SVM-based classification algorithms with interval-valued data

Lev V. Utkin

2015

Author from ...

Saint Petersburg State Forest Technical University

One-class classification (OCC) by precise data

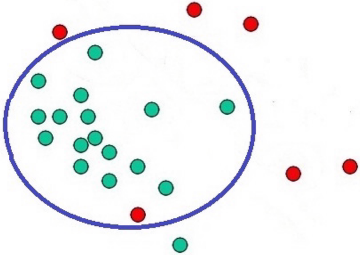
Given:

- an unlabeled training data $\mathbf{x}_1, ..., \mathbf{x}_n \subset \mathcal{X}$
- $\mathbf{x}$ is a multivariate input of m features (examples, patterns, etc.), $\mathcal{X}$ is a compact subset of $\mathbb{R}^m$

The learning problem is:

- to construct a function $f(\mathbf{x})$ which takes the value $+1$ in a "small" region capturing most of the data points and -1 elsewhere

One-class classification: novelty detection



Three main models of the OCC

1. **Schölkopf et al. 2000, 2001**
2. Tax and Duin 1999, 2004
3. Campbell and Bennett 2001

The main idea for solving the OCC problem by precise data

1. Data points lie on the surface of a hypersphere in feature space induced by the map $\phi(\mathbf{x})$.
2. A hyperplane $f(\mathbf{x}, \mathbf{w}) = \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle - \rho = 0$ separates the data from the origin with maximal margin, i.e., we want ρ to be as large as possible so that the volume of the halfspace $\langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle \geq \rho$ is minimized.

The main formal idea for solving the OCC problem

Minimize the risk functional or expected risk

$$R(\mathbf{w},\rho) = \int_{\mathbb{R}^m} l(\mathbf{w}, \phi(\mathbf{x}))\mathrm{d}F_0(\mathbf{x}),$$

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

$v \in [0;1]$ *controls the extent of margin errors (smaller* v *means fewer outliers are ignored)*

The empirical expected risk

$$R_{\mathrm{emp}}(\mathbf{w},\rho) = \frac{1}{n}\sum_{i=1}^{n} l(\mathbf{w}, \phi(\mathbf{x}_i)).$$

SVM for the OCC problem (primal form)

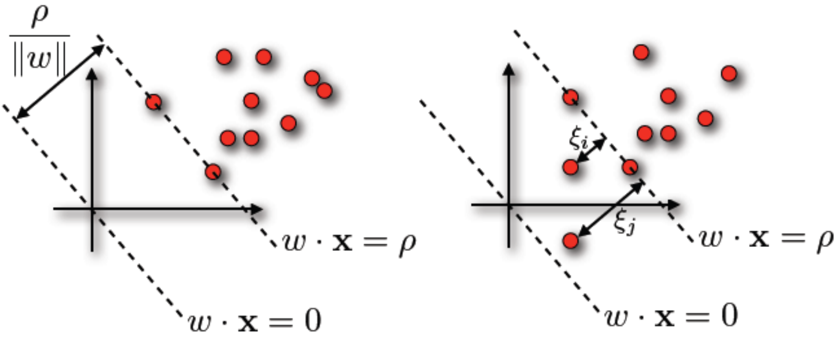
The quadratic program:

$$\min_{w,\xi,\rho} \frac{1}{2}\left\|\mathbf{w}\right\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho,$$

subject to

$$\langle \mathbf{w},\phi(\mathbf{x}_i)\rangle \geq \rho - \xi_i, \quad \xi_i \geq 0, \ \ i = 1, ..., n.$$

SVM for the OCC problem (Schölkopf et al. 2000, 2001)



SVM for the OCC problem (dual form, Lagrangian)

The quadratic program:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j),$$

subject to

$$0 \leq \alpha_i \leq \frac{1}{\nu n}, \ \sum_{i=1}^{n} \alpha_i = 1.$$

The decision function f:

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \mathbf{x}) - \rho \right).$$

$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)\phi(\mathbf{x}_j)$ is the Gaussian (RBF) kernel.

