On Robustness Properties of Support Vector Machines Based on General Loss Functions

Arnout Van Messem Andreas Christmann

Vrije Universiteit Brussel UNIVERSITÄT BAYREUTH

Flexible Modelling: Smoothing and Robustness, Leuven, November 12-14, 2008

Introduction

Known

Support Vector Machines (SVMs) are L-risk consistent and robust, if Lipschitz continiuous loss function and bounded kernel are chosen.

Christmann & Van Messem, 2008, Steinwart & Christmann 2008; Christmann & Steinwart, 2007

Problem

Can the assumptions $f \in L_1(\mathrm{P}_X)$ and $\int |Y| \, d\mathrm{P}$ be weakened?

Notation

Assumptions:

- $X \subseteq \mathbb{R}^d$, $Y \subseteq \mathbb{R}$, $X \neq \varnothing$, $Y \neq \varnothing$
- $D = ((x_1, y_1), \ldots, (x_n, y_n))$, $1 \leq i \leq n$
- (X_i, Y_i) i.i.d. $\sim \mathrm{P} \in \mathcal{M}_1$, P (totally) unknown

Aim:

- $f(x_i)$ = quantity of interest of $\mathrm{P}_{Y_i|X_i=x_i}$
  e.g. conditional median for robust regression

Assumption:

- Loss function: $L : Y \times \mathbb{R} \to [0, \infty)$, $L(y_i, f(x_i))$, convex

Kernel methods

- **Kernel**: $k : X \times X \to \mathbb{R}$, if $\exists$ Hilbert space $\mathcal{H}$ and $\Phi : X \to \mathcal{H}$ such that

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle, \quad \forall x, x' \in X$$

Reproducing Kernel Hilbert Space (RKHS)

$\mathcal{H}$ a Hilbert space of functions $f : X \to \mathbb{R}$. A reproducing kernel for $\mathcal{H}$ is a kernel k with

$$f(x) = \langle f, k(x, \cdot) \rangle \quad \forall\, f \in \mathcal{H}, \forall\, x \in X.$$

- **Canonical feature map**: $\Phi(x) = k(x, \cdot),\ x \in X$
- $k \rightleftarrows$ RKHS unique
- **Bounded**: $||k||_\infty := \sqrt{\sup_{x \in \mathcal{X}} k(x, x)} < \infty$
- **Gaussian RBF**: $k(x, x') = e^{-\gamma ||x - x'||_2^2}$, $\gamma > 0$

Risk

Definition

Risk

$$\mathcal{R}_{L,\mathrm{P}}(f) = \mathbb{E}_{\mathrm{P}} L(Y, f(X))$$

Regularized risk

$$\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f) = \mathbb{E}_{\mathrm{P}} L(Y, f(X)) + \lambda \left\| f \right\|^2_{\mathcal{H}}$$

where $\mathrm{P} \in \mathcal{M}_1$, $\mathcal{H}$ is a RKHS and $\lambda > 0$.

Support Vector Machines

Definition

SVM operator

$$S(\mathrm{P}) = f_{L,\mathrm{P},\lambda} = \arg\min_{f \in \mathcal{H}} \mathbb{E}_{\mathrm{P}} L\big(Y_i, f(X_i)\big) + \lambda \|f\|_{\mathcal{H}}^2 ,$$

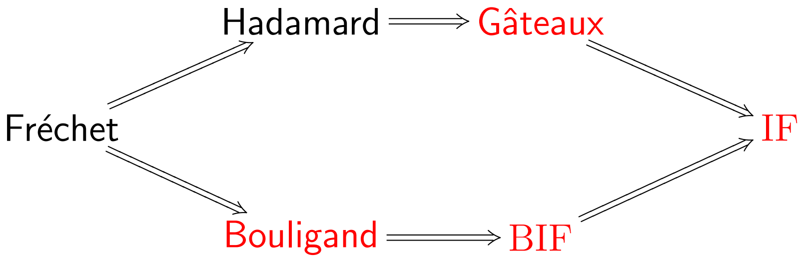
where $\mathrm{P} \in \mathcal{M}_1$, $\mathcal{H}$ is a RKHS and $\lambda > 0$.

