CS189/CS289A
Introduction to Machine Learning
Lecture 13: A Little Convex Optimization and SVMs Revisited

Peter Bartlett

March 5, 2015

Outline

- Convex optimization ideas: primal, Lagrangian, dual.

Outline

- Convex optimization ideas: primal, Lagrangian, dual.
- Weak and strong duality

Outline

- Convex optimization ideas: primal, Lagrangian, dual.
- Weak and strong duality
- Complementary slackness

Outline

- Convex optimization ideas: primal, Lagrangian, dual.
- Weak and strong duality
- Complementary slackness
- Karush-Kuhn-Tucker optimality conditions

Outline

- Convex optimization ideas: primal, Lagrangian, dual.
- Weak and strong duality
- Complementary slackness
- Karush-Kuhn-Tucker optimality conditions
- SVMs

Outline

- Convex optimization ideas: primal, Lagrangian, dual.
- Weak and strong duality
- Complementary slackness
- Karush-Kuhn-Tucker optimality conditions
- SVMs
  - Complementary slackness and support vectors

Outline

- Convex optimization ideas: primal, Lagrangian, dual.
- Weak and strong duality
- Complementary slackness
- Karush-Kuhn-Tucker optimality conditions
- SVMs
  - Complementary slackness and support vectors
  - Dual problem and kernels

A brief detour into optimization

SVM optimization problems

A brief detour into optimization

SVM optimization problems

Hard margin SVM

$$\begin{aligned} \min_{\theta} \quad & \|\theta\|^2 \\ \text{such that} \quad & y^i \theta \cdot x^i \geq 1 \qquad (i = 1, \ldots, n) \end{aligned}$$

A brief detour into optimization

SVM optimization problems

Hard margin SVM

$$\begin{aligned} \min_{\theta} \quad & \|\theta\|^2 \\ \text{such that} \quad & y^i \theta \cdot x^i \geq 1 \qquad (i = 1, \ldots, n) \end{aligned}$$

Soft margin SVM

$$\min_{\theta} \quad \|\theta\|^2 + C \sum_{i=1}^{n} \left(1 - y^i \theta \cdot x^i\right)_+ .$$

A brief detour into optimization

SVM optimization problems: As quadratic programs

A brief detour into optimization

SVM optimization problems: As quadratic programs

Hard margin SVM

$$
\begin{aligned}
\min_{\theta} \quad & \|\theta\|^2 \\
\text{such that} \quad & y^i \theta \cdot x^i \geq 1 \qquad (i = 1, \ldots, n)
\end{aligned}
$$

A brief detour into optimization

SVM optimization problems: As quadratic programs

Hard margin SVM

$$\begin{aligned} \min_{\theta} \quad & \|\theta\|^2 \\ \text{such that} \quad & y^i \theta \cdot x^i \geq 1 \qquad (i = 1, \ldots, n) \end{aligned}$$

Soft margin SVM

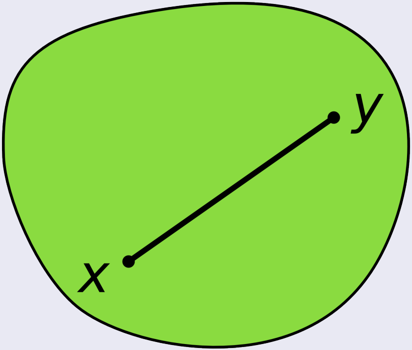
$$\begin{aligned} \min_{\theta, \xi} \quad & \|\theta\|^2 + C \sum_{i=1}^{n} \xi_i \\ \text{such that} \quad & \xi_i \geq 0, \\ & \xi_i \geq 1 - y^i \theta \cdot x^i. \end{aligned}$$