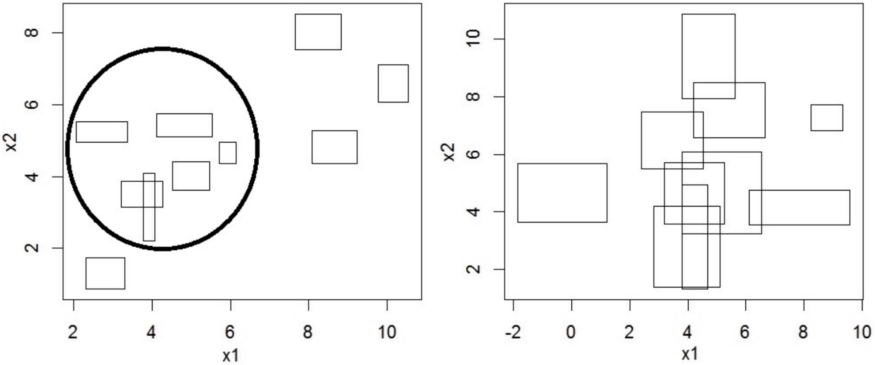
A OCC problem statement by interval data

Training set $(\mathbf{A}_i)$, $i = 1, ..., n$. Every $\mathbf{A}_i \subset \mathbb{R}^m$ 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

Examples of interval data



Approaches to interval-valued data in classification

- Interval-valued data are replaced by precise values based on some assumptions, for example, by taking middle points of intervals (LimaNeto and Carvalho 2008)
- The standard interval analysis (Angulo 2008, Hao 2009):
- Change of the Euclidean distance between two data points in the Gaussian kernel by the Hausdorff distance between two hyper-rectangles (Do and Poulet 2005).
- Similar models with the Hausdorff distance and other distances (Chavent 2006, Souza and Carvalho 2004, Pedrycz et al 2008, Schollmeyer and Augustin 2013)
- Bernstein bounding schemes (Bhadra et al. 2009)

Ideas underlying a new model

1. Interval-valued observations produce a set of expected classification risk measures such that the lower and upper risk measures can be determined by minimizing and by maximizing the risk measure over values of intervals.
2. For the lower risk measure (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. It is proposed to replace the Gaussian kernel by the well-known triangular kernel which can be regarded as an approximation of the Gaussian kernel. This replacement allows us to get a set of linear optimization problems with variables $\mathbf{x}_i$ restricted by intervals $\mathbf{A}_i$, $i = 1, ..., n$.

Interval-valued training data and belief functions (Dempster-Shafer theory)

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):

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

Basic probability assignments

$$m : \mathcal{P}o(\mathcal{X}) \to [0,1], \quad m(\varnothing) = 1, \sum_{\mathbf{A} \in \mathcal{P}o(\mathcal{X})} m(\mathbf{A}) = 1.$$

$$m(\mathbf{A}_i) = c_i / n.$$

Minimax strategy

$$R(\mathbf{w}_{\text{opt}},\rho_{\text{opt}}) = \min_{\mathbf{w},\rho} \overline{R}(\mathbf{w},\rho) = \min_{\mathbf{w},\rho} \left(\sum_{i=1}^{n} m(\mathbf{A}_i) \sup_{\mathbf{x}_i \in \mathbf{A}_i} l(\mathbf{x}_i) \right).$$

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).

Primal optimization problem by interval data (L_2-norm SVM)

$$R=\sup_{\mathbf{x}_i\in\mathbf{A}_i}\min_{\mathbf{w},\rho}\left(\frac{1}{2}\left\|\mathbf{w}\right\|^2+\frac{1}{n\nu}\sum_{i=1}^{n}\max\left\{0,\rho-\left\langle\mathbf{w},\phi(\mathbf{x}_i)\right\rangle\right\}-\rho\right),$$

subject to

$$\xi_i\geq\rho-\left\langle\mathbf{w},\phi(\mathbf{x}_i)\right\rangle,\quad\xi_i\geq 0,\ \ i=1,...,n,$$

