^{\top} + b = 0$ has the form $h(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x}^{\top} + b)$, i.e

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top})$$

s.t.
$$\alpha_i \geq 0$$
 and $\sum_{i=1}^{n} \alpha_i y_i = 0, i \in [1, m]$

Optimal

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$$

• Solution: classifier based on the separating hyperplane $\mathbf{w} \cdot \mathbf{x}^{\top} + b = 0$ has the form $h(\mathbf{x}) = \operatorname{sign}(\mathbf{w} \cdot \mathbf{x}^{\top} + b)$, i.e

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top})$$

s.t.
$$\alpha_i \geq 0$$
 and $\sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m]$

Optimal

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$$

 Solution: classifier based on the separating hyperplane $\mathbf{w} \cdot \mathbf{x}^{\top} + b = 0$ has the form $h(\mathbf{x}) = \operatorname{sign}(\mathbf{w} \cdot \mathbf{x}^{\top} + b)$, i.e.

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top})$$

s.t.
$$\alpha_i \geq 0$$
 and $\sum_{i=1}^{n} \alpha_i y_i = 0, i \in [1,m]$

Optimal

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

• **Solution**: classifier based on the separating hyperplane $\mathbf{w} \cdot \mathbf{x}^\top + b = 0$ has the form $h(\mathbf{x}) = \operatorname{sign}(\mathbf{w} \cdot \mathbf{x}^\top + b)$, i.e.

$$h(\mathbf{x}) = \operatorname{sign}\left(\sum_{i=1}^{m} \alpha_i y_i(\mathbf{x}_i \cdot \mathbf{x}^{\top}) + b\right),$$

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top})$$

s.t.
$$\alpha_i \geq 0$$
 and $\sum_i \alpha_i y_i = 0, \, i \in [1,m]$

Optimal

A.Zaytsev, ML

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

• Solution: classifier based on the separating hyperplane $\mathbf{w} \cdot \mathbf{x}^\top + b = 0$ has the form $h(\mathbf{x}) = \operatorname{sign}(\mathbf{w} \cdot \mathbf{x}^\top + b)$, i.e.

$$h(\mathbf{x}) = \operatorname{sign}\left(\sum_{i=1}^{m} \alpha_i y_i(\mathbf{x}_i \cdot \mathbf{x}^{\top}) + b\right),$$

with $b = y_i - \sum_{i=1}^m \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}_i^\top)$ for any SV \mathbf{x}_i

Skoltech

SVM: Non-separable case

• **Problem**: data often not linearly separable in practice. For any hyperplane there exists \mathbf{x}_i , such that

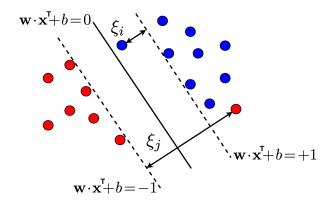
$$y_i[\mathbf{w} \cdot \mathbf{x}_i^{\top} + b] \ngeq 1$$

• Approach: relax constraints using slack variables $\xi_i \geq 0$

$$y_i[\mathbf{w} \cdot \mathbf{x}_i^{\top} + b] \ge 1 - \xi_i$$

A.Zaytsev, ML Skoltech

Soft-margin hyperplane



- Support vectors: points along the margin or outliers
- Soft margin: $\rho = \frac{1}{\|\mathbf{w}\|}$

◆□ ▶ ◆■ ▶ ◆ ■ → りへC

Stations instant of Science and Technology

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

s.t.
$$y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \ge 1 - \xi_i$$
 and $\xi_i \ge 0, i \in [1, m]$

- Properties:
 - Convex optimization
 - Unique solution
 - $-C \ge 0$ is a trade-off parameter

Stations traites at Science and Technology 28/48

Comments

- How to determine C?
- The problem of determining a hyperplane minimizing the train error is NP-complete (as a function of dimension)
- Other convex functions of the slack variables can be used