Convexity

Convex and non-convex sets

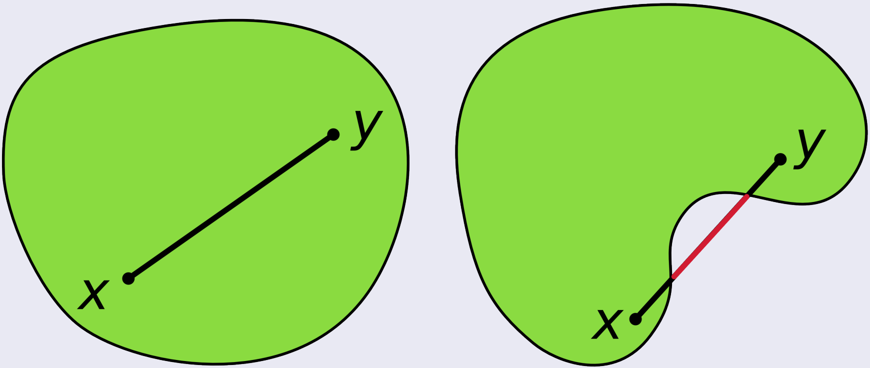
Convexity

Convex and non-convex sets



Convexity

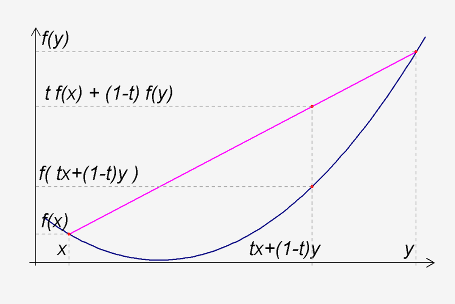
Convex and non-convex sets



Convexity

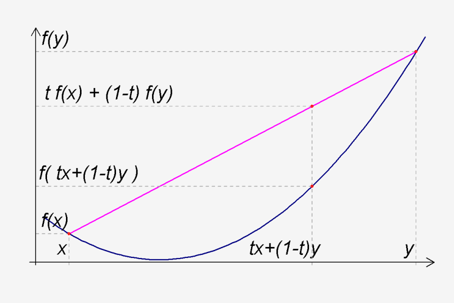
Convex and non-convex functions

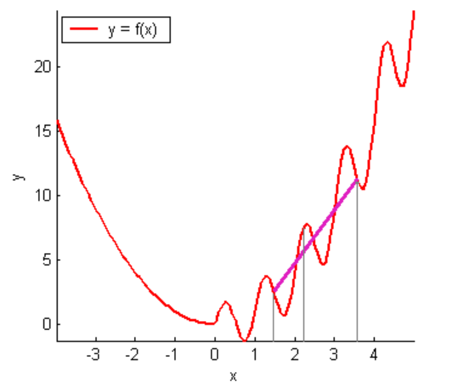
Convex and non-convex functions



Convexity

Convex and non-convex functions





A brief detour into optimization

SVMs: convex criterion, convex constraint set

A brief detour into optimization

SVMs: convex criterion, convex constraint set

Hard margin SVM

$$\begin{aligned} \min_{\theta} \quad & \|\theta\|^2 \\ \text{such that} \quad & y^i \theta \cdot x^i \geq 1 \qquad (i = 1, \ldots, n) \end{aligned}$$

A brief detour into optimization

SVMs: convex criterion, convex constraint set

Hard margin SVM

$$\begin{aligned} \min_{\theta} \quad & \|\theta\|^2 \\ \text{such that} \quad & y^i \theta \cdot x^i \geq 1 \qquad (i = 1, \ldots, n) \end{aligned}$$

Soft margin SVM

$$\begin{aligned} \min_{\theta, \xi} \quad & \|\theta\|^2 + C \sum_{i=1}^{n} \xi_i \\ \text{such that} \quad & \xi_i \geq 0, \\ & \xi_i \geq 1 - y^i \theta \cdot x^i. \end{aligned}$$

Primal, Lagrangian, dual

Consider the convex optimization problem

$$\begin{aligned} p^* = \min_{x \in \mathbb{R}^n} & \ f_0(x) \\ \text{s.t.} & \ f_i(x) \leq 0, \qquad i = 1, 2, \ldots, m. \end{aligned}$$

Primal, Lagrangian, dual

Consider the convex optimization problem

$$\begin{aligned} p^* = \min_{x \in \mathbb{R}^n} \ & f_0(x) \\ \text{s.t.} \ & f_i(x) \leq 0, \qquad i = 1, 2, \ldots, m. \end{aligned}$$

We can rewrite the constraints as penalties:

Primal, Lagrangian, dual

Consider the convex optimization problem

$$p^* = \min_{x \in \mathbb{R}^n} f_0(x)$$
$$\text{s.t. } f_i(x) \le 0, \qquad i = 1, 2, \ldots, m.$$

We can rewrite the constraints as penalties:

$$p^* = \min_{x \in \mathbb{R}^n} f_0(x) + \begin{cases} 0 & \text{if all } f_i(x) \le 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Primal, Lagrangian, dual

Consider the convex optimization problem

$$p^* = \min_{x \in \mathbb{R}^n} f_0(x) \quad \text{s.t. } f_i(x) \leq 0, \qquad i = 1, 2, \ldots, m.$$

We can rewrite the constraints as penalties:

$$p^* = \min_{x \in \mathbb{R}^n} f_0(x) + \begin{cases} 0 & \text{if all } f_i(x) \leq 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Then replace that constraint penalty with something smaller:

Primal, Lagrangian, dual

Consider the convex optimization problem

$$
\begin{aligned}
p^* = \min_{x \in \mathbb{R}^n} \ & f_0(x) \\
\text{s.t.} \ & f_i(x) \le 0, \qquad i = 1, 2, \ldots, m.
\end{aligned}
$$

We can rewrite the constraints as penalties:

$$
p^* = \min_{x \in \mathbb{R}^n} f_0(x) + \begin{cases} 0 & \text{if all } f_i(x) \le 0, \\ +\infty & \text{otherwise.} \end{cases}
$$

Then replace that constraint penalty with something smaller:
Introduce Lagrange multipliers (dual variables) $\lambda_1, \ldots, \lambda_m \ge 0$, and define the Lagrangian $L : \mathbb{R}^{n+m} \to \mathbb{R}$ as

$$
L(x, \lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x).
$$

Primal, Lagrangian, dual

The Lagrangian

$$L(x, \lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x).$$

Primal, Lagrangian, dual

The Lagrangian

$$L(x, \lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x).$$

- Think of the λ_i as the cost of violating the constraint $f_i(x) \leq 0$.

Joseph-Louis Lagrange

1736-1813

Primal, Lagrangian, dual

The Lagrangian

$$L(x,\lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x).$$

- Think of the λ_i as the cost of violating the constraint $f_i(x) \leq 0$.
- L defines a saddle point game: one player (MIN) chooses x to minimize L, the other player (MAX) chooses λ to maximize L. If MIN violates a constraint, $f_i(x) > 0$, then MAX can drive L to infinity.

Joseph-Louis Lagrange

1736-1813

Primal, Lagrangian, dual

- We call the original optimization the *primal* problem. It has value

$$p^* = \min_x \max_{\lambda \geq 0} L(x, \lambda).$$

Primal, Lagrangian, dual

- We call the original optimization the *primal* problem. It has value

$$p^* = \min_x \max_{\lambda \geq 0} L(x, \lambda).$$

(Because for an infeasible x, $L(x, \lambda)$ can be made infinite, and for a feasible x, the $\lambda_i f_i(x)$ terms will become zero.)

