Anomaly Detection
Hyperplane-based methods

Paul Irofti
Cristian Rusu
Andrei Pătraşcu

Computer Science Department
University of Bucharest

Outline

- One class classification: OC-SVM, SVDD. Algorithms.
- Robust versions. Algorithms.

Consider the training data is provided:

$$x_1, x_2, \cdots, x_m \subset X.$$

where m is the number of the observations and X some space (i.e. compact subset of $\mathbb{R}^n$).

Question

What is a "good" binary function f that captures the "region" of the most of datapoints where returns $+1$, and -1 elsewhere ?

One class classification

Consider the training data is provided:

$$x_1, x_2, \cdots, x_m \subset X.$$

where m is the number of the observations and X some space (i.e. compact subset of $\mathbb{R}^n$).

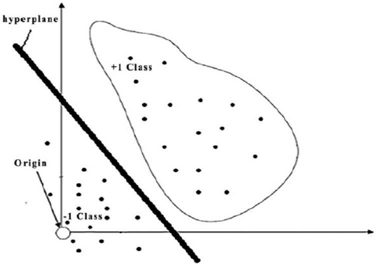
A simple answer: convex bodies such as

- hyperplanes
- spheres
- ellipsoids

Consider the training data is provided:

$$x_1, x_2, \cdots, x_m \subset X.$$

where m is the number of the observations and X some space (e.g. compact subset of $\mathbb{R}^n$).



4. Hyperplane in one-class support vector machine.

Idea: *The inliers are grouped in a region far from the origin.*

One class classification - hyperplane

Consider the training data is provided:

$$x_1, x_2, \cdots, x_m \subset X.$$

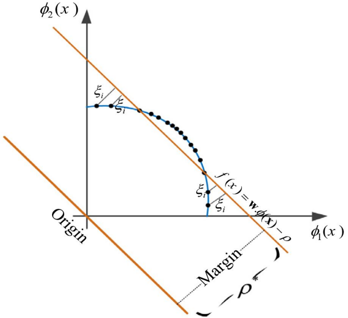
where m is the number of the observations and X some space (e.g. compact subset of $\mathbb{R}^n$).
Prior assumptions:

- the inliers are distributed in a region separable from the origin by a hyperplane
- the outliers lies near the origin

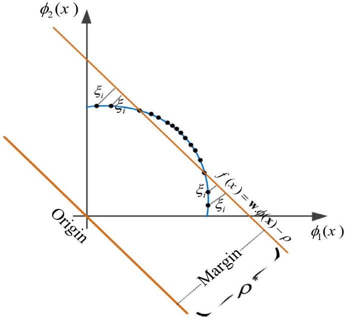
Take the decision function as $f(x) = \mathrm{sgn}(w^T x - \rho)$, then find the optimal paramaters (w, ρ) of the hyperplane such that:

$$w^T x_{inlier} \geq \rho$$
$$w^T x_{outlier} < \rho.$$

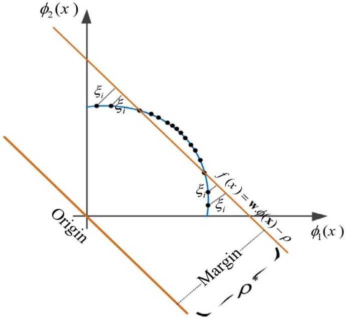
One class classification - hyperplane



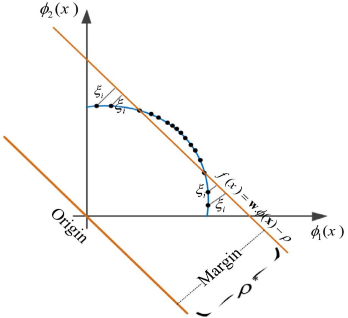
- In the separable case, there is an infinite number of hyperplanes of choice



- In the separable case, there is an infinite number of hyperplanes of choice
- Let the distance from the origin (of a hyperplane) be named as *margin*, then we desire to obtain the hyperplane with the maximal margin.