Skoltech

• Lagrangian: for all $\mathbf{w}, b, \alpha_i \geq 0, \beta_i \geq 0$

$$L(\mathbf{w}, b, \xi, \alpha, \beta) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i$$
$$- \sum_{i=1}^m \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1 + \xi_i] - \sum_{i=1}^m \beta_i \xi_i$$

KKT conditions:

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^{m} \alpha_{i} y_{i} \mathbf{x}_{i} = 0 \Leftrightarrow \mathbf{w} = \sum_{i=1}^{m} \alpha_{i} y_{i} \mathbf{x}_{i}$$

$$\nabla_{b} L = -\sum_{i=1}^{m} \alpha_{i} y_{i} = 0 \Leftrightarrow \sum_{i=1}^{m} \alpha_{i} y_{i} = 0$$

$$\nabla_{\xi_{i}} L = C - \alpha_{i} - \beta_{i} = 0 \Leftrightarrow \alpha_{i} + \beta_{i} = C$$

$$\forall i \in [1, m], \ \alpha_{i} [u_{i}(\mathbf{w} \cdot \mathbf{x}^{T} + b) - 1 + \xi_{i}] = 0 \ \text{and} \ \beta_{i} \xi_{i} = 0$$

4 D > 4 B > 4 B > B 9 9 9

Skoltech

• Lagrangian: for all $\mathbf{w}, b, \alpha_i \geq 0, \beta_i \geq 0$

$$L(\mathbf{w}, b, \xi, \alpha, \beta) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i$$
$$- \sum_{i=1}^m \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1 + \xi_i] - \sum_{i=1}^m \beta_i \xi_i$$

KKT conditions:

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i = 0 \Leftrightarrow \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$$

$$\nabla_b L = -\sum_{i=1}^{m} \alpha_i y_i = 0 \Leftrightarrow \sum_{i=1}^{m} \alpha_i y_i = 0$$

$$\nabla_{\xi_i} L = C - \alpha_i - \beta_i = 0 \Leftrightarrow \alpha_i + \beta_i = C$$

$$\forall i \in [1, m], \ \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1 + \xi_i] = 0 \ \text{and} \ \beta_i \xi_i = 0$$

A.Zaytsev, ML Skoltech

Complementary conditions:

$$\alpha_i[y_i(\mathbf{w}\cdot\mathbf{x}_i^\top+b)-1+\xi_i]=0 \Rightarrow \ \alpha_i=0 \ \text{or} \ y_i(\mathbf{w}\cdot\mathbf{x}_i^\top+b)=1-\xi_i$$

ullet Support vectors: vectors \mathbf{x}_i such that

$$\alpha_i \neq 0$$
 and $y_i(\mathbf{w} \cdot \mathbf{x}_i^{\top} + b) = 1 - \xi_i$

Salou tetra d'Sono and Tetralay

Support Vectors

Complementary conditions:

$$\alpha_i[y_i(\mathbf{w}\cdot\mathbf{x}_i^\top+b)-1+\xi_i]=0 \Rightarrow \ \alpha_i=0 \ \text{or} \ y_i(\mathbf{w}\cdot\mathbf{x}_i^\top+b)=1-\xi_i$$

• Support vectors: vectors x_i such that

$$\alpha_i \neq 0$$
 and $y_i(\mathbf{w} \cdot \mathbf{x}_i^{\top} + b) = 1 - \xi_i$

• Plugging optimal $\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$ in L we get

$$L = \underbrace{\frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \right\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}_{-\frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}$$
$$- \underbrace{\sum_{i=1}^{m} \alpha_i y_i}_{=0} b + \sum_{i=1}^{m} \alpha_i$$

Thus

$$L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)$$

• Since $\beta_i = C - \alpha_i$, the condition $\beta_i \ge 0$ is equivalent to $\alpha_i \le C$

Stations instant of Science and Technology

32/48

• Plugging optimal $\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$ in L we get

$$L = \underbrace{\frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \right\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}_{-\frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}$$
$$- \underbrace{\sum_{i=1}^{m} \alpha_i y_i}_{=0} b + \sum_{i=1}^{m} \alpha_i$$

