METHODEN DES MASCHINELLEN LERNENS FÜR DATEN AUS DER VERSICHERUNGSWIRTSCHAFT

ANDREAS CHRISTMANN
Universität Dortmund, Fachbereich Statistik
christmann@statistik.uni-dortmund.de

DoMuS Kolloquium und Vollversammlung, Dortmund, 24.11.2003

Forschungsband
Modellbildung und Simulation

Vernetzung

Forschungsband

Modellbildung und Simulation

⇕ ⇕ ⇕

Mathematik ⇔ Statistik ⇔ Informatik

Statistik + Informatik	Statistische Methoden und Maschinelle Lernverfahren → Prof. Weihs , Prof. Morik
Statistik + Mathematik	Robuste Modellbildung und Dimensionsreduktion → Prof. Gather , Prof. Davies (Essen)
Statistik + Informatik	Komplexität und Algorithmen in der Statistik → Prof. Gather , Prof. Wegener
Statistik + Mathematik	Risikodifferenzierung in hochdimensionalen Datenstrukturen → Chr. & Dr. Kovac (Essen)

Eingeworbene Projekte und Drittmittel

- Teilprojekt B7 im Rahmen des SFB-475
  'Risikodifferenzierung in hochdimensionalen Datenstrukturen'
  Projektleiter: Chr. und Dr. Kovac (Essen, Mathematik)
  bewilligt für 3 Jahre
  Kooperationsvertrag mit dem Verband öffentlicher Versicherer in Düsseldorf
- Projekt im neubewilligten Graduiertenkolleg 'Statistische Modellbildung'
  zum Thema 'Robustheitsuntersuchungen der Support Vector Machine'

CONTENTS

ZDF: 6. November 2003

Optimieren der KFZ-Versicherung

... Seit über 20 Jahren sind die Tarifmerkmale in der Fahrzeugversicherung (Kasko) nahezu unverändert. Jetzt wird der Kaskoschutz 'runderneuert'. Wichtigste Änderung: Die Typklassen-Struktur wird mit statistischen Verfahren neu ermittelt und berücksichtigt wie die KFZ-Haftpflichtversicherung nun auch Merkmale wie zum Beispiel Fahrleistung und Garage. ...

(Thomas J. Kramer)

1. APPLICATION: INSURANCE TARIFFS

- Project in SFB-475 (with A. Kovac)
- Verband öffentlicher Versicherer, Düsseldorf, Germany
  non-aggregated data from 15 insurance companies:
  3 GB, > 6 millions obs., > 70 explanatory variables, many discrete
- What is the expected claim amount? $\Rightarrow$ insurance tariffs
- What is the probability of a claim?
- complex dependencies, empty cells, missing values
- some extreme high costs

Statistical Objectives

- Y claim size [year], x vector of explanatory variables
- Actual premium charged to the customer:
  pure premium + safety loading + administrative costs + desired profit
- Primary response: Pure premium. $E(Y|X = x)$
- Secondary response: Prob. of claim. $P(Y > 0|X = x)$

EXPLORATORY DATA ANALYSIS

Cost per policy holder per year: total mean ≈ 360 EUR

Cost	% obs.	% of sum	Mean	Median	SD	SD/Mean
total			364	0	21996	60.43
0	94.9	0	0	0	0	
(0,2000]	2.2	6.7	1097	1110	520	0.47
(2000,10000]	2.4	27.1	4156	3496	1940	0.47
(10000,50000]	0.4	19.8	18443	15059	8911	0.48
>50000	0.07	46.4	234621	96417	784365	3.34

Maximum cost: > 27 Million EUR !

