Classification SVM algorithms with interval-valued training data using triangular and Epanechnikov kernels

Lev V. Utkin, Anatoly I. Chekh, Yulia A. Zhuk

Pescara, 2015

Authors ...

Lev **Yulia**

Saint Petersburg State
Forest Technical University

Anatoly

Saint Petersburg State
Electrotechnical University

Authors from ...

Saint Petersburg State
Forest Technical University

Saint Petersburg State
Electrotechnical University

A binary classification problem by precise data

Given:

- a training set $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$, (examples, patterns, etc.)
- $\mathbf{x} \in \mathcal{X}$ is a multivariate input of m features, $\mathcal{X}$ is a compact subset of $\mathbb{R}^m$
- $y \in \{-1, 1\}$ is a scalar output (labels of classes)

The learning problem:

- to select a function $f(\mathbf{x}, w_{\text{opt}})$ from a set of functions $f(\mathbf{x}, w)$ parameterized by a set of parameters $w \in \Lambda$, which separates examples of different classes y.

The expected risk for solving the standard classification problem

Minimize the risk functional or expected risk:

$$R(\mathbf{w},b) = \int_{\mathbb{R}^m} l(\mathbf{w},\phi(\mathbf{x}))\mathrm{d}F(\mathbf{x}),$$

the loss function:

$$l(\mathbf{w},\phi(\mathbf{x})) = \max\left\{0, b - \langle \mathbf{w},\phi(\mathbf{x})\rangle\right\}.$$

The empirical expected risk with the smoothing (Tichonov's) term

$$R_{emp}(\mathbf{w},b) = \frac{1}{n}\sum_{i=1}^{n} l(\mathbf{w},\phi(\mathbf{x}_i)) + C\cdot\|\mathbf{w}\|^2.$$

Support vector machine (SVM): a dual form form

The Lagrangian:

$$\max_{\alpha} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \right),$$

subject to

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \ 0 \leq \alpha_i \leq C, \ i = 1, ..., n.$$

The separating function f in terms of Lagrange multipliers:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b.$$

A binary classification problem by interval-valued data

Training set: $(\mathbf{A}_i, y_i)$, $i = 1, ..., n$. $\mathbf{A}_i \subset \mathbb{R}^m$ is the Cartesian product of m intervals $[\underline{a}_i^{(k)}, \overline{a}_i^{(k)}]$, $k = 1, ..., m$.

Reasons of interval-valued data:

- Imperfection of measurement tools
- Imprecision of expert information
- Missing data

