Dual SVM and Kernels

Machine Learning
CS5824/ECE5424
Bert Huang
Virginia Tech

Outline

- Review soft-margin SVM
- Primals and duals
- Dual SVM and derivation
- The kernel trick
- Popular kernels: polynomial, Gaussian radial basis function (RBF)

Soft-Margin Primal SVM

slack penalty

$$f(w^*) = \min_{\substack{w \in \mathbb{R}^d \\ \xi \geq 0}} \frac{1}{2} w^\top w + C \sum_{i=1}^{n} \xi_i$$

$$\text{s.t.} \quad y_i(w^\top x_i + b) \geq 1 - \xi_i \qquad \forall i \in \{1, \ldots, n\}$$

slack variables

for hard margin: C ← ∞

Duality

- Optimization problems can be viewed from two (or more) perspectives
  - primal problem vs. dual problem
- Solving the dual tells us about the solution to the primal

Lagrangian (KKT) Dual for SVM

Karush-Kuhn-Tucker

Primal SVM

$$\min_{\substack{w \in \mathbb{R}^d \\ \xi \in [0,\infty]^n}} \quad \frac{1}{2} w^\top w + C \sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad y_i(w^\top x_i + b) - 1 + \xi_i \geq 0 \;\; \forall i \in \{1, \ldots, n\}$$

Dual SVM

$$\min_{\alpha} \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^\top x_j - \sum_i \alpha_i$$

$$\text{s.t.} \quad \sum_i \alpha_i y_i = 0, \quad \alpha_i \in [0, C]$$

$$w = \sum_i \alpha_i y_i x_i \qquad\qquad b = y_i - \sum_j \alpha_j y_j x_j^\top x_i$$

for examples i where $0 < \alpha_i < C$

$$\min_{\substack{w \in \mathbb{R}^d \\ \xi \in [0,\infty]^n}} \quad \frac{1}{2} w^\top w + C \sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad y_i(w^\top x_i + b) - 1 + \xi_i \geq 0 \quad \forall i \in \{1, \ldots, n\}$$

$$\min_{\substack{w \in \mathbb{R}^d \\ \xi \in \mathbb{R}^n}} \quad \max_{\substack{\alpha \in [0,\infty]^n \\ \beta \in [0,\infty]^n}} \quad L(w, b, \xi, \alpha, \beta)$$

primal problem

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2} w^\top w + C \sum_i \xi_i$$

$$- \sum_i \alpha_i (y_i(w^\top x_i + b) - 1 + \xi_i) - \sum_i \beta_i \xi_i$$

$-\alpha(-1)$ $\quad -\alpha(+1)$

$\uparrow$ $\quad \uparrow$ $\quad \uparrow\uparrow$

$$\min_{\substack{w \in \mathbb{R}^d \\ \xi \in \mathbb{R}^n}} \max_{\substack{\alpha \in [0,\infty]^n \\ \beta \in [0,\infty]^n}} L(w, b, \xi, \alpha, \beta)$$

primal problem

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2} w^\top w + C \sum_i \xi_i - \sum_i \alpha_i (y_i (w^\top x_i + b) - 1 + \xi_i) - \sum_i \beta_i \xi_i$$

$$\min_{\substack{w \in \mathbb{R}^d \\ \xi \in \mathbb{R}^n}} \max_{\substack{\alpha \in [0, \infty]^n \\ \beta \in [0, \infty]^n}} L(w, b, \xi, \alpha, \beta)$$

primal problem

$$\max_{\substack{\alpha \in [0, \infty]^n \\ \beta \in [0, \infty]^n}} \min_{\substack{w \in \mathbb{R}^d \\ \xi \in \mathbb{R}^n}} L(w, b, \xi, \alpha, \beta)$$

dual problem

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2} w^\top w + C \sum_i \xi_i - \sum_i \alpha_i (y_i (w^\top x_i + b) - 1 + \xi_i) - \sum_i \beta_i \xi_i$$