2. METHODS

- Classical approach (in Germany): 'Marginal Sum Model'
  $\longrightarrow$ Poisson-Regression

- Generalized Linear Models (GLIM):
  - ▶ Y_i has distribution from exponential family
  - ▶ $\mathrm{E}(Y_i) = \mu_i = g^{-1}(x_i'\beta)$ and $\mathrm{Var}(Y_i) = \phi\, V(g^{-1}(x_i'\beta))/w_i$
    g = link function, V = variance function

  Special cases:
  - ▶ Poisson: $g = \log$, $V = id$, i.e. $\mathrm{E}(Y_i) = \mathrm{Var}(Y_i) = e^{x_i'\theta}$
  - ▶ Gamma: $g(\mu_i) = 1/\mu_i$, $g = \log$, $V(\mu_i) = \mu_i^2$
  - ▶ Inverse Gaussian: $g(\mu_i) = \mu_i^{-2}$, $V(\mu_i) = \mu_i^3$
  - ▶ Negative Binomial: $g(\mu_i) = \log(\mu_i)$, $V(\mu_i) = \mu_i + k\mu_i^2$

- Tweedie's compound Poisson Model (Smyth & Jørgensen, '02):
  No. of claims $\sim$ Poisson, size of each claim $\sim$ Gamma. Double GLIM.

PROPOSAL

Denote: Y claim size [year], x vector of explanatory variables

1: Determine $(k+2)$ classes for Y, e.g.

$C=0$ if $Y=0$	'no cost'	94.9	%
$C=1$ if $Y \in (0, 2000]$	'low cost'	2.2	%
$C=2$ if $Y \in (2000, 10000]$	'medium cost'	2.4	%
$C=3$ if $Y \in (10000, 50000]$	'high cost'	0.4	%
$C=4$ if $Y > 50000$	'extreme cost'	0.07	%

2: No claims: $E(Y|X=x, C=0) \equiv 0$ for all x.

$$
\begin{aligned}
& E(Y|X=x) \\
= \; & P(C=k+1) \cdot E(Y|X=x, C=k+1) + \\
& P(C \neq k+1) \cdot \sum_{c=1}^{k} P(C=c|X=x) \cdot E(Y|X=x, C=c)
\end{aligned}
$$

$$E(Y|X=x)$$
$$= P(C=k+1)\cdot E(Y|X=x,C=k+1) +$$
$$P(C\neq k+1)\cdot \sum_{c=1}^{k} P(C=c|X=x)\cdot E(Y|X=x,C=c)$$

- Companies have interest in $P(C=c|X=x)$ or $E(Y|X=x,C=c)$
- Circumvents the problem: most $y_i = 0$, but $P(Y=0)=0$ for many classical approaches: Gamma, Log-Normal, ...
- Reduction of computation time possible
  – approx. 95% of obs. have no claims !
  – Regression estimates only necessary for 5% of obs. !

$$
\begin{aligned}
& E(Y|X=x) \\
= \; & P(C=k+1)\cdot E(Y|X=x, C=k+1) + \\
& P(C\neq k+1)\cdot \sum_{c=1}^{k} P(C=c|X=x)\cdot E(Y|X=x, C=c)
\end{aligned}
$$

- Reduction of interactions possible
- Variable selection: different vectors x for different C groups are possible
- Different estimation techniques can be used for estimating $P(C=c|X=x)$ and $E(Y|X=x, C=c)$, e.g.
  $\longrightarrow$ Multinomial logistic regression + Gamma regression
  $\longrightarrow$ Kernel logistic regression + $\varepsilon-$SVR
  $\longrightarrow$ Classification trees + semiparametric regression
  $\longrightarrow$ extreme value theory based on GPD for extreme claim amounts
  $\longrightarrow$ combination of pairs given above, where additional explanatory variables are constructed via classification and regression trees

3. CONVEX RISK MINIMIZATION BASED ON KERNELS

Vapnik '98

data set $(x_i, y_i) \in \mathbb{R}^p \times \{-1, +1\}$, assume (X_i, Y_i) i.i.d. $\mathbb{P}$, $\mathbb{P}$ unknown,
predictor $\hat{f}(x)$, classifier $\text{sign}(\hat{f}(x) + \hat{b})$, loss function $L(y, f(x) + b)$

