Bouligand Derivatives and Robustness of Support Vector Machines

Arnout Van Messem
joint work with Andreas Christmann

Vrije Universiteit Brussel

Radon Conference on Financial and Actuarial Mathematics for Young Researchers, Linz, May 30-31, 2007

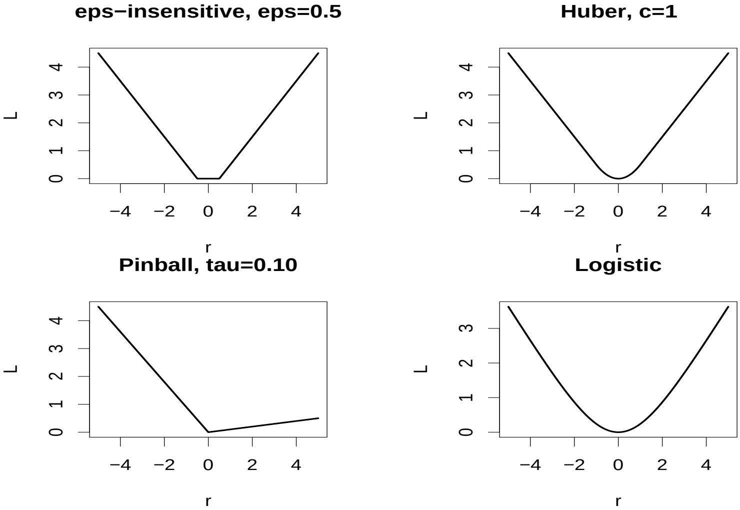
Notation

- Data sample: $\mathcal{D} = ((x_1, y_1), \ldots, (x_n, y_n)) \in Z := X \times Y$, $1 \leq i \leq n$, $n \in \mathbb{N}$
- $X \subseteq \mathbb{R}^d$, $Y \subseteq \mathbb{R}$, closed, $X \neq \varnothing$, $Y \neq \varnothing$
- $f(x_i)$ = quantity of interest of $\mathrm{P}_{Y_i|X_i=x_i}$
- Loss function: $L : Y \times \mathbb{R} \to [0, \infty)$, $L(y_i, f(x_i))$, convex
- Assumption: (X_i, Y_i) i.i.d. $\sim \mathrm{P} \in \mathcal{M}_1$, P unknown

Loss functions for regression

Method	*Loss*, $r := y - f(x)$
ϵ-insensitive	$L_\epsilon(y, f(x)) = \max\{0, \lvert r\rvert - \epsilon\}$
Huber, $c \in (0, \infty)$	$L_{Huber}(y, f(x)) = r^2/2$ if $\lvert r\rvert \leq c$ $= c\lvert r\rvert - c^2/2$ if $\lvert r\rvert > c$
Pinball, $\tau \in (0, 1)$	$L_\tau(y, f(x)) = (\tau - 1)r$ if $r < 0$ $= \tau r$ if $r \geq 0$
Logistic	$L_{log}(y, f(x)) = -\log(4\Lambda(r)[1 - \Lambda(r)])$ $\Lambda(r) := 1/[1 + \exp(-r)]$
Least Squares	$L_{LS}(y, f(x)) = r^2$
$L1$	$L_{L1}(y, f(x)) = \lvert r\rvert$

Loss functions



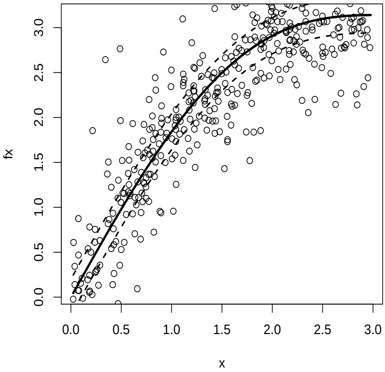
Kernel methods

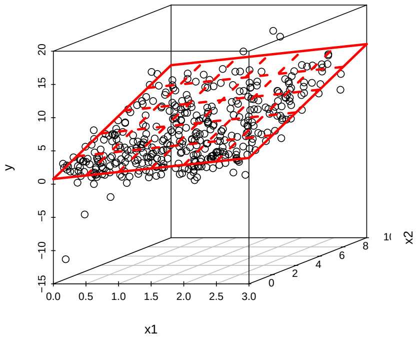
Reproducing Kernel Hilbert Space

Let $\mathcal{H}$ be a Hilbert space of functions $f : X \to \mathbb{R}$. A reproducing kernel for $\mathcal{H}$ is a map $k : X \times X \to \mathbb{R}$ with $\Phi(x) := k(x, \cdot) \in \mathcal{H}$, $f(x) = \langle f, k(x, \cdot) \rangle \; \forall \, x \in X, f \in \mathcal{H}$.