Karush-Kuhn-Tucker Conditions

- At the solution, we will provably have…
- **Stationarity**: gradients for primal and dual variables will be zero
- **Primal feasibility**: constraints on primal constraints will be satisfied
- **Dual feasibility**: constraints on dual variables will be satisfied
- **Complementary slackness**: for all inequality constraints, either the KKT multiplier will be zero or the constraint will be at equality (tight)

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2} w^\top w + C \sum_i \xi_i$$

$$- \sum_i \alpha_i (y_i (w^\top x_i + b) - 1 + \xi_i) - \sum_i \beta_i \xi_i$$

Gradients

$$\nabla_w L = w - \sum_i \alpha_i y_i x_i = 0$$

$$\nabla_b L = - \sum_i \alpha_i y_i = 0$$

$$\nabla_\xi L = C - \alpha - \beta = 0$$

Consequences

$$w = \sum_i \alpha_i y_i x_i$$

$$\sum_i \alpha_i y_i = 0$$

$$\alpha = C - \beta$$

$$\beta = C - \alpha$$

$$w = \sum_i \alpha_i y_i x_i$$

$$\sum_i \alpha_i y_i = 0$$

$$\alpha = C - \beta$$

$$\beta = C - \alpha$$

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2} w^\top w + C \sum_i \xi_i$$

$$- \sum_i \alpha_i (y_i (w^\top x_i + b) - 1 + \xi_i) - \sum_i \beta_i \xi_i$$

$$= \frac{1}{2} w^\top w + C \sum_i \xi_i - w^\top \sum_i \alpha_i y_i x_i$$

$$- b \sum_i \alpha_i y_i + \sum_i \alpha_i - \sum_i \alpha_i \xi_i - \sum_i \beta_i \xi_i$$

$$= \frac{1}{2} w^\top \sum_i \alpha_i y_i x_i + C \sum_i \xi_i - w^\top \sum_i \alpha_i y_i x_i$$

$$+ \sum_i \alpha_i - \sum_i \alpha_i \xi_i - \sum_i \beta_i \xi_i$$

$$w = \sum_i \alpha_i y_i x_i$$

$$\sum_i \alpha_i y_i = 0$$

$$\alpha = C - \beta$$

$$\beta = C - \alpha$$

$$= \frac{1}{2} w^\top \sum_i \alpha_i y_i x_i + C \sum_i \xi_i - w^\top \sum_i \alpha_i y_i x_i$$

$$+ \sum_i \alpha_i - \sum_i \alpha_i \xi_i - \sum_i \beta_i \xi_i$$

$$= -\frac{1}{2} w^\top \left(\sum_i \alpha_i y_i x_i \right) + C \sum_i \xi_i$$

$$+ \sum_i \alpha_i - \sum_i \alpha_i \xi_i - \sum_i (C - \alpha_i) \xi_i$$

$$= -\frac{1}{2} w^\top \left(\sum_i \alpha_i y_i x_i \right) + \sum_i \alpha_i$$

$$w = \sum_i \alpha_i y_i x_i$$

$$\sum_i \alpha_i y_i = 0$$

$$\alpha = C - \beta$$

$$\beta = C - \alpha$$

$$= -\frac{1}{2} w^\top \left(\sum_i \alpha_i y_i x_i \right) + \sum_i \alpha_i$$

$$= -\frac{1}{2} \left(\sum_i \alpha_i y_i x_i \right)^\top \left(\sum_j \alpha_j y_j x_j \right) + \sum_i \alpha_i$$

$$= -\frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^\top x_j + \sum_i \alpha_i$$

$$\max_{\alpha \geq 0} \; -\frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^\top x_j + \sum_i \alpha_i$$

