Dual SVM and Kernels

Machine Learning
CSx824/ECEx242
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

$0 < \alpha_i < C$

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

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

primal problem

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

$$- \sum_{i} \alpha_i (y_i(w^\top x_i + b) - 1 + \xi_i) - \sum_{i} \beta_i \xi_i$$

↑ ↑ ↑↑

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

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

primal problem

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

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

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)

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

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

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

$$\sum_i \alpha_i y_i = 0$$

$$\alpha = C - \beta$$

$$\beta = C - \alpha$$

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

$$\sum_i \alpha_i y_i = 0$$

$$\alpha = C - \beta$$

$$\beta = C - \alpha$$

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

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

$$\sum_i \alpha_i y_i = 0$$

$$\alpha = C - \beta$$

$$\boxed{\beta = C - \alpha}$$

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

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

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

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

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

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

$$\sum_i \alpha_i y_i = 0$$

$$\alpha = C - \beta$$

$$\beta = C - \alpha$$

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

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

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

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

Done? Not quite

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

$$\sum_i \alpha_i y_i = 0$$

$$\alpha = C - \beta$$

$$\beta = C - \alpha$$

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

$$\begin{aligned} \min_{\alpha} \;& \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^\top x_j - \sum_i \alpha_i \\ \text{s.t.} \;& \sum_i \alpha_i = 0, \quad \alpha_i \in [0, C] \end{aligned}$$

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

complementary slackness

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

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

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

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

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

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

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

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

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

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

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

$$\mathcal{Z} = \mathbb{R}^\infty$$

$$\Phi(x) = \left[\exp\left(\frac{1}{\sigma}||x - x_1||^2\right), \exp\left(\frac{1}{\sigma}||x - x_2||^2\right), \dots, \exp\left(\frac{1}{\sigma}||x - x_n||^2\right)\right]$$

(This feature map for RBF is wrong. See next video for correction)

Linear feature map

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

Quadratic feature map

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

Gaussian radial-basis (RBF) feature map

$$\mathcal{Z} = \mathbb{R}^\infty$$

$$\Phi(x) = \left[\exp\left(\frac{1}{\sigma}||x - x_1||^2\right), \exp\left(\frac{1}{\sigma}||x - x_2||^2\right), \dots, \exp\left(\frac{1}{\sigma}||x - x_n||^2\right)\right]$$

(This feature map for RBF is wrong. See next video for correction)

Gram Matrices

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

$$\mathbf{K} = \left[\begin{array}{cccc} \Phi(x_1)^\top \Phi(x_1), & \Phi(x_1)^\top \Phi(x_2), & \dots, & \Phi(x_1)^\top \Phi(x_n) \\ \Phi(x_2)^\top \Phi(x_1), & \Phi(x_2)^\top \Phi(x_2), & \dots, & \Phi(x_2)^\top \Phi(x_n) \\ \vdots, & \vdots, & \dots, & \vdots \\ \Phi(x_n)^\top \Phi(x_1), & \Phi(x_n)^\top \Phi(x_2), & \dots, & \Phi(x_n)^\top \Phi(x_n) \end{array}\right]$$

$$= \left[\begin{array}{c} \Phi(x_1) \\ \vdots \\ \Phi(x_n) \end{array}\right] \left[\begin{array}{ccc} \Phi(x_1), & \dots, & \Phi(x_n) \end{array}\right]$$

positive semidefinite

nonnegative eigenvalues

Linear Kernel

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

$K(x_i, x_j)$

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

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

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
- Next time: how to efficiently compute polynomial and RBF kernels