Thus

$$L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)$$

• Since $\beta_i = C - \alpha_i$, the condition $\beta_i \ge 0$ is equivalent to $\alpha_i \le C$

4 D > 4 B > 4 E > 4 E > 9 Q C

• Plugging optimal $\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$ in L we get

$$L = \underbrace{\frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \right\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}_{-\frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}$$
$$- \underbrace{\sum_{i=1}^{m} \alpha_i y_i}_{=0} b + \sum_{i=1}^{m} \alpha_i$$

Thus

$$L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)$$

• Since $\beta_i = C - \alpha_i$, the condition $\beta_i \ge 0$ is equivalent to $\alpha_i \le C$

◆ロト ◆部 ト ◆ 恵 ト ◆ 恵 ・ 夕 Q C

A.Zaytsev, ML Subortetra di Geno anticara

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top})$$

s.t.
$$0 \le \alpha_i \le C$$
 and $\sum_{i=1}^m \alpha_i y_i = 0, i \in [1,m]$

• **Solution**: separating hyperplane $\mathbf{w} \cdot \mathbf{x}^{\top} + b = 0$

$$h(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i(\mathbf{x}_i \cdot \mathbf{x}^{\top}) + b\right),$$

with $b = y_i - \sum_{j=1}^m \alpha_j y_j(\mathbf{x}_j \cdot \mathbf{x}_i^\top)$ for any SV \mathbf{x}_i with $0 < \alpha_i < C$

4 D > 4 B > 4 E > 4 E > 9 Q O

A.Zaytsev, ML Subunitative distance

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top})$$

s.t.
$$0 \le \alpha_i \le C$$
 and $\sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m]$

• Solution: separating hyperplane $\mathbf{w} \cdot \mathbf{x}^{\top} + b = 0$

$$h(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i(\mathbf{x}_i \cdot \mathbf{x}^\top) + b\right),$$

with $b = y_i - \sum_{i=1}^m \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}_i^\top)$ for any SV \mathbf{x}_i with $0 < \alpha_i < C$

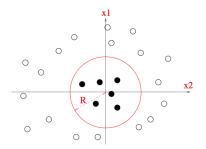
Sation retain of Science and Technology

- Convex optimization and Duality: Basics
- 2 Support Vector Machine
- 3 SVMs with kernels

Support Vector Regression

Skoltech

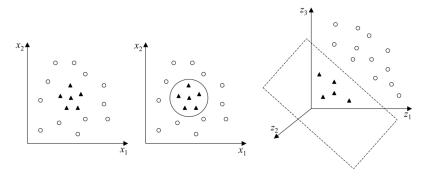
Recall: Non-linear separation



- Linear separation impossible in most problems
- \bullet Non-linear mapping $\varPhi:X\to \mathbb{H}$ from input space to high-dimensional feature space
- \bullet Generalization ability: independent of $\dim(\mathbb{H}),$ depends only on d and m

35/48

A.Zaytsev, ML



For
$$\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$$
, let $\Phi(\mathbf{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2) \in \mathbb{R}^3$. Then
$$K(\mathbf{x}', \mathbf{x}) = \Phi(\mathbf{x}') \cdot \Phi(\mathbf{x})^\top \quad \text{[dot product of features]}$$
$$= x_1^2 (x_1')^2 + 2x_1 x_2 x_1' x_2' + x_2^2 (x_2')^2$$
$$= (x_1 x_1' + x_2 x_2')^2 = (\mathbf{x}' \cdot \mathbf{x}^\top)^2$$

(ㅁㅏㅓ@ㅏㅓㅌㅏㅓㅌㅏ · ㅌ · 쒸٩(

trace of Science and Technology 36/48

Kernel Methods

- Idea:
 - Define $K: X \times X \to \mathbb{R}$ called kernel, such that

$$\Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}')^{\top} = K(\mathbf{x}, \mathbf{x}')$$