Primal, Lagrangian, dual

- We call the original optimization the *primal* problem. It has value

$$p^* = \min_x \max_{\lambda \geq 0} L(x, \lambda).$$

  (Because for an infeasible x, $L(x, \lambda)$ can be made infinite, and for a feasible x, the $\lambda_i f_i(x)$ terms will become zero.)

- Define $g(\lambda) := \min_x L(x, \lambda)$,

Primal, Lagrangian, dual

- We call the original optimization the *primal* problem. It has value

$$p^* = \min_x \max_{\lambda \geq 0} L(x, \lambda).$$

  (Because for an infeasible x, $L(x, \lambda)$ can be made infinite, and for a feasible x, the $\lambda_i f_i(x)$ terms will become zero.)

- Define $g(\lambda) := \min_x L(x, \lambda)$, and define the *dual* problem as

$$d^* = \max_{\lambda \geq 0} g(\lambda) = \max_{\lambda \geq 0} \min_x L(x, \lambda).$$

Primal, Lagrangian, dual

- We call the original optimization the *primal* problem. It has value

$$p^* = \min_x \max_{\lambda \geq 0} L(x, \lambda).$$

(Because for an infeasible x, $L(x, \lambda)$ can be made infinite, and for a feasible x, the $\lambda_i f_i(x)$ terms will become zero.)

- Define $g(\lambda) := \min_x L(x, \lambda)$, and define the *dual* problem as

$$d^* = \max_{\lambda \geq 0} g(\lambda) = \max_{\lambda \geq 0} \min_x L(x, \lambda).$$

- In a zero sum game, it's always better to play second:

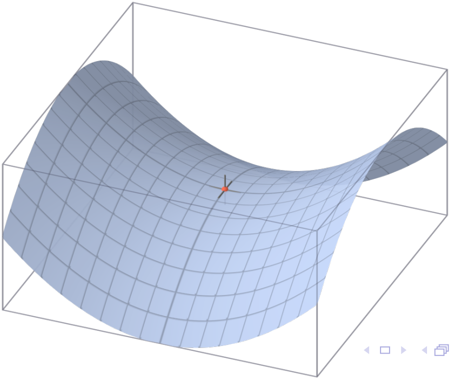
$$p^* = \min_x \max_{\lambda \geq 0} L(x, \lambda) \geq \max_{\lambda \geq 0} \min_x L(x, \lambda) = d^*.$$

This is called *weak duality*.

Strong duality

- If there is a *saddle point* (x^*, λ^*), so that for all x and $\lambda \geq 0$,

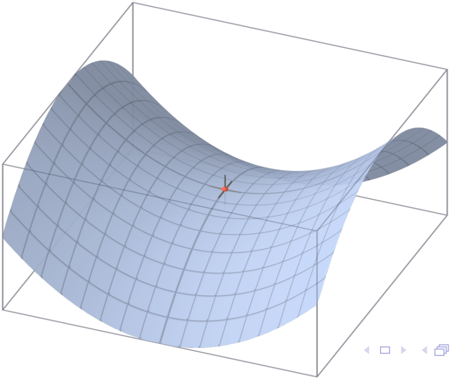
$$L(x^*, \lambda^*)$$



Strong duality

- If there is a *saddle point* (x^*, λ^*), so that for all x and $\lambda \geq 0$,

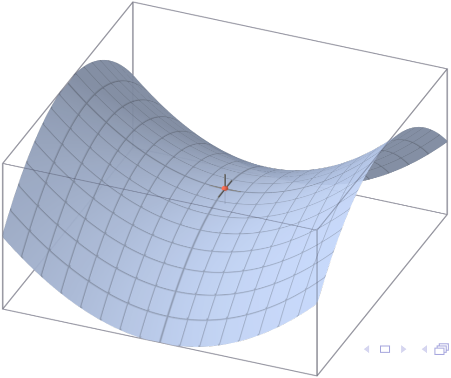
$$L(x^*, \lambda) \leq L(x^*, \lambda^*)$$



Strong duality

- If there is a *saddle point* (x^*, λ^*), so that for all x and $\lambda \geq 0$,

$$L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*),$$



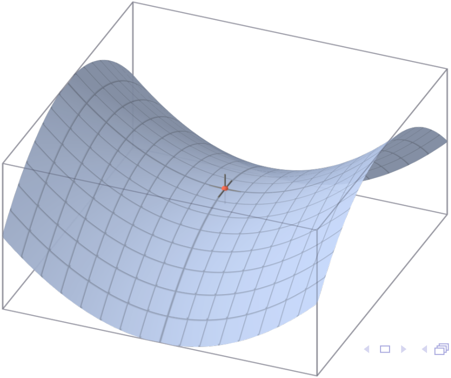
Strong duality

- If there is a *saddle point* (x^*, λ^*), so that for all x and $\lambda \geq 0$,

$$L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*),$$

then we have *strong duality*: the primal and dual have same value,

$$p^* = \min_x \max_{\lambda \geq 0} L(x, \lambda) = \max_{\lambda \geq 0} \min_x L(x, \lambda) = d^*.$$



Outline

- Optimization ideas: primal, Lagrangian, dual.
- Weak and strong duality
- **Complementary slackness**
- Karush-Kuhn-Tucker optimality conditions
- SVMs
  - Complementary slackness and support vectors
  - Dual problem and kernels

Complementary slackness

Suppose $p^* = d^*$. (This will be true for all of our examples.)