$$\mathbf{x}_i\in\mathbf{A}_i,\quad i=1,...,n.$$

Dual optimization problem (Lagrangian) by interval data

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

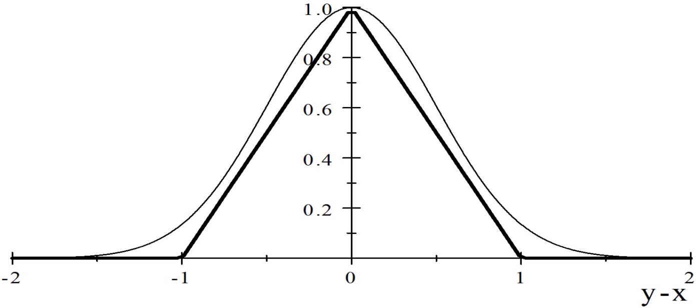
subject to

$$0 \leq \alpha_i \leq \frac{1}{\nu n}, \ \sum_{i=1}^{n} \alpha_i = 1, \ \mathbf{x}_i \in \mathbf{A}_i, \ i = 1, ..., n.$$

How to reduce the problem to the linear (or "approximately" linear) one?

The main idea (1)

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



The main idea (2)

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

$$\Downarrow$$

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

$T(\mathbf{x},\mathbf{y})$ is almost linear or piecewise linear

Dual optimization problem (Lagrangian) by interval data

- If we fix Lagrange multipliers α_i, then we get the following simple linear programming problem:

$$\sup_{\mathbf{x}_i, i=1,\dots,n} \left(-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j T(\mathbf{x}_i, \mathbf{x}_j) \right).$$

subject to $\mathbf{x}_i \in \mathbf{A}_i, \quad i = 1, ..., n.$

- Its optimal solution is achieved at extreme points or vertices of the polytope produced by $\mathbf{A}_i$, i.e., at interval bounds.

A set of Lagrangians

- If the number of extreme points is t, then we solve t quadratic optimization problems by substituting the extreme points (values $\mathbf{x}_1^*,...,\mathbf{x}_n^*$) into every problem:

$$\max_{\alpha}\left(-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\max\{0,1-\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^1/\sigma^2\}\right).$$

subject to $0\leq\alpha_i\leq 1/(vn)$, $\sum_{i=1}^{n}\alpha_i=1$.

- The largest value of the objective function corresponds to the optimal values $\mathbf{x}_1^*,...,\mathbf{x}_n^*$ and to the optimal parameters α_{opt}.

The main virtues and shortcomings

1. If we have n interval-valued data consisting of m features, then the number of extreme points is $t = 2^{nm}$.
2. However, the approach can be applied to arbitrary convex set $\mathcal{M}$ of data values (to imprecise data), for example,
   - comparative data (the first feature is larger than the second feature)
   - functions of data (the sum of two features is less than 1)
   - in fact, this approach is better for the above cases than for intervals.

Again interval-valued data

- What to do when we have many intervals?

Again interval-valued data

- What to do when we have many intervals?
- **Idea:** There are many variants of OCC SVMs.

Again interval-valued data

- What to do when we have many intervals?
- **Idea:** There are many variants of OCC SVMs.

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

Again interval-valued data

- What to do when we have many intervals?
- **Idea:** There are many variants of OCC SVMs.

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

- This is the linear programming OCC SVM by Campbell and Bennett, 2001 for which constraints in **its dual form** do not depend on vectors of observations. This allows us to represent the dual optimization problem as a set of simple optimization problems.

Campbell and Bennett model by interval data

$$W(\varphi, b) = \sup_{\mathbf{x}_i \in \mathbf{A}_i} \min_{\varphi_i, b, \xi_i} \left(\sum_{i=1}^{n} \left(\sum_{j=1}^{n} \varphi_j K(\mathbf{x}_i, \mathbf{x}_j) + b \right) + \frac{1}{v} \sum_{i=1}^{n} \xi_i \right)$$

subject to $\mathbf{x}_i \in \mathbf{A}_i$, $i = 1, ..., n$,

$$\sum_{i=1}^{n} \varphi_i = 1, \;\; \varphi_i \geq 0.$$

$$\sum_{j=1}^{n} \varphi_j K(\mathbf{x}_i, \mathbf{x}_j) + b \geq -\xi_i, \;\; \xi_i \geq 0, \;\; i = 1, ..., n.$$

It turns out that the dual optimization problem and the triangular kernel provide a more or less simple way for solving the OCC problem.