- $-\ K$ is often interpreted as a similarity measure
- Benefits:
 - Efficiency: K is often more efficient to compute than Φ and the dot product
 - Flexibility: K can be chosen arbitrarily so long as the existence of Φ is guaranteed (PDS condition or Mercer's condition)

Skoltech

37/48

Kernel Methods

Idea:

— Define $K: X \times X \to \mathbb{R}$ called kernel, such that

$$\Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}')^{\top} = K(\mathbf{x}, \mathbf{x}')$$

K is often interpreted as a similarity measure

Benefits:

- Efficiency: K is often more efficient to compute than Φ and the dot product
- Flexibility: K can be chosen arbitrarily so long as the existence of Φ is guaranteed (PDS condition or Mercer's condition)

Skoltech 37/48

Standard PDS Kernel

• Gaussian kernel:

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right), \, \sigma \neq 0$$

Skoltech
Stations (1922a) e el Science and Technology 38/48

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top})$$

s.t.
$$0 \le \alpha_i \le C$$
 and $\sum_{i=1}^m \alpha_i y_i = 0, i \in [1, m]$

• Decision function $h(\mathbf{x}) = \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x}^{\top} + b)$ has the form

$$h(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i(\mathbf{x}_i \cdot \mathbf{x}^{\top}) + b\right),$$

with

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i, \ b = y_i - \sum_{j=1}^{m} \alpha_j y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)$$

for any SV \mathbf{x}_i with $0 < \alpha_i < C$

4 D > 4 B > 4 B > B 9 9 9

A.Zaytsev, ML Skidow Ma Constant Consta

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top})$$

s.t.
$$0 \le \alpha_i \le C$$
 and $\sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m]$

 \bullet Decision function $h(\mathbf{x}) = \mathrm{sgn}(\mathbf{w} \cdot \mathbf{x}^\top + b)$ has the form

$$h(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i(\mathbf{x}_i \cdot \mathbf{x}^\top) + b\right),$$

with

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i, \ b = y_i - \sum_{j=1}^{m} \alpha_j y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)$$

for any SV \mathbf{x}_i with $0 < \alpha_i < C$

A.Zaytsev, ML Subservative of concentrations of the Concentration of the Concentration of Concentration

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j \underbrace{\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)^{\top}}_{K(\mathbf{x}_i, \mathbf{x}_j)}$$

s.t.
$$0 \le \alpha_i \le C$$
 and $\sum_{i=1}^m \alpha_i y_i = 0, i \in [1,m]$

• Decision function $h(\mathbf{x}) = \operatorname{sgn}(\mathbf{w} \cdot \Phi(\mathbf{x})^{\top} + b)$ has the form

$$h(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i \underbrace{\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})^{\top}}_{K(\mathbf{x}_i, \mathbf{x})} + b\right),\,$$

with

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \Phi(\mathbf{x}_i), \ b = y_i - \sum_{j=1}^{m} \alpha_j y_j \underbrace{\Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x}_i)^{\top}}_{K(\mathbf{x}_j, \mathbf{x}_i)}$$

for any SV \mathbf{x}_i with $0 < \alpha_i < C$

Salon return of Science and Technology 40

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j \underbrace{\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)^{\top}}_{K(\mathbf{x}_i, \mathbf{x}_j)}$$

s.t.
$$0 \le \alpha_i \le C$$
 and $\sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m]$

• Decision function $h(\mathbf{x}) = \operatorname{sgn}(\mathbf{w} \cdot \Phi(\mathbf{x})^\top + b)$ has the form

$$h(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i \underbrace{\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})^{\top}}_{K(\mathbf{x}_i, \mathbf{x})} + b\right),\,$$

with

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \Phi(\mathbf{x}_i), \ b = y_i - \sum_{j=1}^{m} \alpha_j y_j \underbrace{\Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x}_i)^{\top}}_{K(\mathbf{x}_j, \mathbf{x}_i)}$$