Complementary slackness

Suppose $p^* = d^*$. (This will be true for all of our examples.)

Complementary slackness

If $p^* = d^*$ and we have primal solution x^* and dual solution λ^*, then for the ith constraint ($f_i(x) \leq 0$),

$$\lambda_i^* f_i(x^*) = 0.$$

Complementary slackness

Suppose $p^* = d^*$. (This will be true for all of our examples.)

Complementary slackness

If $p^* = d^*$ and we have primal solution x^* and dual solution λ^*, then for the ith constraint ($f_i(x) \leq 0$),

$$\lambda_i^* f_i(x^*) = 0.$$

That is, if $f_i(x^*) < 0$ then $\lambda_i = 0$.
And if $\lambda_i > 0$ then $f_i(x^*) = 0$.

Complementary slackness

$$\lambda_i^* f_i(x^*) = 0.$$

Complementary slackness

$$\lambda_i^* f_i(x^*) = 0.$$

Why?

Complementary slackness

$$\lambda_i^* f_i(x^*) = 0.$$

Why?

$$f_0(x^*) = g(\lambda^*)$$

Complementary slackness

$$\lambda_i^* f_i(x^*) = 0.$$

Why?

$$f_0(x^*) = g(\lambda^*) = \min_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) \right)$$

Complementary slackness

$$\lambda_i^* f_i(x^*) = 0.$$

Why?

$$f_0(x^*) = g(\lambda^*) = \min_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) \right) \leq f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*).$$

Complementary slackness

$$\lambda_i^* f_i(x^*) = 0.$$

Why?

$$f_0(x^*) = g(\lambda^*) = \min_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) \right) \leq f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*).$$

That is,

$$\sum_{i=1}^m \lambda_i^* f_i(x^*) \geq 0.$$

Complementary slackness

$$\lambda_i^* f_i(x^*) = 0.$$

Why?

$$f_0(x^*) = g(\lambda^*) = \min_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) \right) \leq f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*).$$

That is,

$$\sum_{i=1}^m \lambda_i^* f_i(x^*) \geq 0.$$

But $\lambda_i^* \geq 0$ and $f_i(x^*) \leq 0$, so every term in the sum must be zero.

Optimality conditions

Without constraints, if f_0 is convex and differentiable, the optimal x satisfies $\nabla f_0(x) = 0$. With constraints?

Optimality conditions

Without constraints, if f_0 is convex and differentiable, the optimal x satisfies $\nabla f_0(x) = 0$. With constraints?

Karush-Kuhn-Tucker optimality conditions

Suppose f_0, f_i are convex and differentiable.

Optimality conditions

Without constraints, if f_0 is convex and differentiable, the optimal x satisfies $\nabla f_0(x) = 0$. With constraints?

Karush-Kuhn-Tucker optimality conditions

Suppose f_0, f_i are convex and differentiable.
Then x and λ are optimal

Optimality conditions

Without constraints, if f_0 is convex and differentiable, the optimal x satisfies $\nabla f_0(x) = 0$. With constraints?

Karush-Kuhn-Tucker optimality conditions

Suppose f_0, f_i are convex and differentiable.
Then x and λ are optimal (and $f_0(x) = p^* = d^* = g(\lambda)$)

Optimality conditions

Without constraints, if f_0 is convex and differentiable, the optimal x satisfies $\nabla f_0(x) = 0$. With constraints?

Karush-Kuhn-Tucker optimality conditions

Suppose f_0, f_i are convex and differentiable.
Then x and λ are optimal (and $f_0(x) = p^* = d^* = g(\lambda)$) if and only if

Optimality conditions

Without constraints, if f_0 is convex and differentiable, the optimal x satisfies $\nabla f_0(x) = 0$. With constraints?

Karush-Kuhn-Tucker optimality conditions

Suppose f_0, f_i are convex and differentiable.
Then x and λ are optimal (and $f_0(x) = p^* = d^* = g(\lambda)$) if and only if

1. Primal feasibility: $f_i(x) \leq 0$.

Optimality conditions

Without constraints, if f_0 is convex and differentiable, the optimal x satisfies $\nabla f_0(x) = 0$. With constraints?

Karush-Kuhn-Tucker optimality conditions

Suppose f_0, f_i are convex and differentiable.
Then x and λ are optimal (and $f_0(x) = p^* = d^* = g(\lambda)$) if and only if

1. Primal feasibility: $f_i(x) \leq 0$.
2. Dual feasibility: $\lambda_i \geq 0$.

Optimality conditions

Without constraints, if f_0 is convex and differentiable, the optimal x satisfies $\nabla f_0(x) = 0$. With constraints?

Karush-Kuhn-Tucker optimality conditions

Suppose f_0, f_i are convex and differentiable.
Then x and λ are optimal (and $f_0(x) = p^* = d^* = g(\lambda)$) if and only if

1. Primal feasibility: $f_i(x) \leq 0$.
2. Dual feasibility: $\lambda_i \geq 0$.
3. Complementary slackness: $\lambda_i f_i(x) = 0$.

Optimality conditions

Without constraints, if f_0 is convex and differentiable, the optimal x satisfies $\nabla f_0(x) = 0$. With constraints?