- In the separable case, there is an infinite number of hyperplanes of choice
- Let the distance from the origin (of a hyperplane) be named as *margin*, then we desire to obtain the hyperplane with the maximal margin.
- The distance from x to hyperplane $\{x : w^T x = \rho\}$ is $\frac{|w^T x - \rho|}{\|w\|}$.



- In the separable case, there is an infinite number of hyperplanes of choice
- Let the distance from the origin (of a hyperplane) be named as *margin*, then we desire to obtain the hyperplane with the maximal margin.
- The distance from x to hyperplane $\{x : w^T x = \rho\}$ is $\frac{|w^T x - \rho|}{||w||}$.
- Thus we maximize $\frac{\rho}{||w||}$.

- In order to maximize $\frac{\rho}{\|w\|}$, we minimize $\frac{1}{2}\|w\|^2 - \rho$
- In the nonseparable case we allow slack variables ξ to encode the outlyingness of nonseparable data:

$$
\begin{aligned}
w^T x_{inlier} &\geq \rho \\
\xi = \rho - w^T x_{outlier} &\geq 0.
\end{aligned}
$$

$$
\begin{aligned}
\min_{w, \xi \in \mathbb{R}^m, \rho \in \mathbb{R}} \; & \frac{1}{2}\|w\|^2 + \frac{1}{m\nu}\sum_{i=1}^{m} \xi_i - \rho \\
\text{s.t.} \; & \langle w, x_i \rangle \geq \rho - \xi_i, \xi_i \geq 0 \quad \forall i \in \{1, \cdots, m\}.
\end{aligned}
$$

$$\min_{w,\xi\in\mathbb{R}^m,\rho\in\mathbb{R}} \frac{1}{2}\|w\|^2 + \frac{1}{m\nu}\sum_{i=1}^{m}\xi_i - \rho$$
$$\text{s.t. } \langle w, x_i\rangle \geq \rho - \xi_i, \xi_i \geq 0 \quad \forall i \in \{1,\cdots,m\}.$$

- convex QP with m linear inequalities constraints
- regularization: $\|w\|^2$ (justify the minimum margin hyperplane)
- error slack variables: ξ
- penalty parameter: $1/\nu \in [1,\infty)$

Dual Problem

The primal problem is convex (linear constraints):

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + \frac{1}{m\nu}\sum_{i=1}^{m}\xi_i - \rho$$
$$\text{s.l. } w^T x_i \geq \rho - \xi_i \quad \forall i = 1, \cdots, m$$
$$\xi \geq 0.$$

The large number of constraints makes the primal hard to handle. Therefore, we take steps toward the dual: let the Lagrange multipliers $\lambda, \gamma \geq 0$

$$\mathcal{L}(w, \rho, \xi, \lambda, \gamma) = \frac{1}{2}\|w\|^2 + \frac{1}{m\nu}\sum_{i=1}^{m}\xi_i - \rho - \sum_{i=1}^{m}\lambda_i[w^T x_i - \rho + \xi_i] - \sum_{i=1}^{m}\gamma_i\xi_i.$$

Kuhn-Tucker optimality conditions:

$$\nabla_w \mathcal{L}(w, \rho, \xi, \lambda, \gamma) = w - \sum_{i=1}^{m} \lambda_i x_i = 0$$

$$\nabla_\rho \mathcal{L}(w, \rho, \xi, \lambda, \gamma) = \sum_{i=1}^{m} \lambda_i - 1 = 0$$

$$\nabla_{\xi_i} \mathcal{L}(w, \rho, \xi, \lambda, \gamma) = \frac{1}{m\nu} - \lambda_i - \gamma_i.$$

$$\gamma \geq 0, 0 \leq \lambda \leq \frac{1}{m\nu}.$$

We observe that at optimality we have:

$$w^* = \sum_{i=1}^{m} \lambda_i^* x_i.$$

The datapoints x^i such that $\lambda_i^* = 0$ do not contributes to the problem solution; those for which $\lambda_i^* > 0$ are **support vectors**: $\langle w^*, x^i \rangle - \rho^* = 0$.