- $k(x, x') = \langle \Phi(x), \Phi(x') \rangle, \forall x, x'$
- $k \rightleftarrows$ RKHS unique
- Bounded: $||k||_\infty := \sqrt{\sup_{x \in \mathcal{X}} k(x, x)} < \infty$
- GRBF: $k(x, x') = e^{-\gamma ||x - x'||_2^2}$, $\gamma > 0$

Example for feature map $\Phi(\mathbf{x}) = \mathbf{k}(\mathbf{x}, \cdot)$





Empirical Risk Minimization (ERM)

- Vapnik '98
- $\mathcal{R}_{L,\mathrm{P}}(f) := \mathbb{E}_{\mathrm{P}} L(Y, f(X))$
- $\mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f) := \mathcal{R}_{L,\mathrm{P}}(f) + \lambda ||f||^2_{\mathcal{H}}$, $\lambda \in (0, \infty)$ fixed
- $\mathrm{P}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{(x_i, y_i)}$
- KBR estimator: $S(\mathrm{P}_n) = f_{\mathrm{P}_n,\lambda} = \arg\min_{f \in \mathcal{H}} \mathcal{R}^{reg}_{L,\mathrm{P}_n,\lambda}(f)$
  L convex, $\mathcal{H}$ RKHS with reprod. kernel k, $\lambda > 0$
- $f_{\mathrm{P}_n,\lambda}(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$.
  If $\alpha_i \neq 0$: x_i is support vector.
- KBR functional: $S(\mathrm{P}) := f_{\mathrm{P},\lambda} = \arg\min_{f \in \mathcal{H}} \mathcal{R}^{reg}_{L,\mathrm{P},\lambda}(f)$

Learnability of SVMs

- Universal (weak) consistency:
  $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P}_n}) \xrightarrow{\mathrm{P}} \inf_{f \in \mathcal{H}} \mathcal{R}_{L,\mathrm{P}}(f)$

- $L-$risk consistency:
  $\mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P}_n,\lambda_n}) \xrightarrow{\mathrm{P}} \mathcal{R}_{L,\mathrm{P}}$, where
  $\mathcal{R}_{L,\mathrm{P}} := \inf_{f:\mathcal{X}\to\mathbb{R} \text{ measurable}} \mathcal{R}_{L,\mathrm{P}}(f)$ for suitable $\lambda_n \downarrow 0$

C&S 2007

Question

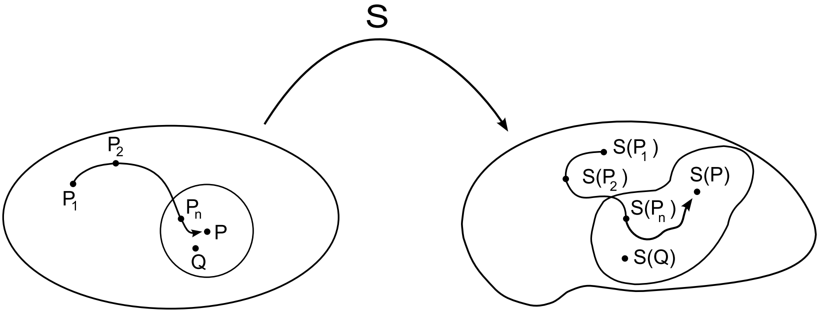
"Which properties must

- the map $S(\mathrm{P}) = f_{\mathrm{P},\lambda}$,
- the kernel k,
- and the loss function L

have for good robustness properties of ERM?"

Robustness

What is the impact on $S(\mathrm{P}) = f_{\mathrm{P},\lambda}$ due to violations from (X_i, Y_i) i.i.d. $\sim \mathrm{P}$, $\mathrm{P} \in \mathcal{M}_1$ unknown ?



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.

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

Strong approximation

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

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

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)

Bouligand influence function

BIF (C&VM '07)

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)

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

Goal: Bounded BIF

Main result

Assumptions

- $X \subset \mathbb{R}^d$, $Y \subset \mathbb{R}$ closed sets,
- $\mathcal{H}$ is RKHS with bounded, measurable kernel k,
- $f_{\mathrm{P},\lambda} \in \mathcal{H}$,
- $L : Y \times \mathbb{R} \to [0, \infty)$ 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. to 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

- $\delta_1 > 0$, $\delta_2 > 0$,
- $\mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda}) := \{f \in \mathcal{H};\, \|f - f_{\mathrm{P},\lambda}\|_{\mathcal{H}} < \delta_1\}$,
- $\lambda > \frac{1}{2}\kappa_2 \left\|\Phi\right\|_{\mathcal{H}}^3$,
- 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 $G : (-\delta_2, \delta_2) \times \mathcal{N}_{\delta_1}(f(\mathrm{P}, \lambda)) \to \mathcal{H}$,

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

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

Theorem (C&VM '07)

The Bouligand influence function of $S(\mathrm{P}) := f_{\mathrm{P},\lambda}$ in direction of Q exists and

$$\begin{aligned}\mathrm{BIF}(\mathrm{Q}; S, \mathrm{P}) &= T^{-1}\Big(\mathbb{E}_{\mathrm{P}}\big(\nabla_2^B L(Y, f_{\mathrm{P},\lambda}(X))\Phi(X)\big) \\ &\qquad -\mathbb{E}_{\mathrm{Q}}\big(\nabla_2^B L(Y, f_{\mathrm{P},\lambda}(X))\Phi(X)\big)\Big),\end{aligned}$$