Karush-Kuhn-Tucker optimality conditions

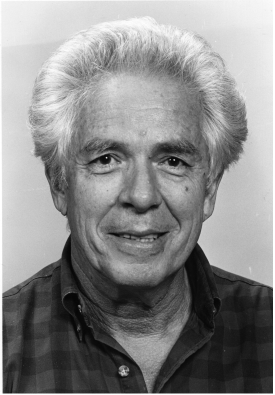
Suppose f_0, f_i are convex and differentiable.
Then x and λ are optimal (and $f_0(x) = p^* = d^* = g(\lambda)$) if and only if

1. Primal feasibility: $f_i(x) \leq 0$.
2. Dual feasibility: $\lambda_i \geq 0$.
3. Complementary slackness: $\lambda_i f_i(x) = 0$.
4. Stationarity: $\nabla f_0(x) + \sum_i \lambda_i \nabla f_i(x) = 0$.

Optimality conditions

William Karush

1917-1997



Harold Kuhn

1925-2014

Albert W. Tucker

1905-1995

Outline

- Optimization ideas: primal, Lagrangian, dual.
- Weak and strong duality
- Complementary slackness
- Karush-Kuhn-Tucker optimality conditions
- **SVMs**
  - Complementary slackness and support vectors
  - Dual problem and kernels

Support vector machines

Hard margin SVM

$$\begin{aligned} \min_{\theta} \quad & \|\theta\|^2 \\ \text{such that} \quad & y^i \theta \cdot x^i \geq 1 \qquad (i = 1, \ldots, n) \end{aligned}$$

Support vector machines

Hard margin SVM

$$\begin{aligned} \min_{\theta} \quad & \|\theta\|^2 \\ \text{such that} \quad & y^i \theta \cdot x^i \geq 1 \qquad (i = 1, \ldots, n) \end{aligned}$$

$$L(\theta, \alpha) = \frac{1}{2}\|\theta\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i \theta' x_i)$$

Support vector machines

Hard margin SVM

$$\begin{aligned} \min_\theta \quad & \|\theta\|^2 \\ \text{such that} \quad & y^i \theta \cdot x^i \geq 1 \qquad (i = 1, \ldots, n) \end{aligned}$$

$$L(\theta, \alpha) = \frac{1}{2}\|\theta\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i \theta' x_i)$$

$$g(\alpha) = \min_\theta L(\theta, \alpha)$$

Support vector machines

Hard margin SVM

$$\begin{aligned} \min_{\theta} \quad & \|\theta\|^2 \\ \text{such that} \quad & y^i \theta \cdot x^i \geq 1 \qquad (i = 1, \dots, n) \end{aligned}$$

$$L(\theta, \alpha) = \frac{1}{2}\|\theta\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i \theta' x_i)$$

$$g(\alpha) = \min_{\theta} L(\theta, \alpha)$$

setting $$\theta^* = \sum_{i=1}^{n} \alpha_i y_i x_i,$$

Hard margin SVM

$$\min_{\theta} \quad \|\theta\|^2$$
$$\text{such that} \quad y^i \theta \cdot x^i \geq 1 \qquad (i = 1, \ldots, n)$$

$$L(\theta, \alpha) = \frac{1}{2}\|\theta\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i \theta' x_i)$$
$$g(\alpha) = \min_{\theta} L(\theta, \alpha)$$
$$\text{setting} \qquad \theta^* = \sum_{i=1}^{n} \alpha_i y_i x_i,$$
$$g(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i' x_j.$$

Support vector machines

It turns out that, if there is a feasible θ (that is, the data are separable), we have strong duality.

Support vector machines

It turns out that, if there is a feasible θ (that is, the data are separable), we have strong duality.
We can express the optimal θ^* in terms of the solution, α^*, to the dual problem:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i' x_j \\ \text{s.t.} \quad & \alpha_i \geq 0, \qquad i = 1, 2, \ldots, n. \end{aligned}$$

Support vector machines

Complementary slackness demonstrates the role of the α_i

Support vector machines

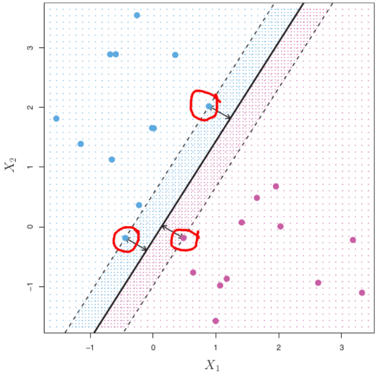
Complementary slackness demonstrates the role of the α_i:

$$\alpha_i > 0 \text{ implies } y_i \theta^{*\prime} x_i = 1,$$

Support vector machines

Complementary slackness demonstrates the role of the α_i:

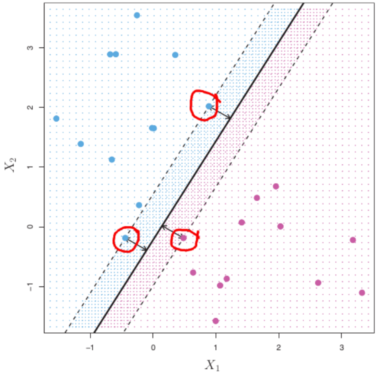
$$\alpha_i > 0 \text{ implies } y_i \theta^{*\prime} x_i = 1,$$



Support vector machines

Complementary slackness demonstrates the role of the α_i:

$$\alpha_i > 0 \text{ implies } y_i\theta^{*\prime}x_i = 1,$$
$$y_i\theta' x_i > 1 \text{ implies } \alpha_i = 0.$$

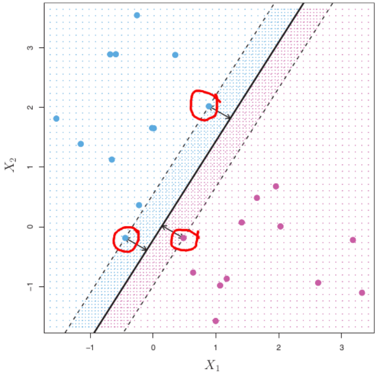


Support vector machines

Complementary slackness demonstrates the role of the α_i:

$$\alpha_i > 0 \text{ implies } y_i \theta^{*\prime} x_i = 1,$$
$$y_i \theta' x_i > 1 \text{ implies } \alpha_i = 0.$$

That is, only the points for which the constraints are tight (support vectors) appear in the sum defining θ^*.