SVM estimator

$$S(\mathrm{P}_n) = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L\big(Y_i, f(X_i)\big) + \lambda \|f\|_{\mathcal{H}}^2 ,$$

where $\mathrm{P}_n := \frac{1}{n} \sum_{i=1}^{n} \Delta_{(x_i, y_i)}$.

Approaches



Christmann & Van Messem (2008)

Bouligand differentiability

Bouligand-derivative

$f : X \to Z$ is **Bouligand-differentiable** at $x_0 \in X$, if $\exists$ a positive homogeneous function $\nabla^B f(x_0) : X \to Z$ such that

$$f(x_0 + h) = f(x_0) + \nabla^B f(x_0)(h) + o(h) ,$$

i.e.

$$\lim_{h \downarrow 0} \frac{\left\| f(x_0 + h) - f(x_0) - \nabla^B f(x_0)(h) \right\|_Z}{\|h\|_X} = 0.$$

Strong approximation

$f : X \to Z$ **strongly approximates** $F : X \times Y \to Z$ in x at (x_0, y_0) (notation: $f \approx_x F$) if $\forall \varepsilon > 0\ \exists$ neighborhoods $\mathcal{N}(x_0)$ of x_0 and $\mathcal{N}(y_0)$ of y_0 such that $\forall x, x' \in \mathcal{N}(x_0), \forall y \in \mathcal{N}(y_0)$

$$\left\| \big(F(x,y) - f(x)\big) - \big(F(x',y) - f(x')\big) \right\|_Z \leq \varepsilon \left\| x - x' \right\|_X .$$

Strong Bouligand-derivative

$F : X \times Y \to Z$ has partial B-derivative $\nabla_1^B F(x_0, y_0)$ w.r.t. x at (x_0, y_0). Then $\nabla_1^B F(x_0, y_0)$ is **strong** if

$$F(x_0, y_0) + \nabla_1^B F(x_0, y_0)(x - x_0) \approx_x F$$

at (x_0, y_0).

Robinson (1991)

The trick

Let $L : Y \times \mathbb{R} \to [0, \infty)$ be a loss function.

Definition

$L^* : Y \times \mathbb{R} \to \mathbb{R}$ with $L^*(y, t) := L(y, t) - L(y, 0)$.

Koenker, 2005; Huber, 1967; Bickel et al, 1993

L^* can be negative!

Properties

- L (strictly) convex, then L^* (strictly) convex.
- L Lipschitz w.r.t. 2^{nd} argument, then L^* Lipschitz w.r.t. 2^{nd} argument.

Reason

Reduce conditions for the existence of the risk

For L Lipschitz w.r.t. 2^{nd} argument

- $\mathbb{E}_{\mathrm{P}} L(Y, f(X)) < \infty$ if $f \in L_1(\mathrm{P}_X)$ and $Y \in L_1(P_{Y|X})$.
- $\mathbb{E}_{\mathrm{P}} L^*(Y, f(X)) < \infty$ if $f \in L_1(\mathrm{P}_X)$.

Properties

- L Lipschitz then

$$|\mathcal{R}_{L^*,\mathrm{P}}(f)| \le |L|_1 \mathbb{E}_{\mathrm{P}_X}|f(X)|.$$

$$|\mathcal{R}^{reg}_{L^*,\mathrm{P},\lambda}(f)| \le |L|_1 \mathbb{E}_{\mathrm{P}_X}|f(X)| + \lambda \|f\|^2_{\mathcal{H}}.$$

- L Lipschitz then

$$\|f_{L^*,\mathrm{P},\lambda}\|_{\mathcal{H}} \le \sqrt{\left(|L|_1 \mathbb{E}_{\mathrm{P}_X}|f_{L^*,\mathrm{P},\lambda}(X)|\right)/\lambda}.$$

- $\nabla^F_2 L^*(y,t) = \nabla^F_2 L(y,t)$ and $\nabla^B_2 L^*(y,t) = \nabla^B_2 L(y,t)$.

Uniqueness of SVM solution

Proposition

- L **Lipschitz continuous** w.r.t 2^{nd} argument,
- $f \in L_1(\mathrm{P}_X)$.