for any SV \mathbf{x}_i with $0 < \alpha_i < C$



A.Zaytsev, ML

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

s.t.
$$0 \le \alpha_i \le C$$
 and $\sum_{i=1}^m \alpha_i y_i = 0, i \in [1,m]$

• Decision function $h(\mathbf{x}) = \operatorname{sgn}(\mathbf{w} \cdot \Phi(\mathbf{x})^{\top} + b)$ has the form

$$h(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right),$$

with

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \Phi(\mathbf{x}_i), \ b = y_i - \sum_{i=1}^{m} \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i)$$

for any SV \mathbf{x}_i with $0 < \alpha_i < C$

A.Zaytsev, ML Subset Police Control

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

s.t.
$$0 \le \alpha_i \le C$$
 and $\sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m]$

• Decision function $h(\mathbf{x}) = \mathrm{sgn}(\mathbf{w} \cdot \Phi(\mathbf{x})^\top + b)$ has the form

$$h(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i \mathbf{K}(\mathbf{x}_i, \mathbf{x}) + b\right),$$

with

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \Phi(\mathbf{x}_i), \ b = y_i - \sum_{j=1}^{m} \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i)$$

for any SV \mathbf{x}_i with $0 < \alpha_i < C$

A.Zaytsev, ML Subscriptor of Con

Convex optimization and Duality: Basics

Support Vector Machine

- SVMs with kernels
- Support Vector Regression

42/48

A.Zaytsev, ML Skoltech

Support Vector Regression I

Hypothesis set

$$\{x \to \mathbf{w} \cdot \Phi(\mathbf{x})^{\top} + b : \mathbf{w} \in \mathbb{R}^p, b \in \mathbb{R}\}$$

• Loss function: ϵ -insensitive loss

$$L(y, y') = |y - y'|_{\epsilon} = \max(0, |y' - y| - \epsilon)$$

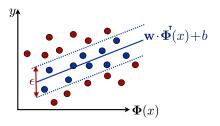


Figure – Fit "tube" with width ϵ to data

Optimization problem: similar to that of SVM

$$\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m |y_i - (\mathbf{w} \cdot \boldsymbol{\Phi}(\mathbf{x}_i)^\top + b)|_{\epsilon} \to \min_{\mathbf{w}, b}$$

Equivalent formulation

$$\begin{aligned} \min_{\mathbf{w}, b, \xi, \xi'} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m (\xi_i + \xi_i') \\ \text{subject to} \quad & (\mathbf{w} \cdot \Phi(\mathbf{x}_i)^\top + b) - y_i \leq \epsilon + \xi_i \\ & y_i - (\mathbf{w} \cdot \Phi(\mathbf{x}_i)^\top + b) \leq \epsilon + \xi_i' \\ & \xi_i \geq 0, \ \xi_i' \geq 0 \end{aligned}$$

A.Zaytsev, ML Subscript Su

Optimization problem: similar to that of SVM

$$\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m |y_i - (\mathbf{w} \cdot \Phi(\mathbf{x}_i)^\top + b)|_{\epsilon} \to \min_{\mathbf{w}, b}$$

Equivalent formulation

$$\begin{aligned} \min_{\mathbf{w},b,\xi,\xi'} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m (\xi_i + \xi_i') \\ \text{subject to} \quad & (\mathbf{w} \cdot \boldsymbol{\Phi}(\mathbf{x}_i)^\top + b) - y_i \leq \epsilon + \xi_i \\ & y_i - (\mathbf{w} \cdot \boldsymbol{\Phi}(\mathbf{x}_i)^\top + b) \leq \epsilon + \xi_i' \\ & \xi_i \geq 0, \, \xi_i' \geq 0 \end{aligned}$$

