SVM part I

Peng Zhang

June 1, 2019

OUTLINES

- General class of regularization problem
- Kernel
- Reproducing kernel Hilbert Space.
- Support Vector Classifier

$$\min_{f \in \mathcal{H}} \Big[\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \Big] \tag{1}$$

- $L(y, f(x))$ is a loss function.
- $J(f)$ is a penalty functional.
- $\mathcal{H}$ Hilbert space.

For case: $J(f) = \int_{\mathcal{R}^p} \frac{|\tilde{f}(s)|^2}{\tilde{G}(s)} ds$,
solutions have the form:

$$f(X) = \sum_{k=1}^{K} \alpha_k \phi_k(X) + \sum_{i=1}^{N} \theta_i G(X - x_i). \tag{2}$$

The solution is finite dimensional, while defined over an infinite-dimensional space

Kernel

Definition

A function $K : R^p \times R^p \to R$ is called kernel if
(1)it is symmetric,i.e $K(x, y) = K(y, x)$
(2)it is positive definite, that is $\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j K(x_i, x_j) \geq 0$ for any $N \in \mathcal{N}$,
$x_1, \cdots, x_n \in R^p$, $c_1, \cdots, c_n \in R$

- Sums of kernels are kernels.
- Products of kernels are kernels.

Kernel

Definition

A function $K : R^p \times R^p \to R$ is called kernel if
(1)it is symmetric,i.e $K(x, y) = K(y, x)$
(2)it is positive definite, that is $\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j K(x_i, x_j) \geq 0$ for any $N \in \mathcal{N}$, $x_1, \cdots, x_n \in R^p$, $c_1, \cdots, c_n \in R$

- Sums of kernels are kernels.
- Products of kernels are kernels.

Samples:

- Polynomial kernel: $K(x, y) = (1+ < x, y >)^d$.
- Exponential kernel: $K(x, y) = \exp(< x, y >)$.
- Gaussian kernel: $K(x, y) = \exp\left(-\nu ||x - y||^2\right)$
- Neural network: $K(x, y) = \tanh(\kappa_1 < x, y > +\kappa_2)$

$\mathcal{H}$:reproducing kernel Hilbert space(RKHS)

Definition

$\mathcal{H}$ is Hilbert space if $\mathcal{H}$ is a complete metric space with respect to the distance function induced by the inner product $< x, y >_{\mathcal{H}}$.

$< x, y >_{\mathcal{H}}$ satisfies:

- conjugate symmetric: $< x, y >= \overline{< y, x >}$
- linear: $< ax_1 + bx_2, y >= a < x_1, y > +b < x_2, y >$
- positive definite: $< x, x >_{\mathcal{H}} \geq 0$ and $< x, x >_{\mathcal{H}} = 0 \Leftrightarrow x = 0$

Theorem

A reproducing Hilbert space defines a positive kernel. Conversely, a positive definite kernel defines a reproducing Hilbert space.

$\mathcal{H}$:reproducing kernel Hilbert space(RKHS)

- Given kernel $K(x, y)$, function $K_x \in \mathcal{H} : R^p \to R$ is $K_x(z) = K(x, z)$, associated with inner product $< K_x, K_y >_{\mathcal{H}} = K(x, y)$–reproducing.
- Suppose kernel K has an eigen-expansion(Mercer's Theorem):

$$K(x, y) = \sum_{i=1}^{\infty} \gamma_i \phi_i(x) \phi_i(y) \tag{3}$$

$\mathcal{H}$:reproducing kernel Hilbert space(RKHS)

- Given kernel $K(x, y)$, function $K_x \in \mathcal{H} : R^p \to R$ is $K_x(z) = K(x, z)$, associated with inner product $< K_x, K_y >_{\mathcal{H}} = K(x, y)$–reproducing.
- Suppose kernel K has an eigen-expansion(Mercer's Theorem):

$$K(x, y) = \sum_{i=1}^{\infty} \gamma_i \phi_i(x) \phi_i(y) \tag{3}$$

- $f \in \mathcal{H}$:

$$f(x) = \sum_{i=1}^{\infty} c_i \phi_i(x). \tag{4}$$

associated with inner product $< f, g >_{\mathcal{H}} = \sum_{i=0}^{\infty} \frac{c_i d_i}{\gamma_i}$

