ST509 Computational Statistics

Lecture 8: ℓ_2-penalized Kernel Machines

Seung Jun Shin

Department of Statistics
Korea University

E-mail: sjshin@korea.ac.kr

Duality I

- Constraint Optimization for $\mathbf{x}$:

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h_i(\mathbf{x}) \leq 0, i = 1, \cdots, m \\ & \ell_j(\mathbf{x}) = 0, j = 1, \cdots, r \end{aligned} \tag{1}$$

- The Lagrangian associated the problem (1) is

$$f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i h_i(\mathbf{x}) + \sum_{j=1}^{r} \nu_j \ell_j(\mathbf{x})$$

where $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_m)$ and $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_r)$ are called the dual variables or Lagrange multipliers associated with the problem (1).

Duality II

- We define the Lagrange dual function of dual variables, $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ as

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} \left\{ f(x) + \sum_{i=1}^{m} \lambda_i h_i(\mathbf{x}) + \sum_{j=1}^{r} \nu_j \ell_j(\mathbf{x}) \right\}$$

which is an affine function and hence always concave (and convex).

Duality III

- Let p^* denote the optimal value of problem (1).
- For any $\boldsymbol{\lambda} \geq 0$ and any $\boldsymbol{\nu}$, we have

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*$$

- This is because and for any feasible point $\tilde{\mathbf{x}}$ and $\boldsymbol{\lambda} \geq 0$

$$\sum_{i=1}^{m} \lambda_i h_i(\tilde{\mathbf{x}}) + \sum_{j=1}^{r} \nu_j \ell_j(\tilde{\mathbf{x}}) \leq 0,$$

  and thus we have

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq L(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f(\tilde{\mathbf{x}}).$$

Duality IV

- Lagrangian dual function provide a lower bound of p^*.
- What is the best lower bound?
- This leads the optimization problem

$$\begin{aligned} \max_{\boldsymbol{\lambda},\boldsymbol{\gamma}} \quad & g(\boldsymbol{\lambda},\boldsymbol{\gamma}), \\ \text{subject to} \quad & \boldsymbol{\lambda} \geq 0. \end{aligned} \tag{2}$$

which we call the Lagrange dual problem associated with the problem.

Duality V

- Let d^* be the optimal value of (2). Then we have

$$d^* \leq p^* \qquad \text{(a.k.a., weak duality)}$$

- We call their difference $p^* - d^*$ the duality gap.
- If the duality gap is 0, i.e. $d^* = p^*$, we say that strong duality holds.

Duality VI

- Suppose strong duality holds (i.e., $d^* = p^*$).
- Let $\mathbf{x}^*$ be a primal optimal and $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ be dual optimal point:

$$
\begin{aligned}
f(\mathbf{x}^*) &= g(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \\
&= \inf_{\mathbf{x}} \left\{ f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i^* h_i(\mathbf{x}) + \sum_{j=1}^{r} \nu_j^* \ell_j(\mathbf{x}) \right\} \\
&\leq f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* h_i(\mathbf{x}^*) + \sum_{j=1}^{r} \nu_j^* \ell_j(\mathbf{x}^*) \\
&\leq f(\mathbf{x}^*)
\end{aligned}
$$

- We must have $\sum_{i=1}^{m} \lambda_i^* h_i(\mathbf{x}^*) = 0$ and thus

$$
\lambda_i^* h_i(\mathbf{x}^*) = 0, \ i = 1, \cdots, m
$$

- This is known as complementary slackness: it holds for any optimal points $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ when strong duality holds.

Duality VII

- Assume all functions in (1) which may not be convex are differentiable.
- Suppose $\mathbf{x}^*$ and $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ are primal and dual optimal points with strong duality.
- We must have
  1. $\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(\mathbf{x}^*) + \sum_{j=1}^{r} \nu_j^* \nabla \ell_j(\mathbf{x}^*) = 0$ (Stantionary)
  2. $\lambda_i^* h_i(\mathbf{x}^*) = 0, i = 1, \cdots, m$ (Complementary Slackness)
  3. $h_i(\mathbf{x}^*) \leq 0, i = 1, \cdots, m$ and
     $\ell_j(\mathbf{x}^*) = 0, j = 1, \cdots, r$ (Primal Feasibility)
  4. $\lambda_i^* \geq 0$ for all $i = 1, \cdots, m$ (Dual Feasibility)
- These are called the Karush-Kuhn-Tucker (KKT) conditions.
- That is KKT condition is a necessary conditions for the optimal points $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ with strong duality.