Support vector machines

And we can express the solution in terms of an arbitrary kernel k:

Support vector machines

And we can express the solution in terms of an arbitrary kernel k:

$$\hat{y}(x) = \operatorname{sign}\left(\langle\theta, \Phi(x)\rangle\right)$$

Support vector machines

And we can express the solution in terms of an arbitrary kernel k:

$$\begin{aligned}\hat{y}(x) &= \operatorname{sign}\left(\langle\theta, \Phi(x)\rangle\right) \\ &= \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i \left\langle \Phi(x_i), \Phi(x) \right\rangle\right)\end{aligned}$$

Support vector machines

And we can express the solution in terms of an arbitrary kernel k:

$$\begin{aligned}
\hat{y}(x) &= \operatorname{sign}\left(\langle \theta, \Phi(x) \rangle\right) \\
&= \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i \left\langle \Phi(x_i), \Phi(x) \right\rangle\right) \\
&= \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i k(x_i, x)\right),
\end{aligned}$$

Support vector machines

And we can express the solution in terms of an arbitrary kernel k:

$$\begin{aligned}\hat{y}(x) &= \operatorname{sign}\left(\langle\theta, \Phi(x)\rangle\right)\\ &= \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i \left\langle \Phi(x_i), \Phi(x)\right\rangle\right)\\ &= \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i k(x_i, x)\right),\end{aligned}$$

where α solves the dual problem

Support vector machines

And we can express the solution in terms of an arbitrary kernel k:

$$\begin{aligned}
\hat{y}(x) &= \operatorname{sign}\left(\langle \theta, \Phi(x) \rangle\right) \\
&= \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i \left\langle \Phi(x_i), \Phi(x) \right\rangle\right) \\
&= \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i k(x_i, x)\right),
\end{aligned}$$

where α solves the dual problem

$$\begin{aligned}
\min_{\alpha} \quad & \frac{1}{2}\alpha^T \operatorname{diag}(y) K \operatorname{diag}(y) \alpha - \alpha^T 1 \\
\text{s.t.} \quad & \alpha \geq 0.
\end{aligned}$$

Soft Margin Support Vector Machine

Soft Margin Support Vector Machine

Soft margin SVM

$$\min_{\theta} \quad \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\left(1 - y^i\theta \cdot x^i\right)_+ .$$

Soft Margin Support Vector Machine

Soft margin SVM

$$\min_{\theta} \quad \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\left(1 - y^i\theta \cdot x^i\right)_+ .$$

As a QP

$$\min_{\theta,\xi} \quad \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i$$

Soft Margin Support Vector Machine

Soft margin SVM

$$\min_{\theta} \qquad \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\left(1 - y^i\theta \cdot x^i\right)_+ .$$

As a QP

$$\begin{aligned} \min_{\theta,\xi} \qquad & \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\ \text{such that} \qquad & \xi_i \geq 0, \end{aligned}$$

Soft Margin Support Vector Machine

Soft margin SVM

$$\min_{\theta} \quad \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\left(1 - y^i\theta \cdot x^i\right)_+ .$$

As a QP

$$\begin{aligned} \min_{\theta,\xi} \quad & \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\ \text{such that} \quad & \xi_i \geq 0, \\ & \xi_i \geq 1 - y^i\theta \cdot x^i. \end{aligned}$$

Soft Margin Support Vector Machine

Soft margin SVM

$$\min_{\theta} \quad \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\left(1 - y^i\theta\cdot x^i\right)_+ .$$

As a QP

$$\begin{aligned} \min_{\theta,\xi} \quad & \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\ \text{such that} \quad & \xi_i \geq 0, \\ & \xi_i \geq 1 - y^i\theta\cdot x^i. \end{aligned}$$

The optimal slack variables ξ_i satisfy $\xi_i = \left(1 - y_i\theta^T x_i\right)_+$.

Soft Margin Support Vector Machine

$L(\theta, \xi, \alpha, \lambda)$

Soft Margin Support Vector Machine

$$L(\theta, \xi, \alpha, \lambda) = \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i +$$

Soft Margin Support Vector Machine

$$L(\theta, \xi, \alpha, \lambda) = \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1 - y_i\theta^T x_i - \xi_i) -$$

Soft Margin Support Vector Machine

$$L(\theta, \xi, \alpha, \lambda) = \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1 - y_i\theta^T x_i - \xi_i) - \sum_{i=1}^{n}\lambda_i\xi_i.$$

Soft Margin Support Vector Machine

$$L(\theta, \xi, \alpha, \lambda) = \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1 - y_i\theta^T x_i - \xi_i) - \sum_{i=1}^{n}\lambda_i\xi_i.$$

Minimizing over θ, ξ (set derivatives to zero) gives

Soft Margin Support Vector Machine

$$L(\theta, \xi, \alpha, \lambda) = \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1 - y_i\theta^T x_i - \xi_i) - \sum_{i=1}^{n}\lambda_i\xi_i.$$