Dual Problem

The dual problem:

$$\max_{\lambda} \ -\frac{1}{2}\lambda^T X X^T \lambda$$
$$\text{s.l. } e^T\lambda = 1,\ 0 \le \lambda_i \le \frac{1}{m\nu}.$$

where $X = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix}$.

- A convex quadratic objective with a high-dimensional Hessian ($m \times m$).
- Strict inliers $\lambda_i^* = 0$, SVs $\lambda_i^* > 0$.
- Once optimal λ^* is computed then $w^* = \sum_{i=1}^m \lambda_i^* x_i$.

Back to our data:

$$x_1, x_2, \cdots, x_m \subset X.$$

In the non-separable case, we map the training data in high-dimensional spaces by choosing a function $\phi : X \to \mathcal{F}$ such that the inner product between the images of ϕ can be evaluated some simple kernel:

$$k(x, y) := \langle \phi(x), \phi(y) \rangle.$$

Example: Gaussian kernel

$$k(x, y) := e^{-\frac{\|x-y\|^2}{\sigma}}$$

$$\phi(x_1), \phi(x_2), \cdots, \phi(x_m) \subset \mathcal{F}.$$

Primal nonlinear problem:

$$\min_{w, \xi \in \mathbb{R}^m, \rho \in \mathbb{R}} \frac{1}{2}\|w\|^2 + \frac{1}{m\nu}\sum_{i=1}^{m} \xi_i - \rho$$
$$\text{s.t. } \langle w, \phi(x_i) \rangle \geq \rho - \xi_i, \xi_i \geq 0 \quad \forall i \in \{1, \cdots, m\}.$$

Now the dimension of w is the dimension of feature space. The problem is still convex and has a similar dual as in the linear case.

The dual problem:

$$\max_{\lambda} \; -\frac{1}{2}\lambda^T K \lambda$$
$$\text{s.l. } e^T\lambda = 1,\; 0 \le \lambda_i \le \frac{1}{m\nu}.$$

where $K = \begin{bmatrix}\phi(x_1) & \phi(x_2) & \cdots & \phi(x_m)\end{bmatrix}^T \begin{bmatrix}\phi(x_1) & \phi(x_2) & \cdots & \phi(x_m)\end{bmatrix}$.

- Hessian $K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle$ and $K_{ii} = 1$.
- Strict inliers $\lambda_i^* = 0$, SVs $\lambda_i^* > 0$.
- Once optimal λ^* is computed then $w^* = \sum_{i=1}^m \lambda_i^* \phi(x_i)$.

How to solve the Dual problem ?

Possible algorithms:

- off-the-shelf QP solvers: cvxpy, quadprog, MOSEK etc. $O(m^3)$

Possible algorithms:

- off-the-shelf QP solvers: cvxpy, quadprog, MOSEK etc. $O(m^3)$
- Simplex-type feasible set ($m\log(m)$ to project on)

Possible algorithms:

- off-the-shelf QP solvers: cvxpy, quadprog, MOSEK etc. $O(m^3)$
- Simplex-type feasible set ($m \log(m)$ to project on)
- The cost of a first-order iteration: gradient evaluation $O(m^2)$ + projection onto the simple $O(m \log(m))$.

Possible algorithms:

- off-the-shelf QP solvers: cvxpy, quadprog, MOSEK etc. $O(m^3)$
- Simplex-type feasible set ($m \log(m)$ to project on)
- The cost of a first-order iteration: gradient evaluation $O(m^2)$ + projection onto the simple $O(m \log(m))$.
- Best suggestion: approach large-scale instances by coordinate descent algorithms (scitkit-learn uses libsvm for training)

$$\min_{\lambda} \frac{1}{2}\lambda^T H\lambda + b^T\lambda$$
$$\text{s.l. } e^T\lambda = 1,\ 0 \leq \lambda_i \leq C.$$

Idea of CD

Instead of approximating the whole λ at each iteration, update only a small block of variables.

$$\min_{\lambda} \frac{1}{2}\lambda^T H\lambda + b^T\lambda$$
$$\text{s.l. } e^T\lambda = 1,\ 0 \leq \lambda_i \leq C.$$

Idea of CD

Instead of approximating the whole λ at each iteration, update only a small block of variables.