Duality VIII

- If (1) is convex (i.e., f is convex and h_i are affine), then the KKT conditions are also sufficient!
- Any points satisfying KKT conditions are primal and dual optimal with strong duality.

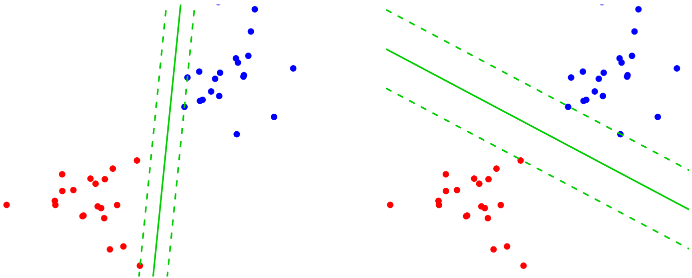
Support Vector Machine I

- Given a set of data $(y_i, \mathbf{x}_i) \in \{-1, 1\} \times \mathbb{R}^p$
- Decision function predicts the response as its sign, i.e., $\hat{y} = f(\mathbf{x})$:
- We assume the decision function is linear in $\mathbf{x}$:

$$f(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}.$$

Support Vector Machine II

- Starts with the separable case.
- Maximal margin classifier seeks an optimal hyperplane of $\mathbf{x}$:



(a) Not optimal (b) Optimal

Support Vector Machine III

- Recall

$$\cos(\theta) = \frac{\langle \mathbf{x} - \mathbf{x}_0, \boldsymbol{\beta} \rangle}{\|\mathbf{x} - \mathbf{x}_0\| \|\boldsymbol{\beta}\|}$$

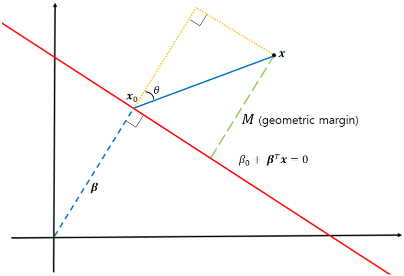


Figure: Geometric margin

Support Vector Machine IV

- Assuming $\|\boldsymbol{\beta}\| = 1$,

  $$M = \cos(\theta)\|\mathbf{x} - \mathbf{x}_0\| = \langle \mathbf{x} - \mathbf{x}_0, \boldsymbol{\beta} \rangle = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}, \quad \text{(if } \mathbf{x} \text{ is on the right)}$$

- Thus

  $$M = y(\beta_0 + \boldsymbol{\beta}^T \mathbf{x})$$

- Maximal margin classifier can be formulated as

  $$\max_{\beta_0, \boldsymbol{\beta}} M$$

  subject to $\|\boldsymbol{\beta}\| = 1$ and $y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) \geq M, i = 1, \cdots, n$

Support Vector Machine V

- Assuming $M = 1/\|\boldsymbol{\beta}\|$, the maximal margin classifier equivalently solves

$$
\begin{aligned}
\min_{\beta_0, \boldsymbol{\beta}} \quad & \frac{1}{2}\|\boldsymbol{\beta}\|^2 \\
\text{subject to} \quad & y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) \geq 1, i = 1, \cdots, n
\end{aligned}
\tag{3}
$$

Support Vector Machine VI

- If the data are not linearly separable, no solution of (3) exists.
- Soft margin classifier solves

$$\begin{aligned} \min_{\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^n \xi_i \\ \text{subject to} \quad & y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) \geq 1 - \xi_i, \quad i = 1, \cdots, n; \\ & \xi_i \geq 0, \qquad\qquad\qquad\quad i = 1, \cdots, n, \end{aligned} \tag{4}$$

  where $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_n)^T$ and $C > 0$ controls the cost for the violation of constraints.
- We call the soft margin classifier linear support vector machine (SVM).

Kernel SVM I

- Linear SVM can be equivalently rewritten as

$$\min \sum_{i=1}^{n} [1 - y_i f(\mathbf{x}_i)]_+ + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2 \tag{5}$$

  where $[a]_+ = \{0, a\}$ and $\lambda = C^{-1}$.

- Familiar "loss and penalty" expression.
- The margin $u_i = y_i f(\mathbf{x}_i)$ plays a role like residual $y_i - f(\mathbf{x}_i)$ in regression.
- The loss function for SVM, $H_1(u) = [1 - u]$ of margin u is called Hinge loss.