A.Zaytsev, ML 44/48

Optimization problem:

$$\begin{aligned} \max_{\boldsymbol{\alpha}, \boldsymbol{\alpha}'} &- \epsilon (\boldsymbol{\alpha}' + \boldsymbol{\alpha})^{\top} \mathbf{1} + (\boldsymbol{\alpha}' - \boldsymbol{\alpha})^{\top} \mathbf{Y} \\ &- \frac{1}{2} (\boldsymbol{\alpha}' - \boldsymbol{\alpha})^{\top} \mathbf{K} (\boldsymbol{\alpha}' - \boldsymbol{\alpha}) \\ \text{s.t. } &(\mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C}) \text{ or } (\mathbf{0} \leq \boldsymbol{\alpha}' \leq \mathbf{C}) \text{ or } ((\boldsymbol{\alpha}' - \boldsymbol{\alpha})^{\top} \mathbf{1} = 0) \end{aligned}$$

Here
$$\mathbf{K} = \{ \boldsymbol{\Phi}(\mathbf{x}_i) \cdot \boldsymbol{\Phi}(\mathbf{x}_j)^\top \}_{i,j=1}^m = \{ K(\mathbf{x}_i, \mathbf{x}_j) \}_{i,j=1}^m \in \mathbb{R}^{m \times m}$$

Solution

$$h(\mathbf{x}) = \sum_{i=1}^{m} (\alpha_i' - \alpha_i) \underbrace{K(\mathbf{x}_i, \mathbf{x})}_{\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})^{\top}} + b$$

with
$$b = \begin{cases} -\sum_{i=1}^{m} (\alpha_j' - \alpha_j) K(\mathbf{x}_j, \mathbf{x}_i) + y_i + \epsilon, & \text{when } 0 < \alpha_i < C \\ -\sum_{i=1}^{m} (\alpha_j' - \alpha_j) K(\mathbf{x}_j, \mathbf{x}_i) + y_i - \epsilon, & \text{when } 0 < \alpha_i' < C \end{cases}$$

Support vectors: points strictly outside the tube

←□ → ←□ → ← □ → ○ へ

45/48

A. Zaytsev, ML Sudan telepa til Green satter

Optimization problem:

$$\begin{aligned} & \max_{\alpha,\alpha'} - \epsilon(\alpha' + \alpha)^{\top} \mathbf{1} + (\alpha' - \alpha)^{\top} \mathbf{Y} \\ & - \frac{1}{2} (\alpha' - \alpha)^{\top} \mathbf{K} (\alpha' - \alpha) \\ & \text{s.t. } (\mathbf{0} \leq \alpha \leq \mathbf{C}) \text{ or } (\mathbf{0} \leq \alpha' \leq \mathbf{C}) \text{ or } ((\alpha' - \alpha)^{\top} \mathbf{1} = 0) \end{aligned}$$

Here
$$\mathbf{K} = \{ \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)^\top \}_{i,j=1}^m = \{ K(\mathbf{x}_i, \mathbf{x}_j) \}_{i,j=1}^m \in \mathbb{R}^{m \times m}$$

Solution

$$h(\mathbf{x}) = \sum_{i=1}^{m} (\alpha_i' - \alpha_i) \underbrace{K(\mathbf{x}_i, \mathbf{x})}_{\mathbf{\Phi}(\mathbf{x}_i) \cdot \mathbf{\Phi}(\mathbf{x})^{\top}} + b$$

with
$$b = \begin{cases} -\sum_{i=1}^{m} (\alpha_j' - \alpha_j) K(\mathbf{x}_j, \mathbf{x}_i) + y_i + \epsilon, & \text{when } 0 < \alpha_i < C \\ -\sum_{i=1}^{m} (\alpha_j' - \alpha_j) K(\mathbf{x}_j, \mathbf{x}_i) + y_i - \epsilon, & \text{when } 0 < \alpha_i' < C \end{cases}$$