Minimizing over θ, ξ (set derivatives to zero) gives

$$\theta = \sum_i \alpha_i y_i x_i,$$

Soft Margin Support Vector Machine

$$L(\theta, \xi, \alpha, \lambda) = \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1 - y_i\theta^T x_i - \xi_i) - \sum_{i=1}^{n}\lambda_i\xi_i.$$

Minimizing over θ, ξ (set derivatives to zero) gives

$$\theta = \sum_i \alpha_i y_i x_i,$$

$$\frac{C}{n} = \alpha_i + \lambda_i,$$

Soft Margin Support Vector Machine

$$L(\theta, \xi, \alpha, \lambda) = \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1 - y_i\theta^T x_i - \xi_i) - \sum_{i=1}^{n}\lambda_i\xi_i.$$

Minimizing over θ, ξ (set derivatives to zero) gives

$$\theta = \sum_i \alpha_i y_i x_i,$$

$$\frac{C}{n} = \alpha_i + \lambda_i,$$

so

$$g(\alpha, \lambda) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i^T x_j.$$

Soft Margin Support Vector Machine

The dual problem is:

$$\begin{aligned} \max_{\alpha,\lambda} \quad & \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ \text{s.t.} \quad & \alpha_i \geq 0 \end{aligned}$$

Soft Margin Support Vector Machine

The dual problem is:

$$\begin{aligned} \max_{\alpha,\lambda} \quad & \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ \text{s.t.} \quad & \alpha_i \geq 0 \\ & \lambda_i \geq 0 \end{aligned}$$

Soft Margin Support Vector Machine

The dual problem is:

$$\begin{aligned}
\max_{\alpha,\lambda} \quad & \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \\
\text{s.t.} \quad & \alpha_i \geq 0 \\
& \lambda_i \geq 0 \\
& \alpha_i + \lambda_i = \frac{C}{n}.
\end{aligned}$$

Soft Margin Support Vector Machine

The dual problem is:

$$\begin{aligned}
\max_{\alpha,\lambda} \quad & \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j x_i^T x_j \\
\text{s.t.} \quad & \alpha_i \geq 0 \\
& \lambda_i \geq 0 \\
& \alpha_i + \lambda_i = \frac{C}{n}.
\end{aligned}$$

Eliminating the λ_i:

Dual problem

Soft Margin Support Vector Machine

The dual problem is:

$$\begin{aligned} \max_{\alpha,\lambda} \quad & \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ \text{s.t.} \quad & \alpha_i \geq 0 \\ & \lambda_i \geq 0 \\ & \alpha_i + \lambda_i = \frac{C}{n}. \end{aligned}$$

Eliminating the λ_i:

Dual problem

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \alpha^T \mathrm{diag}(y) K \mathrm{diag}(y) \alpha - \alpha^T 1 \\ \text{s.t.} \quad & \end{aligned}$$

Soft Margin Support Vector Machine

The dual problem is:

$$
\begin{aligned}
\max_{\alpha,\lambda} \quad & \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j x_i^T x_j \\
\text{s.t.} \quad & \alpha_i \geq 0 \\
& \lambda_i \geq 0 \\
& \alpha_i + \lambda_i = \frac{C}{n}.
\end{aligned}
$$

Eliminating the λ_i:

Dual problem

$$
\begin{aligned}
\min_{\alpha} \quad & \frac{1}{2}\alpha^T \mathrm{diag}(y) K \mathrm{diag}(y)\alpha - \alpha^T 1 \\
\text{s.t.} \quad & 0 \leq \alpha_i \leq \frac{C}{n}.
\end{aligned}
$$

Support vectors

Note: the only change in going from the hard margin to the soft margin is the addition of the upper bound on the α_i.

Support vectors

Note: the only change in going from the hard margin to the soft margin is the addition of the upper bound on the α_i.
Consider the consequences of complementary slackness:

Support vectors

Note: the only change in going from the hard margin to the soft margin is the addition of the upper bound on the α_i.
Consider the consequences of complementary slackness:

$$1 - y_i x_i^T \theta^* - \xi_i^*$$

Support vectors

Note: the only change in going from the hard margin to the soft margin is the addition of the upper bound on the α_i.

Consider the consequences of complementary slackness:

$$\alpha_i^* \left(1 - y_i x_i^T \theta^* - \xi_i^*\right) = 0.$$

Support vectors

Note: the only change in going from the hard margin to the soft margin is the addition of the upper bound on the α_i.
Consider the consequences of complementary slackness:

$$\alpha_i^* \left(1 - y_i x_i^T \theta^* - \xi_i^*\right) = 0.$$

$$\xi_i^*$$

Support vectors

Note: the only change in going from the hard margin to the soft margin is the addition of the upper bound on the α_i.

Consider the consequences of complementary slackness:

$$\begin{aligned} \alpha_i^* \left(1 - y_i x_i^T \theta^* - \xi_i^*\right) &= 0. \\ \lambda_i^* \xi_i^* &= 0. \end{aligned}$$

Support vectors

Note: the only change in going from the hard margin to the soft margin is the addition of the upper bound on the α_i.
Consider the consequences of complementary slackness:

$$\alpha_i^* \left(1 - y_i x_i^T \theta^* - \xi_i^*\right) = 0.$$
$$\lambda_i^* \xi_i^* = 0.$$

1. $\alpha_i^* > 0$ implies $y_i x_i^T \theta^* = 1 - \xi_i^*$

Support vectors

Note: the only change in going from the hard margin to the soft margin is the addition of the upper bound on the α_i.
Consider the consequences of complementary slackness:

$$\begin{aligned} \alpha_i^* \left(1 - y_i x_i^T \theta^* - \xi_i^*\right) &= 0. \\ \lambda_i^* \xi_i^* &= 0. \end{aligned}$$

1. $\alpha_i^* > 0$ implies $y_i x_i^T \theta^* = 1 - \xi_i^* \leq 1$

Support vectors

Note: the only change in going from the hard margin to the soft margin is the addition of the upper bound on the α_i.