Kernel SVM II

- Hinge loss behaves similar to the logistic loss $L(u) = \log(1 + e^{-u})$.
- In fact they are convex surrogates of the 0-1 loss $L^*(u) = \mathbb{1}\{u \geq 0\}$ which yields Bayes classifier.

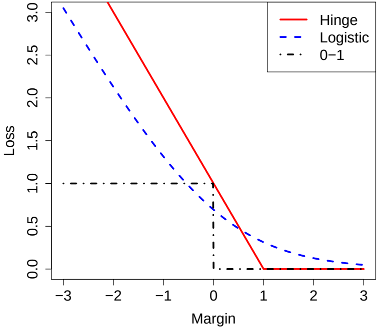


Figure: Hinge vs. Logistic vs. 0-1 Loss

Kernel SVM III

- Now, we assume f is a possibly nonlinear and lies on $\mathcal{F}$, a space of function.
- Nonlinear extension of (5) is

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} [1 - y_i f(\mathbf{x}_i)]_+ + \frac{\lambda}{2} \|f\|_{\mathcal{F}}^2 \tag{6}$$

- However, (6) is infeasible.

Kernel SVM IV

- We let $\mathcal{F}$ be $\mathcal{H}_K$, the reproducing kernel Hilbert space (RKHS) generated by a kernel function $K(\mathbf{x}, \mathbf{x}')$.
- Popular choice of the kernel includes:
  - Linear: $K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}'$
  - Polynomial: $K(\mathbf{x}, \mathbf{x}') = (1 + \mathbf{x}^T\mathbf{x}')$
  - Radial (Gaussian): $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|)$

Kernel SVM V

- (6) becomes

$$\min_{f \in \mathcal{H}_K} \sum_{i=1}^{n} [1 - y_i f(\mathbf{x}_i)]_+ + \frac{\lambda}{2} \|f\|^2_{\mathcal{H}_K} \tag{7}$$

- Representer Theorem states that the solution of (9) has the following finite form:

$$f(\mathbf{x}) = \theta_0 + \sum_{i=1}^{n} \theta_i K(\mathbf{x}, \mathbf{x}_i). \tag{8}$$

Kernel SVM VI

- Plugging (8) into (9), we have

$$\min_{\theta_0, \boldsymbol{\theta}} \sum_{i=1}^{n} [1 - y_i f(\mathbf{x}_i)]_+ + \frac{\lambda}{2} \boldsymbol{\theta}^T \mathbf{K} \boldsymbol{\theta} \tag{9}$$

  where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^T$ and $\mathbf{K}$ denotes the $(n \times n)$-dimensional kernel matrix with $\{\mathbf{K}\}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$.

- Going back to constraint form,

$$\begin{aligned} \min_{\theta_0, \boldsymbol{\theta}, \boldsymbol{\xi}} \quad & \tfrac{1}{2} \boldsymbol{\theta}^T \mathbf{K} \boldsymbol{\theta} + C \textstyle\sum_{i=1}^{n} \xi_i && (10) \\ \text{subject to} \quad & y_i \{\theta_0 + \textstyle\sum_{i=1}^{n} \theta_i K(\mathbf{x}, \mathbf{x}_i)\} \geq 1 - \xi_i, && i = 1, \cdots, n; \\ & \xi_i \geq 0, && i = 1, \cdots, n, \end{aligned}$$

- We call this (kernel) SVM.

Computation of SVM I

- Lagrangian primal function of the linear SVM (4) with $\lambda = C^{-1}$ is

$$L_p : \frac{\lambda}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} + \sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\{1 - y_i(\beta_0 - \boldsymbol{\beta}^T\mathbf{x}_i) - \xi_i\} - \gamma_i\sum_{i=1}^{n}\xi_i \quad (11)$$

- Taking derivative w.r.t primal variables $\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}$:

$$\frac{\partial}{\partial\boldsymbol{\beta}}L_p : \quad \boldsymbol{\beta} = \frac{1}{\lambda}\sum_{i=1}^{n}\alpha_i y_i\mathbf{x}_i \quad (12)$$

$$\frac{\partial}{\partial\beta_0}L_p : \quad \sum_{i=1}^{n}\alpha_i y_i = 0 \quad (13)$$

$$\frac{\partial}{\partial\xi_i}L_p : \quad \alpha_i = 1 - \gamma_i \quad (14)$$

- KKT complementary conditions:

$$\begin{aligned} \alpha_i\{1 - y_i(\beta_0 + \boldsymbol{\beta}^T\mathbf{x}_i) - \xi_i\} &= 0 \\ \gamma_i\xi_i &= 0 \end{aligned}$$