Goal: $\arg\min_f \, \mathrm{E}_{\mathbb{P}} L(Y, f(X) + b)$

1^{st} idea: $\arg\min_f \frac{1}{n} \sum_{i=1}^n I(y_i, f(x_i) + b)$

but: $I(y, f + b)$ not convex, problem often NP-hard (Höffgen et al. '95)

2^{st} idea: minimize regularized empirical risk

$$(\hat{f}_{n,\lambda}, \hat{b}_{n,\lambda}) = \arg\min_{f \in H,\, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i) + b) + \lambda \|f\|_H^2,$$

where L convex, $\lambda > 0$ regularization parameter,

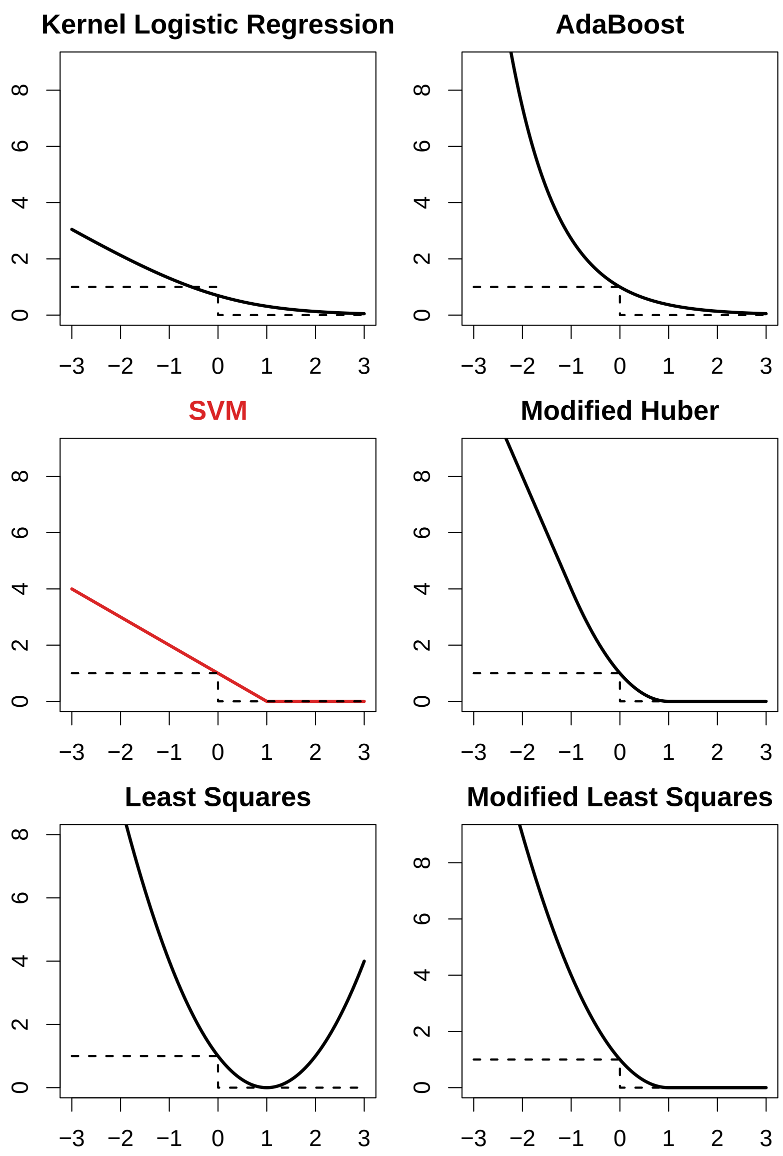
H reproducing kernel Hilbert space (RKHS) of kernel k

reg. theor. risk: $(f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda}) = \arg\min_{f \in H,\, b \in \mathbb{R}} \mathrm{E}_{\mathbb{P}} L(Y, f(X) + b) + \lambda \|f\|_H^2$

Vapnik '98, Zhang '01, Steinwart '02: universal consistency

Special loss functions:

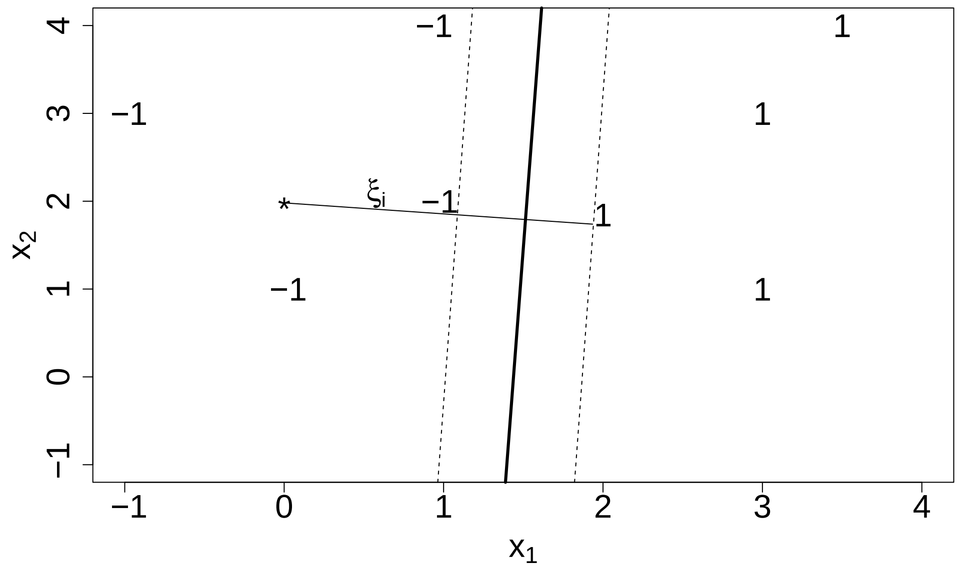
Method	L, $v = y(f(x) + b)$
Kernel Logistic Regression	$\ln(1 + \exp(-v))$
AdaBoost	$\exp(-v)$
Support Vector Machine	$\max(1 - v, 0)$
Modified Huber	$-4v$, if $v < -1$
	$\max(1 - v, 0)^2$, else
Least Squares	$(1 - v)^2$
Modified Least Squares	$\max(1 - v, 0)^2$



Vapnik ’98,
Schölkopf & Smola ’02,
Freund & Schapire ’96,
Friedman, Hastie & Tibshirani ’00,
Hastie, Tibshirani & Friedman ’01,
Suykens et al. ’02,
Zhang ’01, . . .

SUPPORT VECTOR MACHINE (SVM)

Special case: pattern recognition with linear kernel $f(x) = x'\theta$



Optimization problem:

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}\,||\theta||^2 + C\tfrac{1}{n}\textstyle\sum_i \xi_i \\ &\text{subject to:} && \mathbf{x}_i'\theta + b \geq +1 - \xi_i \quad \text{if } y_i = +1 \\ & && \mathbf{x}_i'\theta + b \leq -1 + \xi_i \quad \text{if } y_i = -1 \\ & && \xi_i \geq 0 \end{aligned}$$

Corresponding dual program is convex & quadratic !

$$\arg\min \quad \frac{1}{2}\alpha' Q \alpha - \alpha' \mathbf{1}$$

s.t.: $\frac{1}{n}\sum_i \alpha_i y_i = 0$

$\alpha_i \in [0, C]$

where $(Q)_{ij} = y_i y_j \, k(x_i, x_j)$, $Q \in \mathbb{R}^{n \times n}$!

linear kernel: $k(x_i, x_j) = x_i' x_j$

RBF kernel: $k(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2)$, $\gamma > 0$, $\quad u := ||x_i - x_j||$

Software:

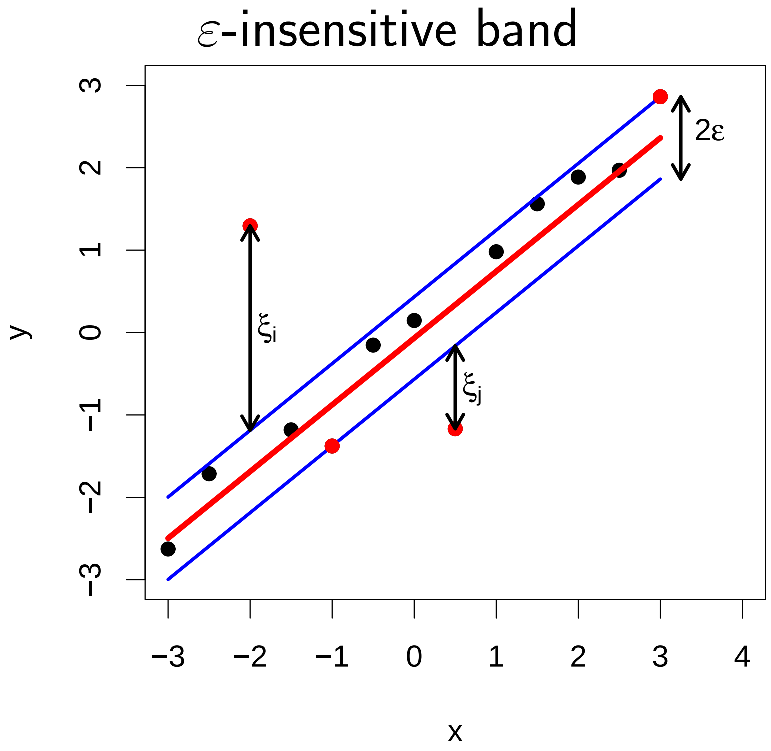
many computational tricks: e.g. Sequential Minimization Optimization (SMO)

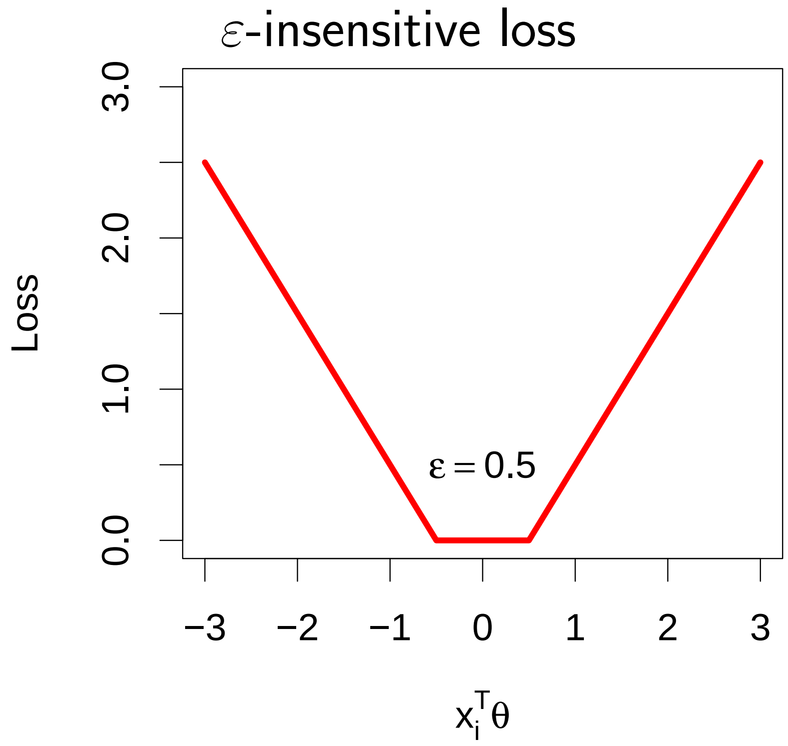
Overview: `http://www.kernel-machines.org`

ε–SV REGRESSION (Vapnik '98)

$$\frac{1}{2}||\theta||^2 + C\sum_i L_\varepsilon(x_i, y_i, f) = \min!,$$

$$L_\varepsilon(x, y, f) = \max\{0, |y - f(x)| - \varepsilon\}$$





KERNEL LOGISTIC REGRESSION (KLR)