Example CD

Exact 2-coordinate descent: choose $(i, j) \in \{1, \cdots, m\}$; let $\lambda_{ij} = [\lambda_i\ \lambda_j]^T$

$$\lambda_{ij}^+ := \arg\min_{\lambda_{ij}} \frac{1}{2}(\lambda_i^2 + \lambda_j^2) + K_{ij}\lambda_i\lambda_j + b_{ij}^T\lambda_{ij}$$
$$\text{s.l. } \lambda_i + \lambda_j = \Delta,\ 0 \leq \lambda_i, \lambda_j \leq C.$$

$$\min_{\lambda} \frac{1}{2}\lambda^T H\lambda + b^T\lambda$$
$$\text{s.l. } e^T\lambda = 1,\ 0 \le \lambda_i \le C.$$

Exact 2-coordinate descent main loop:

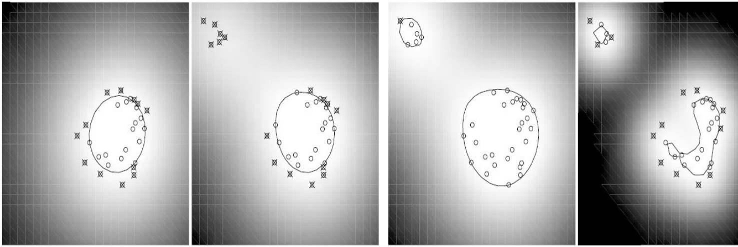
1. choose $(i,j) \in \{1, \cdots, m\}$ (cyclic, random etc.)
2. update:

$$\lambda_{ij}^* := \arg\min_{\lambda_{ij}} \frac{1}{2}(\lambda_i^2 + \lambda_j^2) + K_{ij}\lambda_i\lambda_j + b_{ij}^T\lambda_{ij}$$
$$\text{s.l. } \lambda_i + \lambda_j = \Delta,\ 0 \le \lambda_i, \lambda_j \le C.$$

   where $\Delta = 1 - \sum_{t \ne i,j} \lambda_t$.

3. $\lambda_{ij}^+ := \lambda_{ij}^*$ and $\lambda_{ij}^+ := \lambda_{ij}$

Experiments

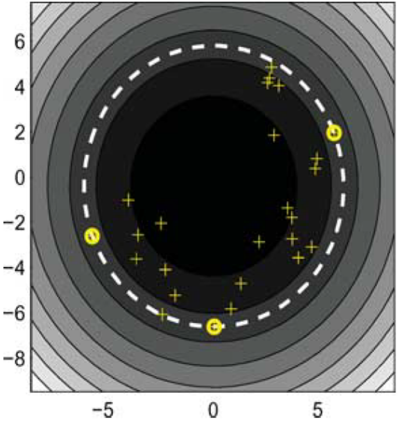


ν, width c	0.5, 0.5	0.5, 0.5	0.1, 0.5	0.5, 0.1

Consider the training data is provided:

$$x_1, x_2, \cdots, x_m \subset X.$$

where m is the number of the observations and X some space (i.e. compact subset of $\mathbb{R}^n$).