Computation of SVM II

- Plugging (18)– (20) into (17), Dual problem is the following QP:

$$\begin{aligned} \max_{\alpha_1,\cdots,\alpha_n} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{subject to} \quad & 0 \leq \alpha_i \leq 1, \quad i = 1, \cdots, n \\ & \sum_{i=1}^{n} \alpha_i y_i = 0. \end{aligned} \tag{15}$$

- By KKT conditions, we must have for all $k \in \{i : 0 < \alpha_i < 1\}$ (a.k.a Support Vectors)

$$1 - y_k(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_k) = 0$$

- The intercept is computed by

$$\beta_0 = y_i - \frac{1}{\lambda} \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i^T \mathbf{x}_k$$

for any support vector $\mathbf{x}_k$.

Computation of SVM III

- It can be shown that the kernel SVM solves the following dual problem:

$$\begin{aligned} \max_{\alpha_1,\cdots,\alpha_n} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq 1, \quad i = 1, \cdots, n \\ & \sum_{i=1}^{n} \alpha_i y_i = 0. \end{aligned} \tag{16}$$

- People refer this kernel trick!

Other Kernel Machine I

- Nonlinear extension of Ridge Regression (RR) is

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \{y_i - f(\mathbf{x}_i)\}^2 + \frac{\lambda}{2} \|f\|_{\mathcal{F}}^2$$

- Kernel Ridge Regression let $\mathcal{F} = \mathcal{H}_K$, and solves

$$\min_{\theta_0, \boldsymbol{\theta}} \sum_{i=1}^{n} \{y_i - f(\mathbf{x}_i)\}^2 + \frac{\lambda}{2} \boldsymbol{\theta}^T \mathbf{K} \boldsymbol{\theta}$$

  where $f(\mathbf{x})$ is defined in (8).

- Solution has the closed form, just like the conventional RR.

Other Kernel Machine II

- Kernel Quantile Regression (KQR) solves

$$\min_{\theta_0,\boldsymbol{\theta}} \sum_{i=1}^{n} \rho_\tau\{y_i - f(\mathbf{x}_i)\} + \frac{\lambda}{2}\boldsymbol{\theta}^T\mathbf{K}\boldsymbol{\theta}$$

  where the check loss

$$\rho_\tau(u) = \begin{cases} \tau u & \text{if } u \geq 0 \\ -(1-\tau)u & \text{if } u < 0 \end{cases} = \tau u - u\mathbb{1}\{u < 0\}$$

  with $\tau \in (0, 1)$.

- $f(\mathbf{x})$ is the τth conditional quantile of Y given $\mathbf{x}$.
- Dual problem of KQR is QP, just like SVM.

Other Kernel Machine III

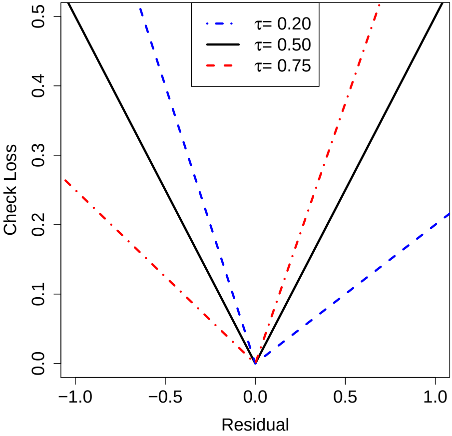


Figure: Check Loss for different values of τ.

Other Kernel Machine IV

- Support Vector Regression (SVR) solves

$$\min_{\theta_0, \boldsymbol{\theta}} \sum_{i=1}^{n} L_\epsilon\{y_i - f(\mathbf{x}_i)\} + \frac{\lambda}{2}\boldsymbol{\theta}^T \mathbf{K} \boldsymbol{\theta}$$

  where the ϵ-intensive loss is

$$L_\epsilon(u) = \begin{cases} 0 & \text{if } |u| \geq \epsilon \\ |u| - \epsilon & \text{if } |u| > \epsilon \end{cases}$$

  for $\epsilon > 0$.

- Dual problem of SVR is QP, just like SVM.