Then $\mathcal{R}_{L^*,\mathrm{P}}(f) \notin \{-\infty, +\infty\}$ and $\mathcal{R}^{reg}_{L^*,\mathrm{P},\lambda}(f) \neq -\infty$.

Theorem

- L **convex**,
- $\mathcal{H}$ RKHS of a measurable kernel k,
- $\mathcal{R}_{L^*,\mathrm{P}}(f) < \infty$ for some $f \in \mathcal{H}$,
- $\mathcal{R}_{L^*,\mathrm{P}}(f) > -\infty$ for all $f \in \mathcal{H}$.

Then $\forall\, \lambda > 0$ there exists at most one SVM solution $f_{L^*,\mathrm{P},\lambda}$.

Existence of SVM solution

Theorem

- L **Lipschitz continuous** w.r.t 2^{nd} argument and **convex**,
- $\mathcal{H}$ RKHS of a **bounded** measurable kernel k.

Then $\forall\, \lambda > 0$ there exists an SVM solution $f_{L^*,\mathrm{P},\lambda}$.

What if trick not needed

- Not needed if $\mathcal{R}_{L,\mathrm{P}}(0) = \mathbb{E}_{\mathrm{P}} L(Y,0) < \infty$.

$$\begin{aligned} \mathcal{R}^{reg}_{L^*,\mathrm{P},\lambda}(f_{L^*,\mathrm{P},\lambda}) &:= \inf_{f \in \mathcal{H}} \mathbb{E}_{\mathrm{P}}(L(Y,f(X)) - L(Y,0)) + \lambda \|f\|^2_{\mathcal{H}} \\ &= \inf_{f \in \mathcal{H}} [\mathbb{E}_{\mathrm{P}} L(Y,f(X)) + \lambda \|f\|^2_{\mathcal{H}}] - \mathbb{E}_{\mathrm{P}} L(Y,0). \end{aligned}$$

- Therefore

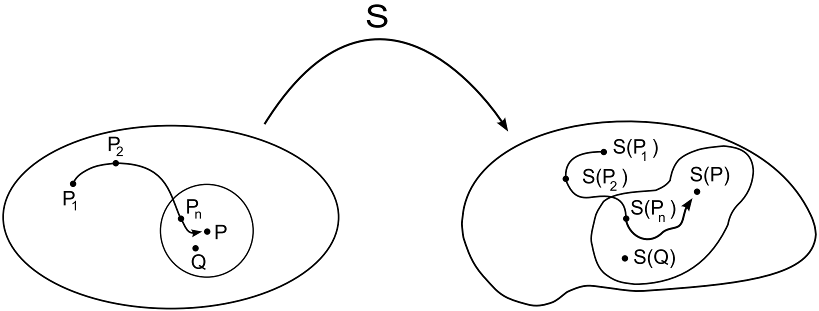
$$\mathcal{R}^{reg}_{L^*,\mathrm{P},\lambda}(f_{L^*,\mathrm{P},\lambda}) = \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f_{L,\mathrm{P},\lambda}) - \mathbb{E}_{\mathrm{P}} L(Y,0).$$

- $f_{L^*,\mathrm{P},\lambda}$ and $f_{L,\mathrm{P},\lambda}$ exist and unique, thus

$$f_{L^*,\mathrm{P},\lambda} = f_{L,\mathrm{P},\lambda}.$$

Robustness

1. What if (X_i, Y_i) i.i.d. $\sim \mathrm{P}$, $\mathrm{P} \in \mathcal{M}_1$ unknown is invalid?
2. What is the impact on $S(\mathrm{P}) = f_{L^*,\mathrm{P},\lambda}$?



Influence Function

Definition (Hampel, '68, Hampel et al. '86)

The influence function of S at P is given by

$$\mathrm{IF}(z; S, \mathrm{P}) := \lim_{\varepsilon \downarrow 0} \frac{S\big((1-\varepsilon)\mathrm{P} + \varepsilon \Delta_z\big) - S(\mathrm{P})}{\varepsilon},$$

in those z where this limit exists.