Done? Not quite

$$w = \sum_i \alpha_i y_i x_i$$

$$\sum_i \alpha_i y_i = 0$$

$$\alpha = C - \beta$$

$$\beta = C - \alpha$$

$$\max_{\alpha \geq 0} \; -\frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^\top x_j + \sum_i \alpha_i$$

$$\min_{\alpha} \; \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^\top x_j - \sum_i \alpha_i$$

$$\text{s.t.} \quad \sum_i \alpha_i y_i = 0, \quad \alpha_i \in [0, C]$$

$$w = \sum_i \alpha_i y_i x_i \qquad b = y_i - \sum_j \alpha_j y_j x_j^\top x_i$$

complementary slackness

$$y_i(x_i^\top \sum_j \alpha_j y_j x_j + b) - 1 = 0$$ for examples i where $0 < \alpha_i < C$

Primal SVM

$$\min_{\substack{w \in \mathbb{R}^d \\ \xi \in [0, \infty]^n}} \quad \frac{1}{2} w^\top w + C \sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad y_i(w^\top x_i + b) - 1 + \xi_i \geq 0 \;\; \forall i \in \{1, \ldots, n\}$$

Dual SVM

$$\min_{\alpha} \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^\top x_j - \sum_i \alpha_i$$

$$\text{s.t.} \quad \sum_i \alpha_i y_i = 0, \quad \alpha_i \in [0, C]$$

$$w = \sum_i \alpha_i y_i x_i \qquad\qquad b = y_i - \sum_j \alpha_j y_j x_j^\top x_i$$

for examples i where $0 < \alpha_i < C$

Dual SVM

$$\min_{\alpha} \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^\top x_j - \sum_i \alpha_i$$

$$\text{s.t.} \quad \sum_i \alpha_i y_i = 0, \quad \alpha_i \in [0, C]$$

$$\sout{w = \sum_i \alpha_i y_i x_i}$$

$$b = y_i - \sum_j \alpha_j y_j x_j^\top x_i$$

for examples i where

$$0 < \alpha_i < C$$

$$f(x) = w^\top x + b = \sum_i \alpha_i y_i x_i^\top x + b$$

Kernel SVM

$$\min_{\alpha} \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_i \alpha_i$$

$$\text{s.t.} \quad \sum_i \alpha_i y_i = 0, \quad \alpha_i \in [0, C]$$

$$f(x) = w^\top x + b = \sum_i \alpha_i y_i K(x_i, x) + b$$

$$b = y_i - \sum_j \alpha_j y_j K(x_i, x_j)$$

for examples *i* where $0 < \alpha_i < C$

K = kernel function

Kernels

$$K(x_i, x_j) := \Phi(x_i)^\top \Phi(x_j) \qquad \Phi : \mathcal{X} \to \mathcal{Z}$$

$$\mathcal{X} = \mathbb{R}^d$$

$$\Phi(x) = [x^1, x^2, x^3, \dots, x^d]^\top \qquad \mathcal{Z} = \mathbb{R}^d$$

$$\Phi(x) = [x^1, \dots, x^d, x^1 x^1, \dots, x^1 x^d, \dots, x^d x^1, \dots, x^d x^d]^\top \qquad \mathcal{Z} = \mathbb{R}^{d^2}$$

Linear feature map

$$\Phi(x) = [x^1, x^2, x^3, \dots, x^d]^\top \qquad \mathcal{Z} = \mathbb{R}^d$$

Quadratic feature map

$$\Phi(x) = [x^1, \dots, x^d, x^1x^1, \dots, x^1x^d, \dots, x^dx^1, \dots, x^dx^d]^\top \qquad \mathcal{Z} = \mathbb{R}^{d^2}$$

Gaussian radial-basis (RBF) feature map $\qquad \mathcal{Z} = \mathbb{R}^\infty$

(Something weird. See in a few slides.)