Consider the consequences of complementary slackness:

$$\alpha_i^* \left(1 - y_i x_i^T \theta^* - \xi_i^*\right) = 0.$$

$$\lambda_i^* \xi_i^* = 0.$$

1. $\alpha_i^* > 0$ implies $y_i x_i^T \theta^* = 1 - \xi_i^* \leq 1$. That is, the 'support vectors' ($y_i x_i$ with $\alpha_i^* > 0$) are in the wrong halfspace $\{x : x^T \theta^* \leq 1\}$.

Support vectors

Note: the only change in going from the hard margin to the soft margin is the addition of the upper bound on the α_i.
Consider the consequences of complementary slackness:

$$\begin{aligned} \alpha_i^* \left(1 - y_i x_i^T \theta^* - \xi_i^*\right) &= 0. \\ \lambda_i^* \xi_i^* &= 0. \end{aligned}$$

1. $\alpha_i^* > 0$ implies $y_i x_i^T \theta^* = 1 - \xi_i^* \leq 1$. That is, the 'support vectors' ($y_i x_i$ with $\alpha_i^* > 0$) are in the wrong halfspace $\{x : x^T \theta^* \leq 1\}$.
2. If $y_i x_i^T \theta^* < 1$, $\xi_i^* > 0$,

Support vectors

Note: the only change in going from the hard margin to the soft margin is the addition of the upper bound on the α_i.
Consider the consequences of complementary slackness:

$$\alpha_i^* \left(1 - y_i x_i^T \theta^* - \xi_i^*\right) = 0.$$
$$\lambda_i^* \xi_i^* = 0.$$

1. $\alpha_i^* > 0$ implies $y_i x_i^T \theta^* = 1 - \xi_i^* \leq 1$. That is, the 'support vectors' ($y_i x_i$ with $\alpha_i^* > 0$) are in the wrong halfspace $\{x : x^T \theta^* \leq 1\}$.
2. If $y_i x_i^T \theta^* < 1$, $\xi_i^* > 0$, so $\lambda_i^* = 0$,

Support vectors

Note: the only change in going from the hard margin to the soft margin is the addition of the upper bound on the α_i.
Consider the consequences of complementary slackness:

$$\alpha_i^* \left(1 - y_i x_i^T \theta^* - \xi_i^*\right) = 0.$$
$$\lambda_i^* \xi_i^* = 0.$$

1. $\alpha_i^* > 0$ implies $y_i x_i^T \theta^* = 1 - \xi_i^* \leq 1$. That is, the 'support vectors' ($y_i x_i$ with $\alpha_i^* > 0$) are in the wrong halfspace $\{x : x^T \theta^* \leq 1\}$.
2. If $y_i x_i^T \theta^* < 1$, $\xi_i^* > 0$, so $\lambda_i^* = 0$, and $\alpha_i^* = C/n$.

Support vectors

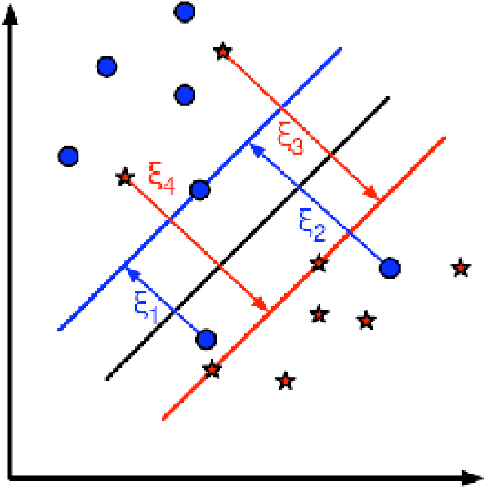
Note: the only change in going from the hard margin to the soft margin is the addition of the upper bound on the α_i.
Consider the consequences of complementary slackness:

$$\alpha_i^* \left(1 - y_i x_i^T \theta^* - \xi_i^*\right) = 0.$$

$$\lambda_i^* \xi_i^* = 0.$$

1. $\alpha_i^* > 0$ implies $y_i x_i^T \theta^* = 1 - \xi_i^* \leq 1$. That is, the 'support vectors' ($y_i x_i$ with $\alpha_i^* > 0$) are in the wrong halfspace $\{x : x^T \theta^* \leq 1\}$.
2. If $y_i x_i^T \theta^* < 1$, $\xi_i^* > 0$, so $\lambda_i^* = 0$, and $\alpha_i^* = C/n$. That is, the support vectors in the open halfspace $\{x : x^T \theta^* < 1\}$ have $\alpha_i^* = C/n$.

Support vectors



Role of C

Role of C

- In the primal, increasing C penalizes errors more (and puts less emphasis on minimizing $\|\theta\|$, that is, maximizing the margin).

Role of C

- In the primal, increasing C penalizes errors more (and puts less emphasis on minimizing $\|\theta\|$, that is, maximizing the margin).
- In the dual, decreasing C forces the α_is to be small. So the solution is not strongly influenced by a single outlier.

Outline

- Optimization ideas: primal, Lagrangian, dual.
- Weak and strong duality
- Complementary slackness
- Karush-Kuhn-Tucker optimality conditions
- SVMs
    - Complementary slackness and support vectors
    - Dual problem and kernels