Take the decision function as $f(x) = \text{sgn}(\|c - x_i\| - R)$, then finding the optimal paramaters of the hypersphere reduces to solving:

$$\min_{c, R \geq 0, \xi \in \mathbb{R}^m} R^2 + C \sum_{i=1}^{m} \xi_i$$

$$\text{s.t.} \; \|c - x_i\| \leq R^2 + \xi_i, \xi_i \geq 0 \quad \forall i \in \{1, \cdots, m\}.$$

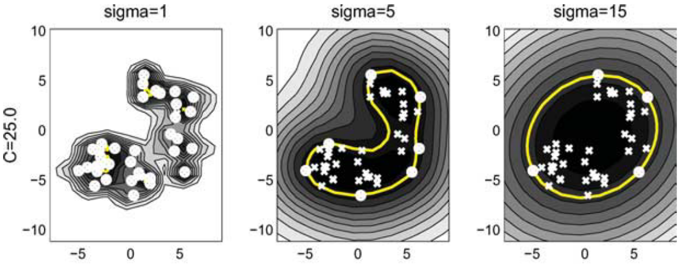
The dual problem:

$$\max_{\lambda} \quad -\frac{1}{2}\lambda^T X^T X \lambda + \lambda^T diag(X^T X)$$
$$\text{s.l. } e^T\lambda = 1,\ 0 \leq \lambda_i \leq C.$$

where $X = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix}$.

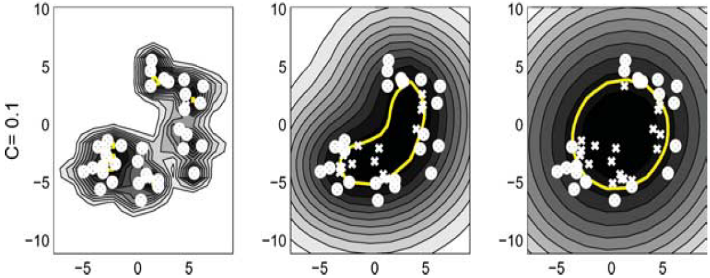
1. $\|x - c\|^2 < R^2 \rightarrow \lambda_i = 0, \gamma_i = 0$
2. $\|x - c\|^2 = R^2 \rightarrow 0 < \lambda_i < C, \gamma_i = 0$
3. $\|x - c\|^2 > R^2 \rightarrow \lambda_i = C, \gamma_i > 0$

When the data is normalized then SVDD is equivalent with OCSVM.



- large C means high penalty of outlyingness (thus large coverage of data)

Experiments



- large C means high penalty of outlyingness (thus large coverage of data)
- low C means high margin of the hyperplane (thus large robustness)
- however, any datapoint has a certain influence on the decision boundary

- One class classification: OC-SVM, SVDD. Algorithms.
- **Robust versions. Algorithms.**

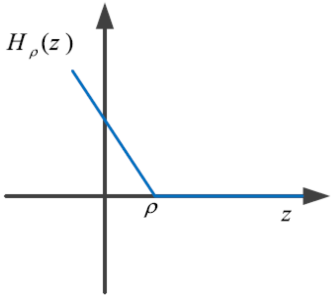
$$\min_{w,\xi\in\mathbb{R}^m,\rho\in\mathbb{R}} \frac{1}{2}\|w\|^2 + \frac{1}{m\nu}\sum_{i=1}^{m}\xi_i - \rho$$

$$\text{s.t. } \langle w, x_i\rangle \geq \rho - \xi_i, \xi_i \geq 0 \quad \forall i \in \{1,\cdots,m\}.$$

By elimination of ξ we obtain:

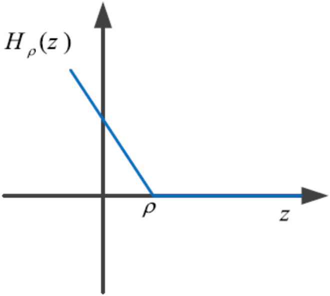
$$\min_{w,\rho\in\mathbb{R}} \frac{1}{2}\|w\|^2 + \frac{1}{m\nu}\sum_{i=1}^{m}\underbrace{\max\{0,\rho-\langle w,x_i\rangle\}}_{H_\rho(\langle w,x_i\rangle)} - \rho.$$

We denote hinge penalty (convex) function: $H_\rho(z) := \max\{0, \rho - z\}$.