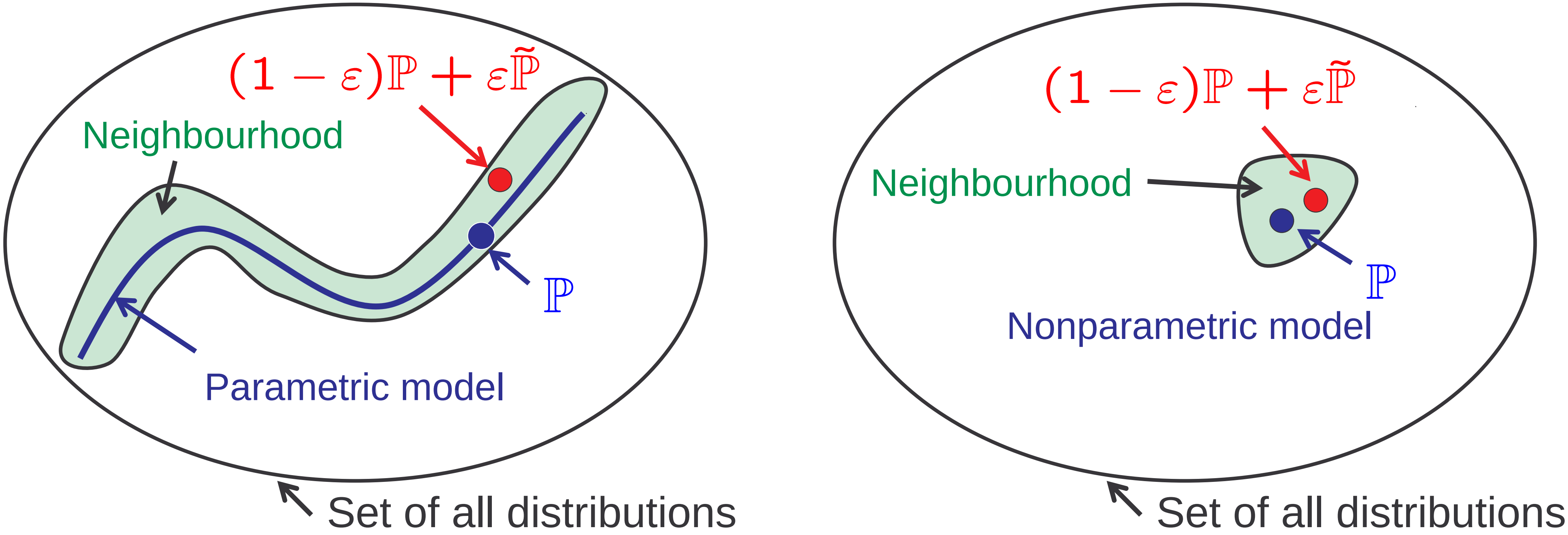
- SVM estimates $\text{sign}\left(P(Y=1|X=x)-\frac{1}{2}\right)$
- KLR estimates $f(x)=\log\left(\frac{P(Y=1|X=x)}{P(Y=-1|X=x)}\right)$, i.e. $P(Y=1|X=x)=\frac{1}{1+e^{-f(x)}}$
- KLR vs. SVM:

  $\longrightarrow$ offers estimate of class probabilities: risk scoring

  $\longrightarrow$ computationally more expensive.
  Keerthi et al. '02: fast dual algorithm with pseudo-code
  myKLR: Rüping '03 (Computer Science, Univ. of Dortmund)
  $n=10^5$ manageable on PC

  $\longrightarrow$ in general: no. SV's $\approx$ no. obs. for KLR fit $\hat{f}(x)=\hat{b}+\sum_i \hat{\alpha}_i k(x,x_i)$,
  i.e. no data compression. For SVM: in general no. SV's $\ll$ no. obs.