Penalty functional: $J(f) =< f, f >_{\mathcal{H}}$

Problem:

$$\min_{f \in \mathcal{H}} \Big[\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \Big]$$

becomes into:

$$\min_{\{c_i\}_1^\infty} \Big[\sum_{i=1}^{N} L(y_i, \sum_{j=1}^{\infty} c_j \phi_j(x_i)) + \lambda \sum_{j=1}^{\infty} c_j^2 / \gamma_j \Big]. \tag{5}$$

Solution form, which is proved in Ex.5.15, is:

$$f(x) = \sum_{i=1}^{N} \alpha_i K_{x_i}(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i). \tag{6}$$

It is finite-dimensional.

Matrix form

$$
\begin{aligned}
J(f) &=< f, f >_{\mathcal{H}} =< \sum_{i=1}^{N} \alpha_i K_{x_i}(z), \sum_{j=1}^{N} \alpha_j K_{x_j}(z) > \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j < K_{x_i}(z), K_{x_j}(z) >= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j K(x_i, x_j) \\
&= \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha}
\end{aligned}
$$

$$
f(x_i) = \sum_{j=1}^{N} \alpha_i K(x_i, x_j)
$$

$$
[f(x_1), \cdots, f(x_N)]^T = \boldsymbol{K} \boldsymbol{\alpha}
$$

$L(y, f(x))$: squared error loss

Penalized least squares problem (PLSP):

$$\min_{\boldsymbol{\alpha}} L(y, \boldsymbol{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^T \boldsymbol{K}\boldsymbol{\alpha}$$
$$\min_{\boldsymbol{\alpha}} (y - \boldsymbol{K}\boldsymbol{\alpha})^T (y - \boldsymbol{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^T \boldsymbol{K}\boldsymbol{\alpha}$$

Solution of $\boldsymbol{\alpha}$, $f(x)$ are

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{K} + \lambda I)^{-1} y \tag{7}$$
$$\hat{f}(x) = \sum_{i=1}^{N} \hat{\alpha}_i K(x, x_i). \tag{8}$$

Polynomial kernel: $K(x, y) = (< x, y > +1)^d$

For, $x, y \in R^p$, has $M = \binom{p+d}{d}$ eigen-functions.
Sample $(p = 2, d = 2, M = 6)$:

- $K(x, y) = 1 + 2x_1y_1 + 2x_2y_2 + x_1^2y_1^2 + x_2^2y_2^2 + 2x_1x_2y_1y_2$
- $h(x)^T = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$

$K(x, y) = \sum_{m=1}^{M} h_m(x)h_m(y)$
$f(x) = \sum_{m=1}^{M} \beta_m h_m(x)$

Polynomial kernel: $K(x, y) = (< x, y > +1)^d$

Penalized polynomial regression problem (PPRP):

$$\min_{\{\beta_m\}_1^M} \sum_{i=1}^{N} \left(y_i - \sum_{m=1}^{M} \beta_m h_m(x_i)\right)^2 + \lambda \sum_{m=1}^{M} \beta_m^2 \tag{9}$$

$$\min_{\boldsymbol{\beta}} (y - \boldsymbol{H}\boldsymbol{\beta})^T (y - \boldsymbol{H}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}. \tag{10}$$

Polynomial kernel: $K(x, y) = (< x, y > +1)^d$

Penalized polynomial regression problem (PPRP):

$$\min_{\{\beta_m\}_1^M} \sum_{i=1}^{N} \left(y_i - \sum_{m=1}^{M} \beta_m h_m(x_i)\right)^2 + \lambda \sum_{m=1}^{M} \beta_m^2 \tag{9}$$

$$\min_{\boldsymbol{\beta}} (y - \boldsymbol{H}\boldsymbol{\beta})^T (y - \boldsymbol{H}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}. \tag{10}$$

Solution of β and $f(x)$ are:

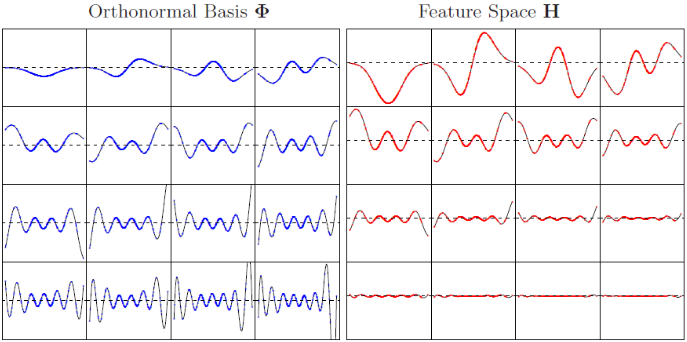
$$\hat{\boldsymbol{\beta}} = (\lambda I + \boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T y$$

$$\hat{f}(x) = \sum_{m=1}^{M} \hat{\beta}_m h_m(x)$$

This problem is equivalent to penalized least squares problem (PLSP) by Ex.5.16

Gaussian kernel: $K(x, y) = e^{-\nu||x-y||^2}$

- Eigen-decomposition: $\boldsymbol{K} = \boldsymbol{\Phi} \boldsymbol{D}_\gamma \boldsymbol{\Phi}^T$.
- The ith columns of $\boldsymbol{\Phi}$ is the empirical estimates of the eigen expansion function $\hat{\phi}_i(x)$.
- Feature space representation: $h_i(x) = \sqrt{(\hat{\gamma}_i)}\hat{\phi}_i(x)$, $i = 1, \cdots, N$.



$L(y, f(x))$: SVM Hinge Loss

$L(y, f(x)) = [1 - yf(x)]_+$

$$\min_{\alpha_0, \boldsymbol{\alpha}} \left(\sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \frac{\lambda}{2} \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} \right)$$

A finite dimensional solution of the form

$$f(x) = \alpha_0 + \sum_{i=1}^{N} \alpha_i K(x, x_i) \tag{11}$$

OUTLINES

- General class of regularization problem
- Kernel
- Reproducing kernel Hilbert Space.
- Support Vector Classifier

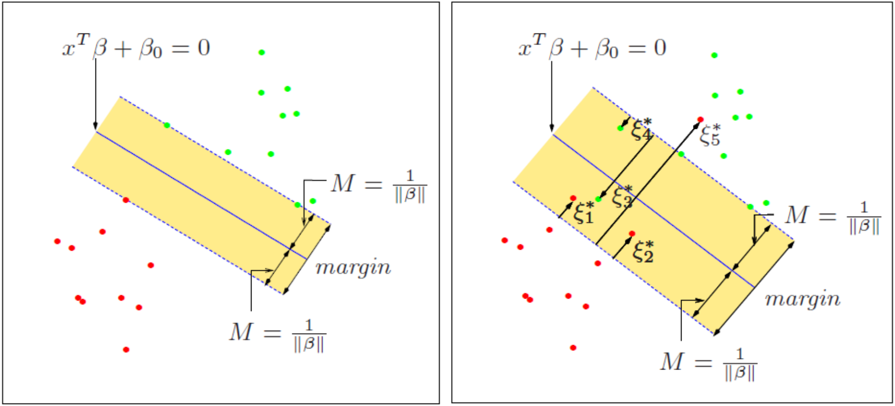


FIGURE 12.1. *Support vector classifiers. The left panel shows the separable case. The decision boundary is the solid line, while broken lines bound the shaded maximal margin of width $2M = 2/\|\beta\|$. The right panel shows the nonseparable (overlap) case. The points labeled ξ_j^* are on the wrong side of their margin by an amount $\xi_j^* = M\xi_j$; points on the correct side have $\xi_j^* = 0$. The margin is maximized subject to a total budget $\sum \xi_i \leq$ constant. Hence $\sum \xi_j^*$ is the total distance of points on the wrong side of their margin.*