a) The Hinge Loss
for OC-SVM

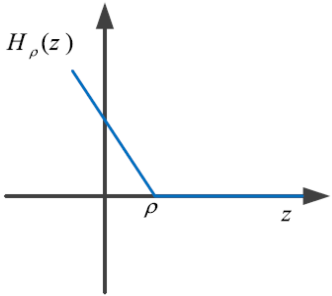
- if a datapoint falls above the hyperplane $w^T z \geq \rho$, then no penalty $H_\rho(z) = 0$



a) The Hinge Loss for OC-SVM

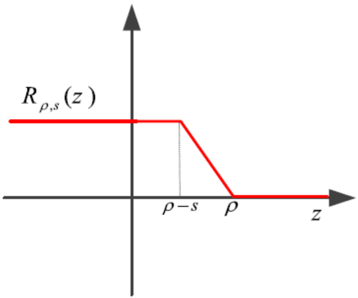
- if a datapoint falls above the hyperplane $w^T z \geq \rho$, then no penalty $H_\rho(z) = 0$
- otherwise, if $w^T z < \rho$, then a penalty corresponding to the distance of this point to the hyperplane will be applied

One class classification - reformulation



a) The Hinge Loss
for OC-SVM

- if a datapoint falls above the hyperplane $w^T z \geq \rho$, then no penalty $H_\rho(z) = 0$
- otherwise, if $w^T z < \rho$, then a penalty corresponding to the distance of this point to the hyperplane will be applied
- one can "robustify" $H_\rho(z)$ by limiting the penalty to a given threshold



b)The Ramp Loss for OC-SVM

- Ramp function $R_{\rho,s}(z) = \begin{cases} 0, & \text{if } z \geq \rho \\ \rho - z, & \text{if } \rho - s < z < \rho \\ s, & \text{if } z < \rho - s. \end{cases}$

Given parameters s, ν then

$$\min_{w,\rho\in\mathbb{R}} \frac{1}{2}\|w\|^2 + \frac{1}{m\nu}\sum_{i=1}^{m} R_{\rho,s}(\langle w, x_i\rangle) - \rho.$$

- This new problem is nonconvex nondifferentiable

Given parameters s, ν then

$$\min_{w,\rho\in\mathbb{R}} \frac{1}{2}\|w\|^2 + \frac{1}{m\nu}\sum_{i=1}^{m} R_{\rho,s}(\langle w, x_i\rangle) - \rho.$$

- This new problem is nonconvex nondifferentiable
- Notice that $R_{\rho,s}(z) = H_\rho(z) - H_{\rho-s}(z)$ (difference of convex function)

Given parameters s, ν then

$$\min_{w,\rho\in\mathbb{R}} \frac{1}{2}\|w\|^2 + \frac{1}{m\nu}\sum_{i=1}^{m} R_{\rho,s}(\langle w, x_i\rangle) - \rho.$$

- This new problem is nonconvex nondifferentiable
- Notice that $R_{\rho,s}(z) = H_\rho(z) - H_{\rho-s}(z)$ (difference of convex function)
- Based on this observation we can derive a simple iterative first-order algorithm.

Given parameters s, ν then

$$\begin{aligned}
&\min_{w,\rho\in\mathbb{R}} \frac{1}{2}\|w\|^2 + \frac{1}{m\nu}\sum_{i=1}^{m} R_{\rho,s}(\langle w, x_i\rangle) - \rho \\
&= \underbrace{\frac{1}{2}\|w\|^2 + \frac{1}{m\nu}\sum_{i=1}^{m} H_{\rho}(\langle w, x_i\rangle) - \rho}_{convex} - \underbrace{\frac{1}{m\nu}\sum_{i=1}^{m} H_{\rho-s}(\langle w, x_i\rangle)}_{convex}.
\end{aligned}$$

DC algorithm

1. Initialize (w_1, ρ_1) and $k := 0$
2. $(w_{k+1}, \rho_{k+1}) =$
$\arg\min \frac{1}{2}\|w\|^2 + \frac{1}{m\nu}\sum_{i=1}^{m} H_\rho(\langle w, x_i\rangle) - \rho - \frac{1}{m\nu}\sum_{i=1}^{m}\langle \begin{bmatrix} x_i \\ 1 \end{bmatrix} H'_{\rho_k - s}(\langle w_k, x_i\rangle), (w, \rho)\rangle$
3. If (w_{k+1}, ρ_{k+1}) satisfies the convergence criterion, then STOP; otherwise, $k := k + 1$ and reiterate.