The dual form

A set of n optimization problems

$$\sup_{\mathbf{x}_i \in \mathbf{A}_i} \left(\max_{\alpha} \sum_{i=1}^{n} (1 - n\alpha_i) K(\mathbf{x}_i, \mathbf{x}_j) \right) \rightarrow \min_{j=1,\dots,n},$$

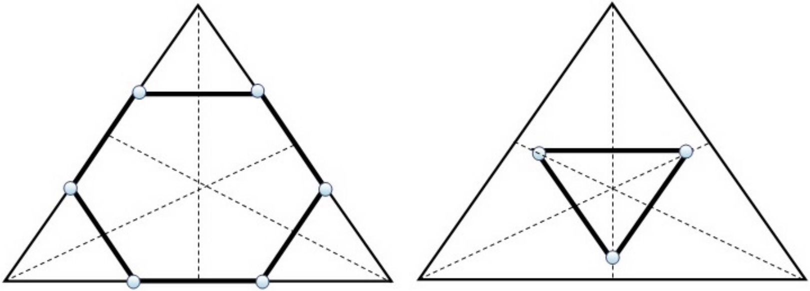
subject to

$$0 \leq \alpha_i \leq \frac{1}{vn}, \ \ i = 1, \dots, n, \ \ \sum_{i=1}^{n} \alpha_i = 1.$$

Let us fix $\mathbf{x}_1, \dots, \mathbf{x}_n$.

The convex sets of solutions

$$0 \leq \alpha_i \leq \frac{1}{\nu n}, \;\; i = 1, ..., n, \;\; \sum_{i=1}^{n} \alpha_i = 1.$$



The convex set of solutions (extreme points)

Proposition

1. *If* $v \geq (n-1)n^{-1}$, *then* $T = n$ *extreme points: the k-th element is* $v^{-1}(n^{-1} + v - 1)$ *and* $n - 1$ *elements are* $v^{-1}n^{-1}$.
2. *If* $n^{-1} < v < (n-1)n^{-1}$, *then* $T = s\binom{n}{s}$ *extreme points:* $s \in \mathbb{N}$ *is defined by*

$$\frac{1}{n-s+1} \leq \frac{1}{vn} \leq \frac{1}{n-s}.$$

Extreme points: $n - s$ *elements are* $v^{-1}n^{-1}$, *one element is* $1 - (n-s)v^{-1}n^{-1}$, *and* $s - 1$ *elements are* 0.

3. *If* $v \leq n^{-1}$, *then the unit simplex.*

Change of the kernel

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

$$\Downarrow$$

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

A set of "linear" problems

If $\alpha^{(k)} = (\alpha_n^{(k)}, ..., \alpha_n^{(k)})$ is the k-th extreme point, then we get the linear optimization problems:

$$\max_{\mathbf{x}_i \in \mathbf{A}_i} \left(\max_{k=1,...,T} \min_{j=1,...,n} O(j,k) \right)$$

$$= \max_{\mathbf{x}_i \in \mathbf{A}_i} \left(\max_{k=1,...,T} \min_{j=1,...,n} \sum_{i=1}^{n} (1 - n\alpha_i^{(k)}) \max \left\{ 0, 1 - \frac{\left\| \mathbf{x}_i - \mathbf{x}_j \right\|^1}{\sigma^2} \right\} \right),$$

subject to $\mathbf{x}_i \in \mathbf{A}_i$, $i = 1, ..., n$.

Auxiliary lemma

Lemma (Beaumont,1998)

If $[\underline{x}, \overline{x}] \subset \mathbb{R}$, $\underline{x} < \overline{x}$, *and, if*

$$u = \frac{|\overline{x}| - |\underline{x}|}{\overline{x} - \underline{x}}, \quad v = \frac{\overline{x}|\underline{x}| - \underline{x}|\overline{x}|}{\overline{x} - \underline{x}},$$

we have

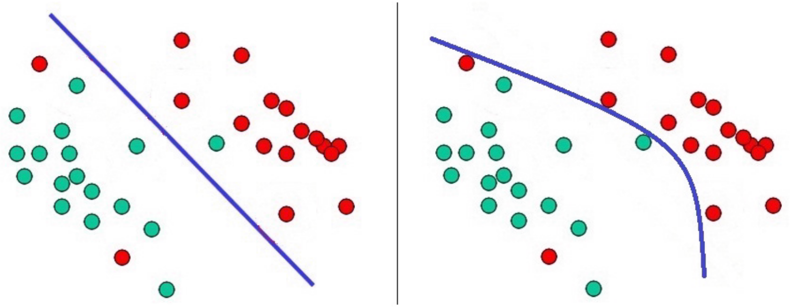
$$\forall x \in [\underline{x}, \overline{x}], \quad |x| \leq ux + v.$$