Support vectors: points strictly outside the tube

45/48

A.Zaytsev, ML Susan indian and in

Optimization problem:

$$\begin{aligned} \max_{\boldsymbol{\alpha}, \boldsymbol{\alpha}'} &- \epsilon (\boldsymbol{\alpha}' + \boldsymbol{\alpha})^{\top} \mathbf{1} + (\boldsymbol{\alpha}' - \boldsymbol{\alpha})^{\top} \mathbf{Y} \\ &- \frac{1}{2} (\boldsymbol{\alpha}' - \boldsymbol{\alpha})^{\top} \mathbf{K} (\boldsymbol{\alpha}' - \boldsymbol{\alpha}) \\ \text{s.t. } &(\mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C}) \text{ or } &(\mathbf{0} \leq \boldsymbol{\alpha}' \leq \mathbf{C}) \text{ or } &((\boldsymbol{\alpha}' - \boldsymbol{\alpha})^{\top} \mathbf{1} = 0) \end{aligned}$$

Here
$$\mathbf{K} = \{ \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)^\top \}_{i,j=1}^m = \{ K(\mathbf{x}_i, \mathbf{x}_j) \}_{i,j=1}^m \in \mathbb{R}^{m \times m}$$

Solution

$$h(\mathbf{x}) = \sum_{i=1}^{m} (\alpha_i' - \alpha_i) \underbrace{K(\mathbf{x}_i, \mathbf{x})}_{\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})^{\top}} + b$$

with
$$b = \begin{cases} -\sum_{i=1}^{m} (\alpha_j' - \alpha_j) K(\mathbf{x}_j, \mathbf{x}_i) + y_i + \epsilon, & \text{when } 0 < \alpha_i < C \\ -\sum_{i=1}^{m} (\alpha_j' - \alpha_j) K(\mathbf{x}_j, \mathbf{x}_i) + y_i - \epsilon, & \text{when } 0 < \alpha_i' < C \end{cases}$$

Support vectors: points strictly outside the tube

45/48

A.Zaytsev, ML

Comments

- Advantages
 - strong theoretical guarantees (for that loss)
 - sparser solution
 - use of kernels
- Disadvantages
 - selection of two parameters: C and ϵ . Heuristics for that:
 - * search C near maximum y, ϵ near average difference of y-s, measure of no. of SVs
 - large matrices: low-rank approximations of kernel matrix

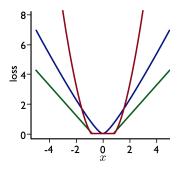
46/48

A.Zaytsev, ML Skoltech

Comments

- Advantages
 - strong theoretical guarantees (for that loss)
 - sparser solution
 - use of kernels
- Disadvantages
 - selection of two parameters: C and ϵ . Heuristics for that:
 - * search C near maximum y, ϵ near average difference of y-s, measure of no. of SVs
 - large matrices: low-rank approximations of kernel matrix

Alternative Loss Functions (similar formulations and results)



ullet quadratic ϵ -insensitive

$$x \to \max(0, |x| - \epsilon)^2$$

Huber

$$x
ightarrow egin{cases} x^2, & \text{if } |x| \leq c \ 2c|x|-c^2, & \text{otherwise} \end{cases}$$

ullet ϵ -insensitive

$$x \to \max(0, |x| - \epsilon)$$

47/48

A.Zaytsev, ML Skoltech

• SVR in case of quadratic ϵ -insensitive for $\epsilon=0$ coincides with Kernel Ridge Regression (see lecture 2)

$$h(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}_i, \mathbf{x}), \tag{*}$$

where

$$\boldsymbol{\alpha} = (\boldsymbol{\Phi}(\mathbf{X}) \cdot \boldsymbol{\Phi}(\mathbf{X})^{\top} + \lambda \mathbf{I})^{-1} \mathbf{Y} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{Y},$$

where

$$\mathbf{X} = {\mathbf{x}_1, \dots, \mathbf{x}_m} \in \mathbb{R}^{m \times d}, \, \mathbf{Y} = (y_1, \dots, y_m) \in \mathbb{R}^{m \times 1}$$

• In case of $\epsilon>0$ SVR allows to reduce a number of terms in (*) above thanks to the support vector concept: explicit solution vs. sparsity!

4ロ > 4回 > 4 き > 4 き > き り < ○</p>

48/48

A.Zaytsev, ML