If Gâteaux derivative $\nabla^G(z; S, \mathrm{P})$ exists:

$\nabla^G = \mathrm{IF}$ and IF is linear and continuous

Goal: Bounded IF

Bouligand Influence Function

BIF (C&VM '08)

The **Bouligand influence function** (BIF) of a function $S : \mathcal{M}_1 \to \mathcal{H}$ for a distribution P in the direction of a distribution $\mathrm{Q} \neq \mathrm{P}$ is the special B-derivative (if it exists)

$$\lim_{\varepsilon \downarrow 0} \frac{\left\| S\big((1-\varepsilon)\mathrm{P} + \varepsilon \mathrm{Q}\big) - S(\mathrm{P}) - \mathrm{BIF}(\mathrm{Q}; S, \mathrm{P}) \right\|_{\mathcal{H}}}{\varepsilon} = 0.$$

If BIF exists and $\mathrm{Q} = \Delta_z$: IF exists and BIF = IF (C&VM '08)

Goal: Bounded BIF

Result IF

Assumptions

- $\mathcal{H}$ is RKHS with **bounded**, measurable kernel k,
- How do we put $f \in L_1(\mathrm{P}_X)$ in the assumptions or is $f_{L^*,\mathrm{P},\lambda} \in \mathcal{H}$ sufficient?,
- $L : Y \times \mathbb{R} \to [0, \infty)$ **convex** and **Lipschitz continuous** w.r.t. the 2^{nd} argument with uniform Lipschitz constant $|L|_1 := \sup_{y \in Y} |L(y, \cdot)|_1 \in (0, \infty)$,
- $\mathrm{P} \in \mathcal{M}_1(X \times Y)$.

Theorem IF

Then $\mathrm{IF}(z; S, \mathrm{P})$ with $S(\mathrm{P}) := f_{L^*,\mathrm{P},\lambda}$ and $z := (x, y)$

1. exists and
2. equals

$$T^{-1}\Big(\mathbb{E}_{\mathrm{P}} \nabla_2^F L^*(Y, f_{L^*,\mathrm{P},\lambda}(X))\Phi(X)\Big) \\ -\nabla_2^F L^*(y, f_{L^*,\mathrm{P},\lambda}(x))T^{-1}\Phi(x)\,,$$

where $T : \mathcal{H} \to \mathcal{H}$ with
$T = 2\lambda\,\mathrm{id}_{\mathcal{H}} + \mathbb{E}_{\mathrm{P}} \nabla_{2,2}^F L^*(Y, f_{L^*,\mathrm{P},\lambda}(X))\langle\Phi(X), \cdot\rangle_{\mathcal{H}}\Phi(X)$.

Sketch of proof

- Define
  $G(\varepsilon, f) := 2\lambda f + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\Delta_z} \nabla_2^F L^*(Y, f(X))\Phi(X)$
- $G(\varepsilon, f)$ fulfills the conditions for an implicit function theorem

$$G(\varepsilon, f) = \frac{\partial \mathcal{R}^{reg}_{L^*,(1-\varepsilon)\mathrm{P}+\varepsilon\Delta_z,\lambda}}{\partial \mathcal{H}}(f) = \nabla_2^F \mathcal{R}^{reg}_{L^*,(1-\varepsilon)\mathrm{P}+\varepsilon\Delta_z,\lambda}(f)\,,$$

$$\varepsilon \in [0, 1]$$

Result BIF

Assumptions

- $X \subset \mathbb{R}^d$, $Y \subset \mathbb{R}$ closed sets,
- $\mathcal{H}$ is RKHS with **bounded**, measurable kernel k,
- $f_{L^*,\mathrm{P},\lambda} \in \mathcal{H}$,
- $L : Y \times \mathbb{R} \to [0, \infty)$ **convex** and **Lipschitz continuous** w.r.t. the 2^{nd} argument with uniform Lipschitz constant $|L|_1 := \sup_{y \in Y} |L(y, \cdot)|_1 \in (0, \infty)$,
- L has measurable partial B-derivatives w.r.t. the 2^{nd} argument with $\kappa_1 := \sup_{y \in Y} \left\| \nabla_2^B L(y, \cdot) \right\|_\infty \in (0, \infty)$, $\kappa_2 := \sup_{y \in Y} \left\| \nabla_{2,2}^B L(y, \cdot) \right\|_\infty < \infty$,