An algorithm

Step 1. $\mathcal{E}(\mathcal{M}_v)$ is the set of extreme points $\alpha^{(1)}, ..., \alpha^{(T)}$.
Step 2. Select the k-th extreme point $\alpha^{(k)}$ from $\mathcal{E}(\mathcal{M}_v)$.
Step 3. For $j \in \{1, ..., n\}$ and $k \in \{1, ..., T\}$, solve linear problems over $\mathbf{x}_i \in \mathbf{A}_i$.
Step 3. For $j \in \{1, ..., n\}$, select $k_j^* \leftarrow \arg_k \max O(j, k)$.
Step 5. Select $j^* \leftarrow \arg_j \min O(j, k_j^*)$. As a result, we get an optimal vector $(\mathbf{x}_1^*, ..., \mathbf{x}_n^*)$.
Step 6. Now we solve the original Campbell and Bennett model with $(\mathbf{x}_1^*, ..., \mathbf{x}_n^*)$.

A binary classification problem by precise data

- **Given:** a training set $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$
- $\mathbf{x} \in \mathcal{X}$ is a multivariate input of m features (examples, patterns, etc.), $\mathcal{X}$ 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 $f(\mathbf{x}, w) = \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle + b$ parameterized by a set of parameters w, b, which separates examples of different classes y.



L_2-norm SVM

Primal:

$$\min_{\xi,\mathbf{w},b} R = \min_{\xi,\mathbf{w},b} \left(\frac{1}{2} \|\mathbf{w}\|_2 + C \sum_{i=1}^{n} \xi_i \right),$$

s.t. $\xi_i \geq 0, \quad y_i \left(\langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle + b \right) \geq 1 - \xi_i, \; i = 1, ..., n.$

Dual (Lagrangian):

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

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

A binary classification problem by interval-valued data

- **Given:** a training set $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$
- $\mathbf{x}_i \in \mathbf{A}_i$, $i = 1, ..., n$.
- $y \in \{-1, 1\}$ is a scalar output (labels of classes)
- **The learning problem is:** to construct a function $f(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle + b$

L_2-norm SVM by interval data

This case totally coincides with the OOC SVM and is limited by the number of extreme points of $\mathbf{A}_i$: $t = 2^{nm}$

L_infinite-norm SVM by interval data

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

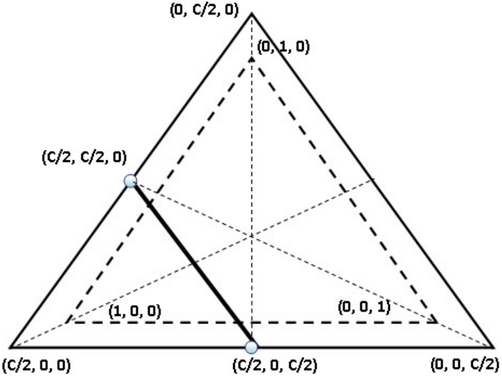
$$\sum_{i=1}^{n} z_i \geq 1,\ 0 \leq z_j \leq C,\ j = 1, ..., n,\ \ \sum_{i=1}^{n} z_i y_i = 0.$$

All $\mathbf{x}_1, ..., \mathbf{x}_n$ in the objective function, constraints are have only variables $z_1, ..., z_n$

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:

$$N_1 = \sum\nolimits_{t=\lceil 1/2C \rceil}^{\min(n_-,n_+)} \binom{n_-}{t}\binom{n_+}{t}$$

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

$$\min_{\text{extreme points } z^{*}} \min_{\mathbf{x}_i \in \mathbf{A}_i, i=1,\dots,n} \sum_{i=1}^{n} y_i \left(\sum_{j=1}^{n} z_j^{*} y_j K(\mathbf{x}_i, \mathbf{x}_j) \right),$$

where we use the triangle kernel

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

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)

Questions

?