Given N pairs $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$, with $x_i \in \mathcal{R}^p$, $y_i \in \{-1, 1\}$, the hyperplane is

$$\{x : f(x) = x^T \beta + \beta_0 = 0\}.$$

Class are separable

Purpose: find a function $f(x) = x^T\beta + \beta_0$, which meet,

$$\left\{\begin{array}{ll} f(x_i) > 0, \text{if} & y_i > 0 \\ f(x_i) < 0, \text{if} & y_i < 0 \end{array}\right. \Leftrightarrow y_i f(x_i) > 0$$

and the margin as big as possible.

$$\max_{\beta,\beta_0,||\beta||=1} M$$
$$\text{subject to:} y_i(x_i^T\beta + \beta_0) \geq M, i = 1, \cdots, N.$$

or

$$\min_{\beta,\beta_0} ||\beta||$$
$$\text{subject to:} y_i(x_i^T\beta + \beta_0) \geq 1, i = 1, \cdots, N.$$

Class are overlap

Introduce the slack variables $\xi = (\xi_1, \xi_2, \cdots, \xi_N)$,

$$\min_{\beta,\beta_0,\xi} ||\beta||$$

$$\text{subject to:} y_i(x_i\beta^T + \beta_0) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \sum \xi_i \leq constant, \forall i$$

ξ_i is the proportional amount by which the prediction $f(x_i) = x_i^T\beta + \beta_0$ is on the wrong side of its margin.

$\xi_i = 0$: correct side;

$\xi_i > 1$: Misclassifications.

Computing the Support Vector Classifier

$$\min_{\beta,\beta_0,\xi} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\xi_i$$

subject to: $\xi_i \geq 0, \quad y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \forall i$

Computing the Support Vector Classifier

$$\min_{\beta,\beta_0,\xi} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{subject to:}\xi_i \geq 0, \quad y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \forall i$$

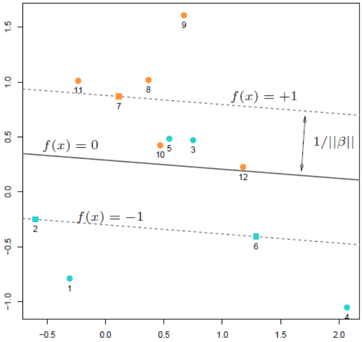
$$L_p = \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i[y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^{N}\mu_i\xi_i$$

$$L_D(\beta_0, \xi, \alpha, \mu) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\alpha_i\alpha_{i'}y_iy_i'x_i^Tx_{i'}$$

with $\beta = \sum_{i=1}^{N}\alpha_i y_i x_i$.
Maximizing L_D is a simpler convex quadratic programming problem than the primal.

Support vectors

The solution for β has the form, $\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i$
Support vectors: observations with nonzero coefficients $\hat{\alpha}_i$.
points on the wrong side of the boundary;
points on the correct side of the boundary but close to it.
The number of support vector should be as small as possible.



Support Vector Machines

Basis functions $h_m(x), m = 1, \cdots, M$, with $h(x_i) \equiv (h_1(x_i), \cdots, h_M(x_i))$, try to produce the function $f(x) = h(x)^T\beta + \beta_0$.

$$\min_{\beta,\beta_0,\xi} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{subject to:} \xi_i \geq 0, \quad y_i(h(x_i)^T\beta + \beta_0) \geq 1 - \xi_i, \forall i$$

Support Vector Machines

Basis functions $h_m(x), m = 1, \cdots, M$, with $h(x_i) \equiv (h_1(x_i), \cdots, h_M(x_i))$, try to produce the function $f(x) = h(x)^T\beta + \beta_0$.

$$\min_{\beta,\beta_0,\xi} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{subject to:}\xi_i \geq 0, \quad y_i(h(x_i)^T\beta + \beta_0) \geq 1 - \xi_i, \forall i$$

$$L_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\alpha_i\alpha_{i'}y_iy_i' < h(x_i), h(x_{i'}) >$$

$$\beta = \sum_{i=1}^{N}\alpha_iy_ih(x_i)$$

$$f(x) = h(x)^T\beta + \beta_0 = \sum_{i=1}^{N}\alpha_iy_i < h(x), h(x_i) > +\beta_0$$