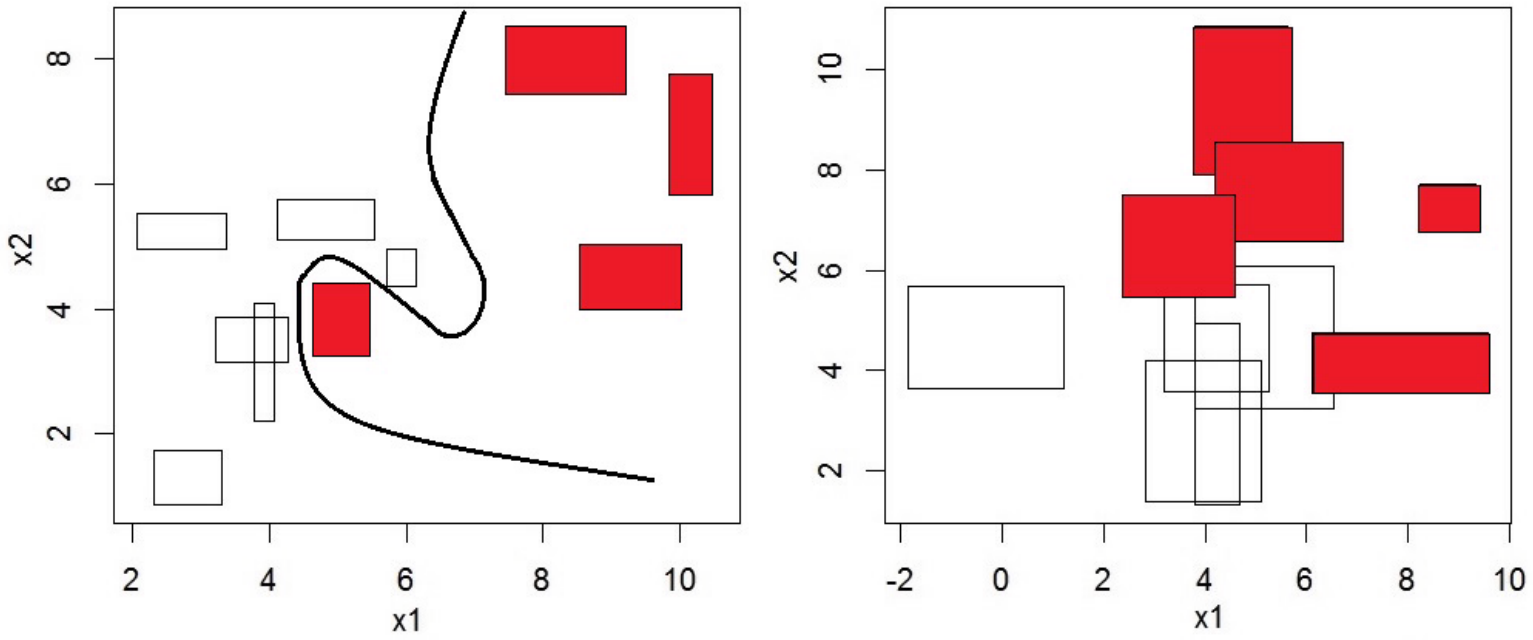
Approaches to interval-valued data in classification and regression (1)

- Interval-valued data are replaced by precise values based on some assumptions, for example, by taking middle points of intervals (LimaNeto and Carvalho 2008): *a very popular approach, unjustified, especially, by large intervals*
- The standard interval analysis (Angulo 2008, Hao 2009): *only linear separating or regression functions*
- Bernstein bounding schemes (Bhadra et al. 2009): *incorporate probability distributions over intervals.*

Approaches to interval-valued data in classification and regression (2)

- The Euclidean distance between two data points in the Gaussian kernel is replaced by the Hausdorff distance and other distances between two hyper-rectangles (Do and Poulet 2005, Chavent 2006, Souza and Carvalho 2004, Pedrycz et al 2008, Schollmeyer and Augustin 2013): *a nice and simple idea, but with some questions.*
- Minimizing and maximizing the risk measure over values of intervals (Utkin and Coolen 2011, Cattaneo and Wienzierz 2015): *only monotone separating functions* (Utkin and Coolen 2011) *or only interval-valued response variables y in regression models* (Cattaneo and Wienzierz 2015).

Classification problems by interval-valued data



Ideas underlying two new algorithms

1. Interval observations produce a set of expected risk measures such that the lower and upper risk measures are determined by minimizing and by maximizing the risk measure over values of intervals (*this is an old idea used in Utkin and Coolen 2011, Cattaneo and Wienzierz 2015*).
2. By applying the lower risk (**the minimax strategy**), it would be nice to isolate a "linear" programm from the SVM with variables $\mathbf{x}_i \in \mathbf{A}_i$ and then to work with extreme points $\mathbf{x}_i^*$.
3. **Important idea**: We replace the Gaussian kernel by the triangular kernel which can be regarded as an approximation of the Gaussian kernel (Utkin and Chekh 2015). This replacement allows us to get a set of linear programms with variables $\mathbf{x}_i$ restricted by $\mathbf{A}_i$, $i = 1, ..., n$.