Gram Matrices

$$\mathbf{K}_{ij} = K(x_i, x_j)$$

$$\mathbf{K} = \begin{bmatrix} \Phi(x_1)^\top \Phi(x_1), & \Phi(x_1)^\top \Phi(x_2), & \ldots, & \Phi(x_1)^\top \Phi(x_n) \\ \Phi(x_2)^\top \Phi(x_1), & \Phi(x_2)^\top \Phi(x_2), & \ldots, & \Phi(x_2)^\top \Phi(x_n) \\ \vdots, & \vdots, & \ldots, & \vdots \\ \Phi(x_n)^\top \Phi(x_1), & \Phi(x_n)^\top \Phi(x_2), & \ldots, & \Phi(x_n)^\top \Phi(x_n) \end{bmatrix}$$

$$= \begin{bmatrix} \Phi(x_1) \\ \vdots \\ \Phi(x_n) \end{bmatrix} \begin{bmatrix} \Phi(x_1), & \ldots, & \Phi(x_n) \end{bmatrix}$$

positive semidefinite

nonnegative eigenvalues

Linear Kernel

$\mathbf{X} = [x_1, \ldots, x_n]$ $\quad\quad$ $\Phi(x) = x$ $\quad\quad$ $\mathbf{K} = \mathbf{X}^\top \mathbf{X}$

$K(x_i, x_j)$

$$\min_{\alpha} \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^\top x_j - \sum_i \alpha_i$$

$$\text{s.t.} \quad \sum_i \alpha_i y_i = 0, \quad \alpha_i \in [0, C]$$

Efficient Kernel Computation

$\Phi(x)$ $\quad$ $K(x_i, x_j) = \Phi(x_i)^\top \Phi(x_j)$

$$\Phi(x_i) = [x_i^1, \ldots, x_i^d, x_i^1 x_i^1, \ldots, x_i^1 x_i^d, \ldots, x_i^d x_i^1, \ldots, x_i^d x_i^d, \ldots]^\top$$

$$| \quad | \quad | \quad | \quad | \quad | \quad | \quad | \quad | \quad | \quad |$$

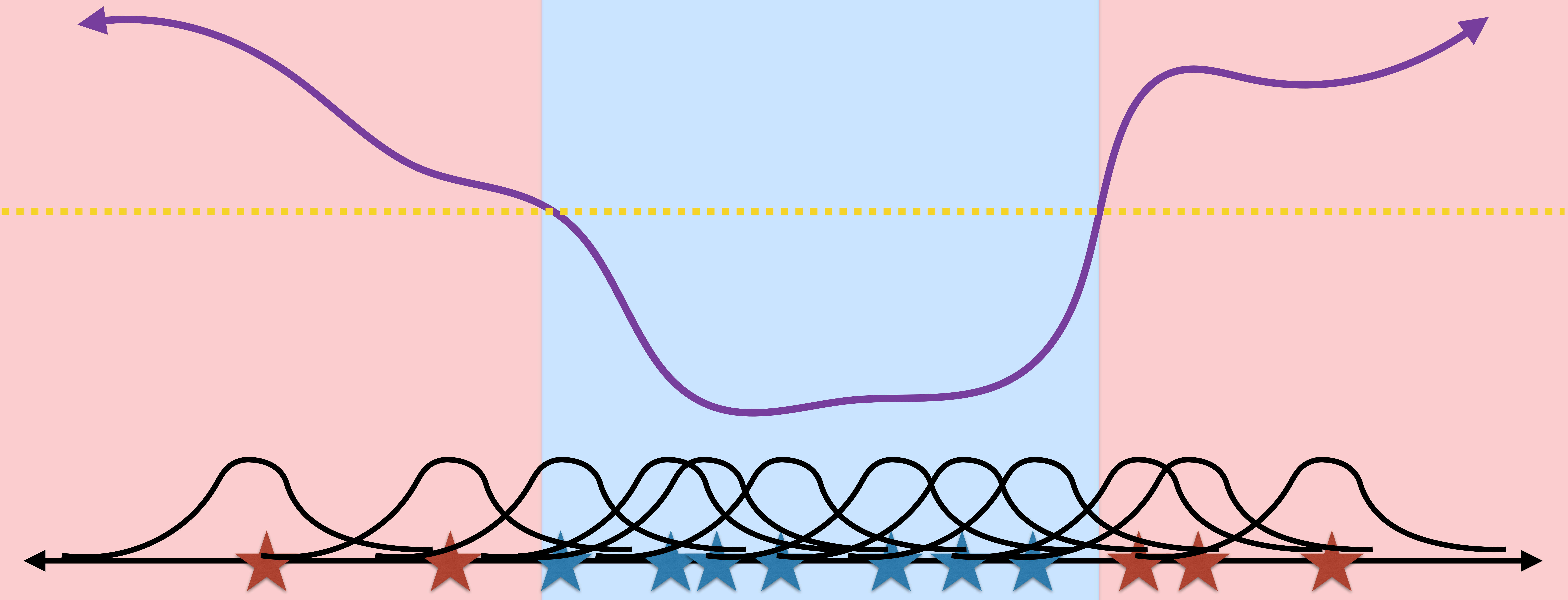
$$\Phi(x_j) = [x_j^1, \ldots, x_j^d, x_j^1 x_j^1, \ldots, x_j^1 x_j^d, \ldots, x_j^d x_j^1, \ldots, x_j^d x_j^d, \ldots]^\top$$