SVM as a Penalization Method

$$\min_{\beta,\beta_0,\xi} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{subject to:}\xi_i \geq 0, \quad y_i f(x_i) \geq 1 - \xi_i, \forall i$$

Constraint: $\xi_i \geq 0$ and $\xi_i \geq 1 - y_i f(x_i)$, minimal value of ξ_i:
$\xi_i = \max(0, 1 - y_i f(x_i))$

SVM as a Penalization Method

$$\min_{\beta,\beta_0,\xi} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\xi_i$$
$$\text{subject to:} \xi_i \geq 0, \quad y_i f(x_i) \geq 1 - \xi_i, \forall i$$

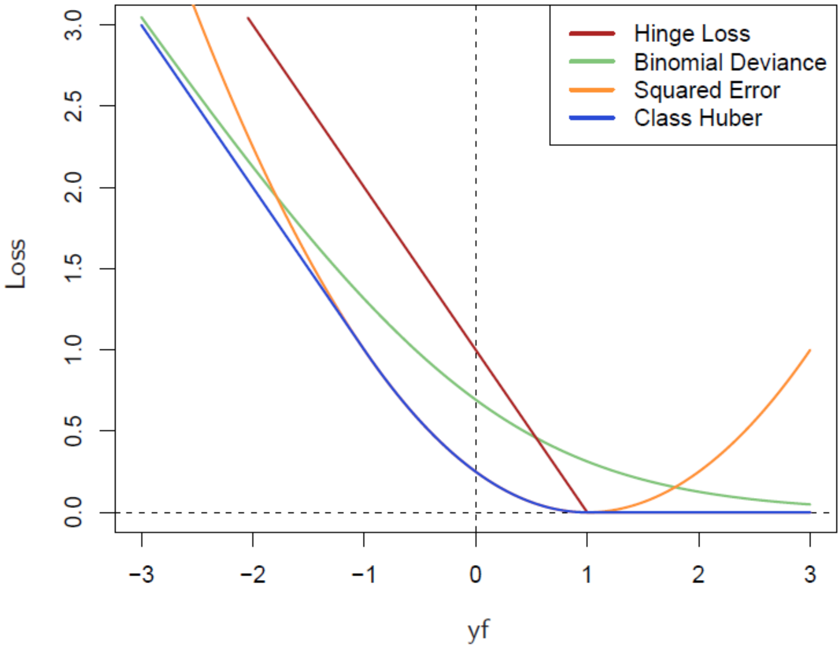
Constraint: $\xi_i \geq 0$ and $\xi_i \geq 1 - y_i f(x_i)$, minimal value of ξ_i:
$\xi_i = \max(0, 1 - y_i f(x_i))$
Equivalently to penalization method:

$$\min_{\beta,\beta_0} \sum_{i=1}^{N}[1 - y_i f(x_i)]_+ + \frac{\lambda}{2}||\beta||^2$$

with $\lambda = 1/C$.

TABLE 12.1. *The population minimizers for the different loss functions in Figure 12.4. Logistic regression uses the binomial log-likelihood or deviance. Linear discriminant analysis (Exercise 4.2) uses squared-error loss. The SVM hinge loss estimates the mode of the posterior class probabilities, whereas the others estimate a linear transformation of these probabilities.*

Loss Function	$L[y, f(x)]$	Minimizing Function
Binomial Deviance	$\log[1 + e^{-yf(x)}]$	$f(x) = \log \frac{\Pr(Y = +1\|x)}{\Pr(Y = \text{-}1\|x)}$
SVM Hinge Loss	$[1 - yf(x)]_+$	$f(x) = \text{sign}[\Pr(Y = +1\|x) - \frac{1}{2}]$
Squared Error	$[y - f(x)]^2 = [1 - yf(x)]^2$	$f(x) = 2\Pr(Y = +1\|x) - 1$
"Huberised" Square Hinge Loss	$-4yf(x)$, $\quad yf(x) < \text{-}1$ $[1 - yf(x)]_+^2 \quad$ otherwise	$f(x) = 2\Pr(Y = +1\|x) - 1$


Hinge Loss
Binomial Deviance
Squared Error
Class Huber
Loss
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−3
−2
−1
0
1
2
3
yf

Thank You