Assumptions

- $\delta_1 > 0$, $\delta_2 > 0$,
- $\mathcal{N}_{\delta_1}(f_{L^*,\mathrm{P},\lambda}) := \{f \in \mathcal{H}; \|f - f_{L^*,\mathrm{P},\lambda}\|_{\mathcal{H}} < \delta_1\}$,
- $\lambda > \frac{1}{2}\kappa_2 \|\Phi\|_{\mathcal{H}}^3$, (Note: $\kappa_2 = 0$ for L_ϵ, L_τ)
- P, Q probability measures on $\big(X \times Y, \mathcal{B}(X \times Y)\big)$ with $\mathbb{E}_{\mathrm{P}}|Y| < \infty$ and $\mathbb{E}_{\mathrm{Q}}|Y| < \infty$.
- Define $G : (-\delta_2, \delta_2) \times \mathcal{N}_{\delta_1}(f(L^*, \mathrm{P}, \lambda)) \to \mathcal{H}$,

$$G(\varepsilon, f) := 2\lambda f + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} \nabla_2^B L^*(Y, f(X))\Phi(X)\,,$$

- $G(0, f_{L^*,\mathrm{P},\lambda}) = 0$ and $\nabla_2^B G(0, f_{L^*,\mathrm{P},\lambda})$ is **strong**.

Theorem BIF

Then $\mathrm{BIF}(\mathrm{Q}; S, \mathrm{P})$ with $S(\mathrm{P}) := f_{L^*,\mathrm{P},\lambda}$

1. exists,
2. equals

$$T^{-1}\Big(\mathbb{E}_{\mathrm{P}}\nabla_2^B L^*(Y, f_{L^*,\mathrm{P},\lambda}(X))\Phi(X) - \mathbb{E}_{\mathrm{Q}}\nabla_2^B L^*(Y, f_{L^*,\mathrm{P},\lambda}(X))\Phi(X)\Big),$$

   where $T : \mathcal{H} \to \mathcal{H}$ with
   $T = 2\lambda\,\mathrm{id}_{\mathcal{H}} + \mathbb{E}_{\mathrm{P}}\nabla_{2,2}^B L^*(Y, f_{L^*,\mathrm{P},\lambda}(X))\langle\Phi(X), \cdot\rangle_{\mathcal{H}}\Phi(X)$,
   and
3. is bounded.

Sketch of proof

- $G(\varepsilon, f)$: $\nabla_2^B L^*(Y, f(X)) = \nabla_2^B L(Y, f(X))$ hence

$$G(\varepsilon, f) = 2\lambda f + \mathbb{E}_{(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q}} \nabla_2^B L(Y, f(X))\Phi(X)$$

  $\longrightarrow$ proof identical as in C&VM '08
- $G(\varepsilon, f)$ fulfills the conditions of Robinson's (1991) implicit function theorem on Bouligand-derivatives

$$G(\varepsilon, f) = \frac{\partial \mathcal{R}^{reg}_{L^*,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}}{\partial \mathcal{H}}(f) = \nabla_2^B \mathcal{R}^{reg}_{L^*,(1-\varepsilon)\mathrm{P}+\varepsilon\mathrm{Q},\lambda}(f)\,,\ \varepsilon \in [0,1]$$

Conclusions

Support Vector Machines based on $L^* := L - L(\cdot, 0)$ fulfill

- Weakens assumptions on P: only $f \in L_1(\mathrm{P}_X)$ is needed
- Existence and uniqueness of $f_{L^*,\mathrm{P},\lambda}$
- $\mathbb{E}_{\mathrm{P}} L(Y, 0) < \infty \quad \Longrightarrow \quad f_{L^*,\mathrm{P},\lambda} = f_{L,\mathrm{P},\lambda}$

Robustness

- Existence of IF and BIF
- Robust: $\mathrm{BIF}(\mathrm{Q}; T, \mathrm{P})$ bounded for regression if $\nabla_2^B L$ and k bounded

References

- Christmann & Van Messem (2008). Bouligand derivatives and robustness of support vector machines for regression. *Journal of Machine Learning Research*, **9**, 915-936.
- Steinwart & Christmann (2008). Support Vector Machines. Springer, New York.
- Christmann & Steinwart (2007). Consistency and robustness of kernel based regression. *Bernoulli*, **13**, 799-819.
- Hampel (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, **69**, 383-393.
- Schölkopf & Smola (2002). Learning with kernels. MIT Press.
- Koenker (2005). Quantile regression. Cambridge University Press.
- Huber (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the* 5^{th} *Berkeley Symposium*.