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

Examples

The assumptions of the theorem are valid and thus $\mathrm{BIF}(\mathrm{Q}; S, \mathrm{P})$ exists and is bounded, if

ϵ-insensitive loss L_ϵ, pinball loss L_τ

$\forall \delta > 0 \, \exists$ positive constants ξ_{P}, ξ_{Q}, c_{P}, and c_{Q} such that $\forall t \in \mathbb{R}$ with $|t - f_{\mathrm{P},\lambda}(x)| \leq \delta \|k\|_\infty$ the following inequalities hold $\forall a \in [0, 2\delta\|k\|_\infty]$ and $\forall x \in X$:

$$\mathrm{P}\big(Y \in [t, t+a] \,\big|\, x\big) \leq c_{\mathrm{P}} a^{1+\xi_{\mathrm{P}}}$$
$$\mathrm{Q}\big(Y \in [t, t+a] \,\big|\, x\big) \leq c_{\mathrm{Q}} a^{1+\xi_{\mathrm{Q}}} .$$

The assumptions of the theorem are valid and thus $\mathrm{BIF}(\mathrm{Q}; S, \mathrm{P})$ exists and is bounded, if

Huber loss L_{Huber}

$$\begin{aligned} \forall x \in X: & \\ & \mathrm{P}\big(Y \in \{f_{\mathrm{P},\lambda}(x) - c, f_{\mathrm{P},\lambda}(x) + c\} \,\big|\, x\big) \\ = & \ \mathrm{Q}\big(Y \in \{f_{\mathrm{P},\lambda}(x) - c, f_{\mathrm{P},\lambda}(x) + c\} \,\big|\, x\big) \\ = & \ 0\,. \end{aligned}$$

Logistic loss L_{log}

No special assumptions on the probabilities needed.

Project

Description

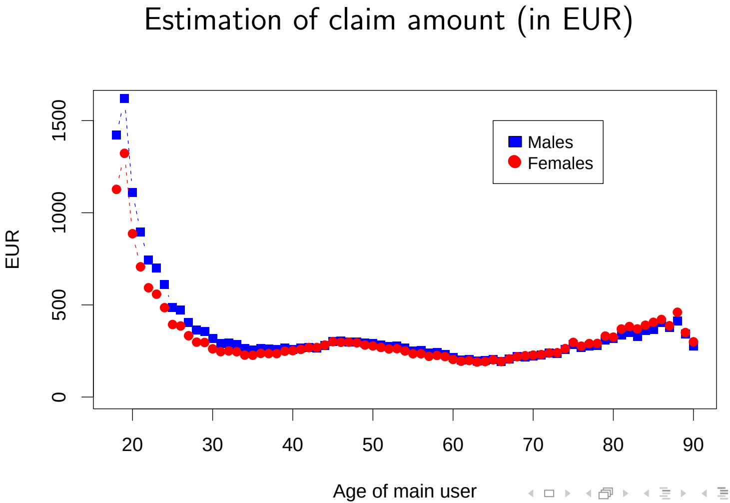
- cooperation partner: union of 15 German insurance companies (Verband öffentlicher Versicherer, Düsseldorf)
- $n \approx 4.5$ million customers
- primary goal: model the claim amount (in EUR)
- secondary goal: model the probability for a claim
- many potential risk variables and complex dependency structures

Questions

- Can we improve the current insurance tariff?
- Which groups of persons have especially often claims?
- Does the data set contain unknown information?
- If yes, how can we extract and model the information?
- Can we develop a method which is able to "learn" to do this automatically?

Arnout Van Messem - avmessem@vub.ac.be

Project


Estimation of claim amount (in EUR)
EUR
0
500
1000
1500
20
30
40
50
60
70
80
90
Age of main user
Males
Females

Summary

Kernel based methods like SVM:

- Non-parametric, flexible
- Can model complex high-dimensional dependency structures
- Robustness of SVM:
  - Bounded $\mathrm{BIF}(\mathrm{Q}; T, \mathrm{P})$
  - Robustness for regression if $\nabla_2^B L$ and k bounded
- Applications: insurance tariffs, credit scoring in banks, fraud detection, data mining, genomics, ...

References

- Christmann & Van Messem (2007). Bouligand derivatives and robustness of support vector machines. Submitted.
- Christmann & Steinwart (2004). Robust properties of convex risk minimization methods for pattern recognition. *JMLR*, **5**, 1007-1034.
- Christmann & Steinwart (2007). Consistency and robustness of kernel based regression. To appear: *Bernoulli*.
- Robinson (1991). An implicit-function theorem for a class of non-smooth functions. *Mathematics of Operations Research*, **16**, 292-309.
- Schölkopf & Smola (2002). Learning with kernels. MIT Press.
- Vapnik (1998). Statistical learning theory. Wiley.

More on the theorem

For the proof of the theorem we showed:

i. For some χ and each $f \in \mathcal{N}_{\delta_1}(f_{\mathrm{P},\lambda})$, $G(\cdot\,, f)$ 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 $(0, f_{\mathrm{P},\lambda})$.

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

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

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

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