Interval-valued training data, belief functions and minimax strategy

Lower $\underline{R}$ and upper $\overline{R}$ expectations of the loss function $l(\mathbf{x})$ in the framework of belief functions (Nguyen-Walker 1994, Strat 1990):

$$\underline{R} = \sum_{i=1}^{n} m(\mathbf{A}_i) \inf_{\mathbf{x}_i \in \mathbf{A}_i} l(\mathbf{x}_i) = \frac{1}{n} \sum_{i=1}^{n} \inf_{\mathbf{x}_i \in \mathbf{A}_i} l(\mathbf{x}_i),$$

$$\overline{R} = \sum_{i=1}^{n} m(\mathbf{A}_i) \sup_{\mathbf{x}_i \in \mathbf{A}_i} l(\mathbf{x}_i) = \frac{1}{n} \sum_{i=1}^{n} \sup_{\mathbf{x}_i \in \mathbf{A}_i} l(\mathbf{x}_i).$$

The minimax strategy (Γ-minimax): we do not know a precise value of the loss function l, but we take the "worst" value providing the largest value of the expected risk (Berger 1994, Gilboa and Schmeidler 1989, Robert 1994): $R(\mathbf{w}_{\text{opt}}, b_{\text{opt}}) = \min_{\mathbf{w},\rho} \overline{R}(\mathbf{w}, b)$.

Support vector machine (SVM): a dual form form

The Lagrangian:

$$\max_{\mathbf{x}_i \in \mathbf{A}_i} \max_{\alpha} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \right),$$

subject to

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \ 0 \leq \alpha_i \leq C, \ i = 1, ..., n.$$

The separating function f in terms of Lagrange multipliers:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b.$$

The first algorithm

An obvious way is to fix α and to replace the Gaussian kernel

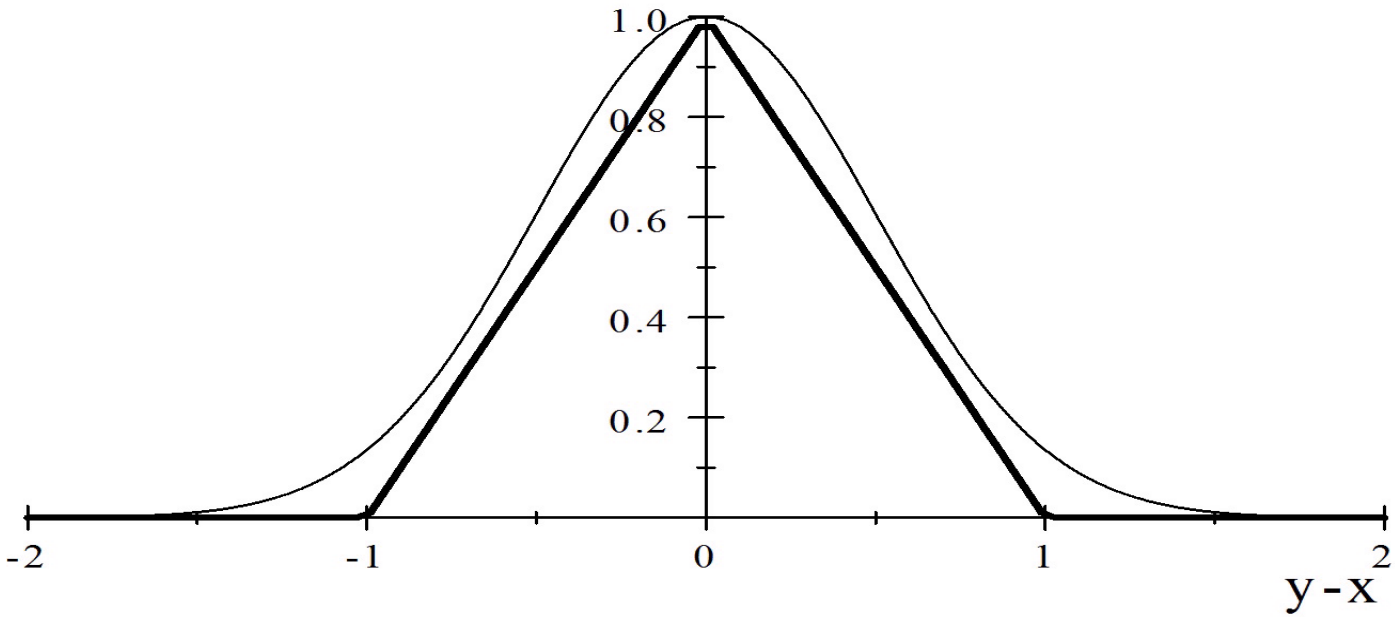
$$K(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}\right)$$