Sketch: Proof for IF

For the proof of the theorem about the BIF we showed:

i. $G(0, f) = 0 \Leftrightarrow f = f_{L^*,\mathrm{P},\lambda}$.

ii. G continuously F-differentiable.

iii. $\frac{\partial G}{\partial \mathcal{H}}(0, f_{L^*,\mathrm{P},\lambda})$ invertible.

iv. Then there exist $\delta > 0$, a neighborhood $\mathcal{N}_\delta(f_{L^*,\mathrm{P},\lambda}) := \{f \in \mathcal{H};\, \|f - f_{L^*,\mathrm{P},\lambda}\|_{\mathcal{H}} < \delta\}$, and a function $f^* : (-\delta, \delta) \to \mathcal{N}_\delta(f_{L^*,\mathrm{P},\lambda})$ satisfying

iv.1) $f^*(0) = f_{L^*,\mathrm{P},\lambda}$.

iv.2) It holds

$$\nabla^F f^*(0) = -\left(\nabla_2^F G(0, f_{L^*,\mathrm{P},\lambda})\right)^{-1} - \nabla_1^B G(0, f_{L^*,\mathrm{P},\lambda}).$$

Sketch: Proof for BIF

For the proof of the theorem about the BIF we showed:

i. For some χ and each $f \in \mathcal{N}_{\delta_1}(f_{L^*,\mathrm{P},\lambda})$, $G(\cdot\,, f)$ is Lipschitz continuous on $(-\delta_2, \delta_2)$ with Lipschitz constant χ.

ii. G has partial B-derivatives with respect to ε and f at $(0, f_{L^*,\mathrm{P},\lambda})$.

iii. $\nabla_2^B G(0, f_{L^*,\mathrm{P},\lambda})\big(\mathcal{N}_{\delta_1}(f_{L^*,\mathrm{P},\lambda}) - f_{L^*,\mathrm{P},\lambda}\big)$ is a neighborhood of $0 \in \mathcal{H}$.

iv. $\delta\big(\nabla_2^B G(0, f_{L^*,\mathrm{P},\lambda}), \mathcal{N}_{\delta_1}(f_{L^*,\mathrm{P},\lambda}) - f_{L^*,\mathrm{P},\lambda}\big) \;=:\; d_0 > 0.$

v. For each $\xi > d_0^{-1}\chi$ there exist $\delta_3, \delta_4 > 0$, a neighborhood $\mathcal{N}_{\delta_3}(f_{L^*,\mathrm{P},\lambda}) := \{f \in \mathcal{H}; \|f - f_{L^*,\mathrm{P},\lambda}\|_{\mathcal{H}} < \delta_3\}$, and a function $f^* : (-\delta_4, \delta_4) \to \mathcal{N}_{\delta_3}(f_{L^*,\mathrm{P},\lambda})$ satisfying

- **v.1)** $f^*(0) = f_{L^*,\mathrm{P},\lambda}$.
- **v.2)** $f^*(\cdot)$ is Lipschitz continuous on $(-\delta_4, \delta_4)$ with Lipschitz constant $|f^*|_1 = \xi$.
- **v.3)** For each $\varepsilon \in (-\delta_4, \delta_4)$ is $f^*(\varepsilon)$ the unique solution of $G(\varepsilon, f) = 0$ in $(-\delta_4, \delta_4)$.
- **v.4)** It holds $\nabla^B f^*(0)(u) = \left(\nabla_2^B G(0, f_{L^*,\mathrm{P},\lambda})\right)^{-1} \left(-\nabla_1^B G(0, f_{L^*,\mathrm{P},\lambda})(u)\right)$.