Other Kernel Machine V

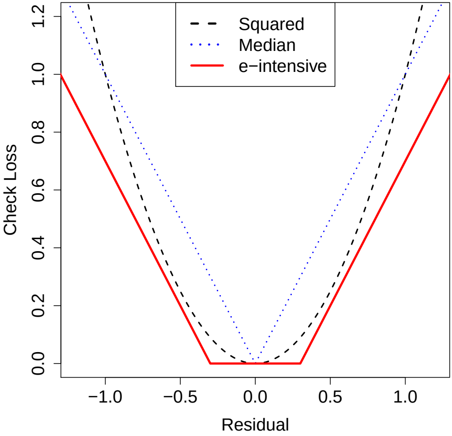


Figure: ϵ-intensive loss for SVR.

Other Kernel Machine VI

- Kernel Logistic Regression (KLR) solves

$$\min_{\theta_0, \boldsymbol{\theta}} \sum_{i=1}^{n} \log[1 + \exp\{-y_i f(\mathbf{x}_i)\}] + \frac{\lambda}{2} \boldsymbol{\theta}^T \mathbf{K} \boldsymbol{\theta}$$

  where $f(\mathbf{x})$ is defined in (8).

- We can apply Newton Raphson method to solve KLR.

Regularization path of SVM I

- Recall Lagrangian primal function of the linear SVM (4):

$$L_p : \frac{\lambda}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} + \sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\{1 - y_i(\beta_0 - \boldsymbol{\beta}^T\mathbf{x}_i) - \xi_i\} - \gamma_i\sum_{i=1}^{n}\xi_i \quad (17)$$

- Taking derivative w.r.t primal variables $\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}$:

$$\frac{\partial}{\partial\boldsymbol{\beta}}L_p : \quad \boldsymbol{\beta} = \frac{1}{\lambda}\sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i \quad (18)$$

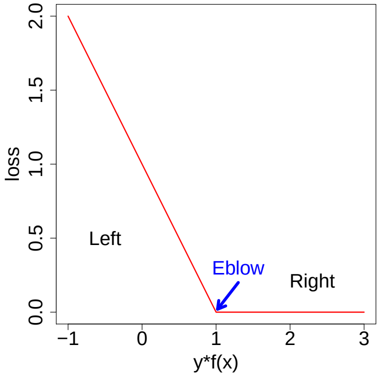
$$\frac{\partial}{\partial\beta_0}L_p : \quad \sum_{i=1}^{n}\alpha_i y_i = 0 \quad (19)$$

$$\frac{\partial}{\partial\xi_i}L_p : \quad \alpha_i = 1 - \gamma_i \quad (20)$$

- KKT complementary conditions:

$$\begin{aligned} \alpha_i\{1 - y_i(\beta_0 + \boldsymbol{\beta}^T\mathbf{x}_i) - \xi_i\} &= 0 \\ \gamma_i\xi_i &= 0 \end{aligned}$$

Regularization path of SVM II



- Define
  - $\mathcal{E} = \{i : y_i f(\mathbf{x}_i) = 1, \text{ and } 0 \le \alpha \le 1\}$
  - $\mathcal{L} = \{i : y_i f(\mathbf{x}_i) < 1, \text{ and } \alpha = 1\}$
  - $\mathcal{R} = \{i : y_i f(\mathbf{x}_i) > 1, \text{ and } \alpha = 0\}$

Regularization path of SVM III

- SVM decision function is

$$f(\mathbf{x}) = \frac{1}{\lambda}\left(\alpha_0 + \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i)\right)$$

- As λ changes, $(\alpha_0, \boldsymbol{\alpha})$ and hence the sets do. We call it event.

$$\mathcal{L} \quad \Leftrightarrow \quad \mathcal{E} \quad \Leftrightarrow \quad \mathcal{R}$$

Regularization path of SVM IV

- Let $\lambda_\ell > \lambda_{\ell+1}$ are the two adjacent event points. $\alpha_i^\ell, i = 0, 1, \cdots, n$, $f^\ell(\mathbf{x})$, $\mathcal{E}_\ell, \mathcal{L}_\ell$, and $\mathcal{R}_\ell$ are obtained at λ_ℓ.
- For $\lambda \in (\lambda_{\ell+1}, \lambda_\ell)$, we have

$$
\begin{aligned}
f(\mathbf{x}) &= \left[f(\mathbf{x}) - \frac{\lambda_\ell}{\lambda} f^\ell(\mathbf{x}) \right] + \frac{\lambda_\ell}{\lambda} f^\ell(\mathbf{x}) \\
&= \frac{1}{\lambda} \left[\sum_{j \in \mathcal{E}_\ell} (\alpha_j - \alpha_j^\ell) y_j K(\mathbf{x}, \mathbf{x}_j) + (\alpha_0 - \alpha_0^\ell) + \lambda_\ell f^\ell(\mathbf{x}) \right]
\end{aligned}
$$