Dual DC algorithm

1. Compute $\delta_i = \begin{cases} -\frac{1}{m\nu} & \rho - (w^k)^T\phi(x_i) > s \\ 0, & \text{otherwise} \end{cases}$.
2. $\lambda^{k+1} := \max_{\lambda} \; -\frac{1}{2}\lambda^T K\lambda$ s.l. $e^T\lambda = 1, \; -\nu m\delta_i \le \lambda_i \le \frac{1}{m\nu} - \nu m\delta_i$.
3. If λ^{k+1} satisfies the convergence criterion, then STOP; otherwise, $k := k + 1$ and reiterate.

- If the number of iterations is T then Dual DC solves T QP dual problems.

Dual DC algorithm

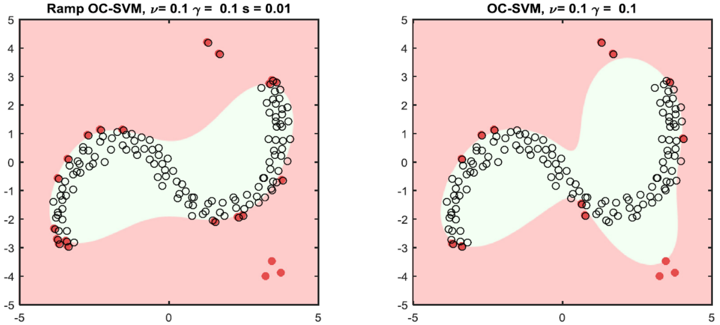
1. Compute $\delta_i = \begin{cases} -\frac{1}{m\nu} & \rho - (w^k)^T \phi(x_i) > s \\ 0, & \text{otherwise} \end{cases}.$
2. $\lambda^{k+1} := \max_{\lambda} \ -\frac{1}{2}\lambda^T K \lambda$ s.l. $e^T \lambda = 1, \ -\nu m \delta_i \leq \lambda_i \leq \frac{1}{m\nu} - \nu m \delta_i.$
3. If λ^{k+1} satisfies the convergence criterion, then STOP; otherwise, $k := k + 1$ and reiterate.

- If the number of iterations is T then Dual DC solves T QP dual problems.
- DC provides the pair λ^*, ρ^*.

Dual DC algorithm

1. Compute $\delta_i = \begin{cases} -\frac{1}{m\nu} & \rho - (w^k)^T \phi(x_i) > s \\ 0, & \text{otherwise} \end{cases}$.
2. $\lambda^{k+1} := \max\limits_{\lambda} \; -\frac{1}{2}\lambda^T K \lambda$ s.l. $e^T\lambda = 1, \; -\nu m \delta_i \leq \lambda_i \leq \frac{1}{m\nu} - \nu m \delta_i$.
3. If λ^{k+1} satisfies the convergence criterion, then STOP; otherwise, $k := k+1$ and reiterate.