$(x_i^\top x_j + 1)^M$ $\quad$ $(x_i^\top x_j)(x_i^\top x_j) + 2x_i^\top x_j + 1$

$X \in \mathbb{R}^{d \times n}$ $\quad$ $K = (X^\top X + 1)^M$

elementwise exponentiation

Radial Basis Functions



Taylor Expansion of RBF Kernel

$$K(x_i, x_j) = \exp\left(-\frac{1}{2\sigma^2}||x_i - x_j||^2\right)$$

$$\sigma = 1/\sqrt{2}$$

$$= \exp\left(-||x_i - x_j||^2\right)$$

$$= \exp\left(-x_i^\top x_i\right) \exp\left(-x_j^\top x_j\right) \exp(2x_i^\top x_j)$$

$$\exp(z) = \sum_{n=0}^{\infty} \frac{z^n}{n!}$$

$$= \exp\left(-x_i^\top x_i\right) \exp\left(-x_j^\top x_j\right) \sum_{n=0}^{\infty} \frac{2^n (x_i^\top x_j)^n}{n!}$$

order-n polynomial kernel* $\Phi^n(x)$

$$\Phi^{\text{rbf}} = \exp(-x^\top x) \left[\Phi^1(x)^\top, \Phi^2(x)^\top, \ldots, \Phi^\infty(x)^\top\right]^\top$$

Kernel Formulas

Linear $K(x_i, x_j) = {x_i}^\top x_j$ $X_i \in \mathbb{R}^{d \times m}$ $X_j \in \mathbb{R}^{d \times n}$

$K = {X_i}^\top X_j$

Polynomial $K(x_i, x_j) = ({x_i}^\top x_j + 1)^M$ $K = ({X_i}^\top X_j + 1)^M$

RBF $K(x_i, x_j) = \exp\left(-\frac{1}{2\sigma^2}||x_i - x_j||^2\right)$

$$K = \exp\left(-\frac{1}{2\sigma^2}\left(\mathrm{diag}({X_i}^\top X_i)\vec{1}^\top + \vec{1}\mathrm{diag}({X_j}^\top X_j)^\top - 2{X_i}^\top X_j\right)\right)$$

Kernels

- Map input data to new feature space (usually higher dimensional)
- Efficient method for computing inner product in mapped space
- Methods using inner products can directly use kernel
  - E.g., dual SVM

Summary

- SVM primal problem has a dual optimization
- Dual has box constraints on dual variables
- Dual only considers inner products of data vectors
- Kernel trick: replace inner products with kernel functions
    - Inner products in mapped feature space