$$\Downarrow$$

$$T(\mathbf{x},\mathbf{y}) = \max\left\{0, 1-\frac{\|\mathbf{x}-\mathbf{y}\|^1}{\sigma^2}\right\}$$

Triangular kernel

We approximate the Gaussian kernel by the **triangular kernel** *in order to get a "piecewise" linear programm!*



A set of standard quadratic problems

- By fixed Lagrangian multipliers α and the triangular kernel, we get a linear problem with constraints $\mathbf{x}_i \in \mathbf{A}_i$.
- Its optimal solution is achieved at extreme points or vertices of the hyperrectangles produced by $\mathbf{A}_i$, i.e., at interval bounds.
- For every extreme point, we solve the standard quadratic problem.

The main problem of the algorithm:
If we have n interval-valued data consisting of m features, then the number of extreme points (quadratic programms) is $t = 2^{nm}$.

What to do when we have many intervals?

Idea: There are many variants of SVMs.

It would be nice to find a SVM for which constraints for classification parameters do not depend on interval observations x_i.

L_infinite-norm SVM

An interesting L_∞-norm SVM proposed by Zhou et al. 2002:

$$\min R = \min \left(-r + C \sum_{i=1}^{n} \xi_i \right),$$

subject to

$$y_j \left(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_j) + b \right) \geq r - \xi_j, \ j = 1, ..., n,$$

$$-1 \leq \alpha_i \leq 1, \ i = 1, ..., n, \quad r \geq 0, \ \xi_j \geq 0, \ j = 1, ..., n.$$

α_j, ξ_j, $j = 1, ..., n$, r, b are optimization variables

The dual form is more interesting

The dual form by fixed $\mathbf{x}_1, ..., \mathbf{x}_n$:

$$\min_z \sum_{i=1}^{n} y_i \left(\sum_{j=1}^{n} z_j y_j K(\mathbf{x}_i, \mathbf{x}_j) \right),$$

subject to $\sum_{i=1}^{n} z_i \geq 1,\ 0 \leq z_j \leq C,\ j = 1, ..., n,\ \ \sum_{i=1}^{n} z_i y_i = 0.$