Regularization path of SVM V

- For any $\mathbf{x}_i$ where $i \in \mathcal{E}_\ell$,

$$\frac{1}{\lambda}\left[\sum_{j\in\mathcal{E}_\ell}(\alpha_j - \alpha_j^\ell)y_i y_j K(\mathbf{x}_i,\mathbf{x}_j) + y_i(\alpha_0 - \alpha_0^\ell) + \lambda_\ell\right] = 1 \qquad (21)$$

- Let $\delta_j = \alpha_j^\ell - \alpha_j$ then

$$\sum_{j\in\mathcal{E}_\ell}\delta_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j) + y_u\delta_0 = \lambda_\ell - \lambda,\ \forall i \in \mathcal{E}_\ell \qquad (22)$$

- Due to (19), we must have

$$\sum_{j\in\mathcal{E}_\ell} y_j\delta_j = 0 \qquad (23)$$

Regularization path of SVM VI

- Equations (18) and (19) constitute $m+1$ linear equations in $m+1$ unknown $\delta_j, j \in \{0\} \cup \mathcal{E}_\ell$ with $m = |\mathcal{E}_\ell|$:

$$\begin{aligned} \mathbf{K}_\ell^* \boldsymbol{\delta} + \delta_0 \mathbf{y}_\ell &= (\lambda_\ell - \lambda)\mathbb{1} \\ \mathbf{y}_\ell^T \boldsymbol{\delta} &= 0 \end{aligned}$$

- This yields

$$\mathbf{A}_\ell \boldsymbol{\delta}^a = (\lambda_\ell - \lambda)\mathbb{1}^a.$$

where

$$\mathbf{A}_\ell = \begin{pmatrix} 0 & \mathbf{y}_\ell^T \\ \mathbf{y}_\ell & \mathbf{K}_\ell^* \end{pmatrix}, \boldsymbol{\delta}^a = \begin{pmatrix} \delta_0 \\ \boldsymbol{\delta} \end{pmatrix}, \text{ and } \mathbb{1}^a = \begin{pmatrix} 0 \\ \mathbb{1} \end{pmatrix}$$

Regularization path of SVM VII

- If $\mathbf{A}_\ell$ is of full rank, we have

$$\boldsymbol{\delta}^a = (\lambda_\ell - \lambda)\mathbf{b}^a$$

  where $\mathbf{b}^a = \{b_j\} = \mathbf{A}_\ell^{-1}\mathbb{1}^a$.

- Finally, we have

$$\boldsymbol{\alpha}_j = \boldsymbol{\alpha}_j^\ell - (\lambda_\ell - \lambda)b_j, \quad j \in \{0\} \cup \mathcal{E}_\ell$$

- Thus, α is linear in λ when $\lambda \in [\lambda_{\ell+1}, \lambda_\ell]$, thus, piecewise linear in λ

Regularization path of SVM VIII

- `svmpath`-package in `R`.

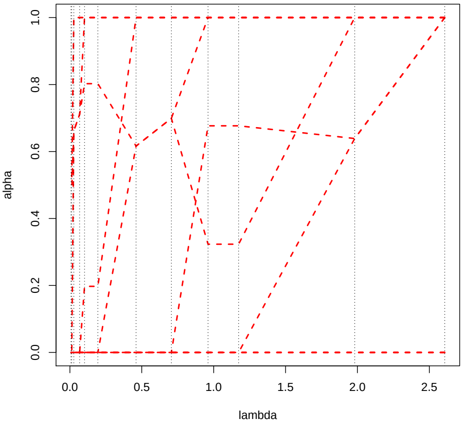


Figure: Illustration

Regularization path of SVM IX

- Piecewise linearity comes form the combination of L_1-type loss + L_2-tyep penalty.
- SVR, WSVM, KQR, and many other kernel machines possess this property.

Reference

- Cristianini & Shawe-Taylor (2000) An introduction to support vector machines and other kernel-based learning methods Cambridge University Press.
- Boyd & Vandenberghe (2004) Convex optimization Cambridge university press.
- Wahba (1990) Spline models for observational data SIAM.
- Takeuchi, Le, Sears & Smola (2006) Nonparametric quantile estimation JMLR 7, p.1231-1264.
- Zhu & Hastie (2002) Kernel logistic regression and the import vector machine In Advances in NIPS (pp. 1081-1088).