4. ROBUSTNESS ASPECTS



Goal: Statistic $T(\mathbb{P})$

But: $T((1-\varepsilon)\mathbb{P}+\varepsilon\tilde{\mathbb{P}}) \approx T(\mathbb{P})$?

Hampel's influence function:

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

Here: $T(\mathbb{P})$ is regularized theoretical risk:

$$(f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda}) = \arg \min_{f\in H,\, b\in\mathbb{R}} \mathrm{E}_{\mathbb{P}}\, L(Y, f(X)+b) + \lambda \|f\|_H^2$$

Tukey's sensitivity curve:

$$SC_n(z) = n\,[T_n(z_1,\ldots,z_{n-1},z) - T_{n-1}(z_1,\ldots,z_{n-1})]$$

PROP.: Uniform bounds on the difference quotient used by IF.
Assume: $L : \{-1,+1\} \times \mathbb{R} \to [0,\infty)$ continuous and convex loss function, $X \subset \mathbb{R}^p$ compact, H is RKHS of continuous kernel.
Then for all $\lambda > 0$ there exists a constant $c_L(\lambda) > 0$ (explicitly known) such that for ALL distributions $\mathbb{P}$ and $\tilde{\mathbb{P}}$ on $X \times \{-1,+1\}$ we have

$$\left\| \frac{f_{(1-\varepsilon)\mathbb{P}+\varepsilon\tilde{\mathbb{P}},\lambda} - f_{\mathbb{P},\lambda}}{\varepsilon} \right\|_H \leq c_L(\lambda)\, \|\mathbb{P} - \tilde{\mathbb{P}}\|_{\mathcal{M}}\,, \quad \varepsilon > 0\,.$$

Proof: Chr & Steinwart '03

Applications:

- SVM, KLR, ...
- Tukey's sensitivity curve: $\mathbb{P} = \mathbb{P}_n$, $\tilde{\mathbb{P}} = \Delta_z\,,\ \varepsilon = \frac{1}{n}$
- upper bound of max-bias curve

Some results on robustness properties of the influence function.

5. SUMMARY

- convex risk minimization methods based on kernels have many desirable properties
- robustness properties for SVM, KLR, AdaBoost have been studied

Current research together with Dipl.-Stat. M. Marin-Galiano (DoMuS):

- combination KLR + $\varepsilon-$SVR (or $\nu-$SVR)
- simulations to study robustness properties

REFERENCES

Christmann, Steinwart (2003). *On robust properties of convex risk minimization methods for pattern recognition.* Submitted.

Christmann, Fischer, Joachims (2002). Comparison between various regression depth methods and the support vector machine to approximate the minimum number of misclassifications. *Computational Statistics*, 17, 273-287.

Rousseeuw, Christmann (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics & Data Analysis*, 43, 315-332.

Friedman, Hastie, Tibshirani (2000). Additive logistic regression: a statistical view of boosting (with discussion). *Ann. Statist.*, 28, 337-407.

Hampel, Ronchetti, Rousseeuw, Stahel (1986). *Robust statistics: The Approach Based on Influence Functions*. Wiley.

Hastie, Tibshirani, Friedman (2001). *The Elements of Statistical Learning. Data Mining, Inference and Prediction.* Springer.

Höffgen, Simon, van Horn (1995). Robust Trainability of Single Neurons. *J. Computer and System Sciences*, 50, 114-125.

Schölkopf et al. (2000). New support vector algorithms. *Neural Computation*, 12, 1207-1245.

Schölkopf, Smola (2002) *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Smyth, Jørgensen (2002). Fitting Tweedie's Compound Poisson Model to Insurance Claims Data: Dispersion Modelling. *ASTIN Bulletin*, 32, 143-157.

Vapnik (1998). *Statistical Learning Theory*. Wiley.

Zhang (2001). Statistical behaviour and consistency of classification methods based on convex risk minimization. To appear in *Ann. Statist.*

christmann@statistik.uni-dortmund.de
www.statistik.uni-dortmund.de