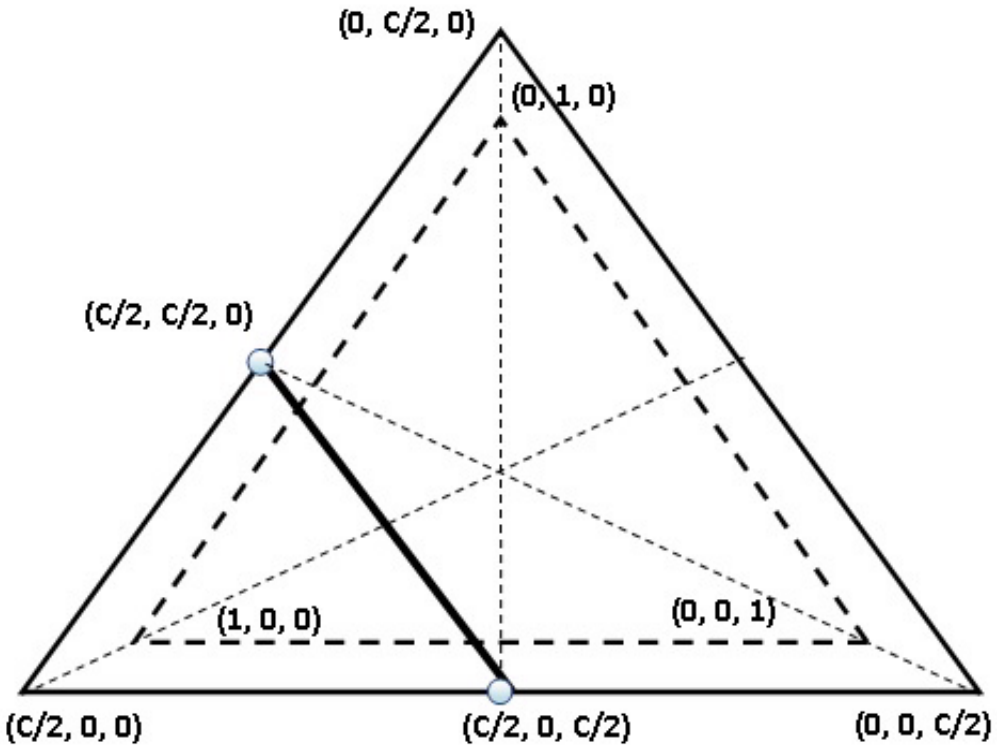
All $\mathbf{x}_1, ..., \mathbf{x}_n$ are in the objective function

Constraints have only variables $z_1, ..., z_n$ which produce the convex set Z of an interesting form.

The convex sets of solutions

$$\sum_{i=1}^{n} z_i \geq 1,\ 0 \leq z_j \leq C,\ j = 1, ..., n,\ \ \sum_{i=1}^{n} z_i y_i = 0.$$

$$z_1 \rightarrow y_1 = -1,\ \ z_2 \rightarrow y_2 = 1,\ \ z_3 \rightarrow y_3 = 1$$



The convex sets of solutions

Proposition

Let n_- *and* n_+ *be numbers of* $y = -1$ *and* $y = 1$. t *and* s:

$$(2C)^{-1} < t \leq \min(n_-, n_+),$$
$$(2C)^{-1} - 1 \leq s < \min\left((2C)^{-1}, n_-, n_+\right),$$

The first subset: $N_1 = \sum_{t=\lceil 1/2C \rceil}^{\min(n_-,n_+)} \binom{n_-}{t}\binom{n_+}{t}$ *extreme points:* t *elements from every class are* C, *others are* 0.
If $s \geq 0$, *then the second subset:* $N_2 = (n_- - s)(n_+ - s)\binom{n_-}{s}\binom{n_+}{s}$ *extreme points:* s *elements from every class are* C, *one element from every class is* $1/2 - sC$, *others are* 0.

The final optimization problems

By using again the triangular kernel, we get a set of N_1+N_2 (the number of extreme points of Z) linear programms with variables $\mathbf{x}_i \in \mathbf{A}_i, i = 1, ..., n$.

The number of linear programms does not depend on the number m of features!

The Epanechnikov kernel

Another kernel:

$$T_2(\mathbf{x}, \mathbf{y}) = \max\{0, 1 - \|\mathbf{x} - \mathbf{y}\|^2 / \sigma^2\}.$$

We get a quadratically constrained linear program (QCLP). Tools: the sequential quadratic programming (Boggs and Tolle 1995), SNOP (Gill et al. 2002)

Advantages of the algorithms

1. The algorithms allows us to construct non-linear separating functions.
2. The algorithms are justified from the decision point of view (minimax strategy).
3. The algorithms produce unique and consistent precise points of intervals corresponding to the largest value of the expected classification risk. The points compose a single probability distribution among a set of distributions produced by intervals in the framework of Dempster-Shafer theory.
4. The algorithms can be extended on the support vector regression algorithms when dependent as well as independent variables are interval-valued.

Questions

?