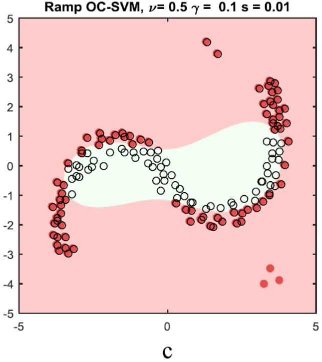
- If the number of iterations is T then Dual DC solves T QP dual problems.
- DC provides the pair λ^*, ρ^*.
- Test on new sample x: evaluate $sgn\left(\sum_i \lambda_i^* k(x_i, x) - \rho^*\right)$

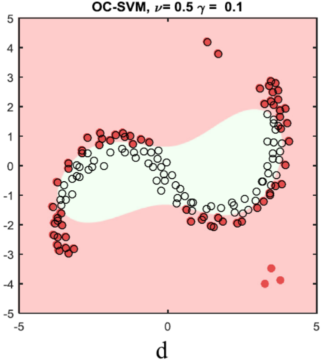


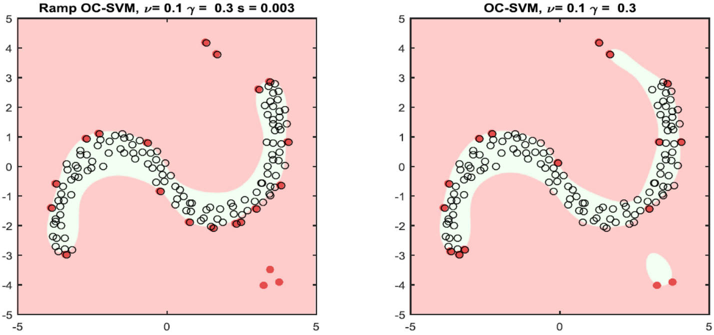
a b

- ν estimate the ratio of outliers
- outliers have a lower impact over Ramp-OCSVM

Experiments (synthetic in 2D)



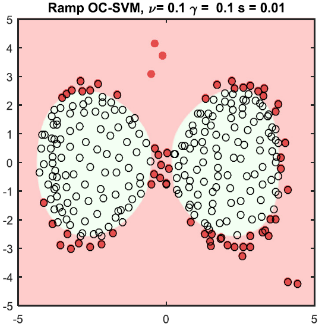




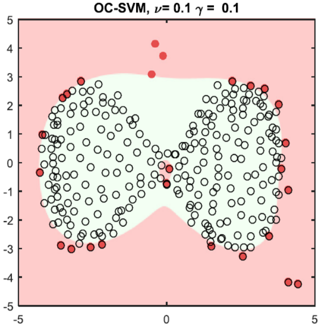
e f

- for small ν OCSVM shift towards outliers
- Ramp-OCSVM controls this shifting through parameter s

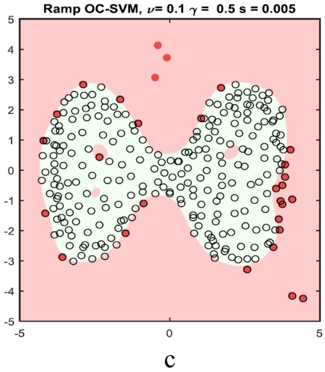
Experiments (synthetic in 2D)

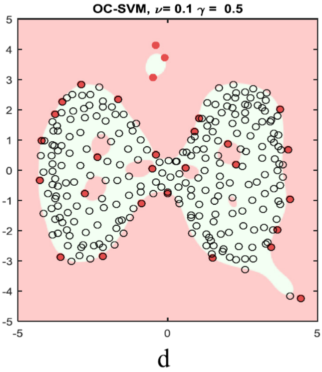


a



b





- behaviour comparison against changing the kernel parameter γ

1. Tian, Yingjie, et al. ”Ramp loss one-class support vector machine; a robust and effective approach to anomaly detection problems.” Neurocomputing 310 (2018): 223-235.
2. Alam, Shamshe, et al. ”One-class support vector classifiers: A survey.” Knowledge-Based Systems 196 (2020): 105754.
3. Tax, David MJ, and Robert PW Duin. ”Support vector data description.” Machine learning 54 (2004): 45-66.
4. Kaiser, Ferdinand. Robust Support Vector Machines For Implicit Outlier Removal. MS thesis. 2013.
5. Schölkopf, Bernhard, et al. ”Estimating the support of a high-dimensional distribution.” Neural computation 13.7 (2001): 1443-1471.