Support vector machines

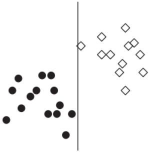
Victor Kitov

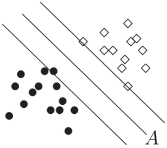
v.v.kitov@yandex.ru

Table of contents

1. Support vector machines
   - Linearly separable case
   - Linearly non-separable case

Support vector machines


(a)



(b)

Support vector machines

(a) (b)

Main idea

Select hyperplane maximizing the margin - the sum of distances from nearest ω_1 object to hyperplane and from nearest ω_2 object to hyperplane.

Support vector machines

Objects x_i for $i = 1, 2, ...n$ lie at distance $b/|w|$ from discriminant hyperplane if

$$\begin{cases} x_i^T w + w_0 \geq b, & y_i = +1 \\ x_i^T w + w_0 \leq -b & y_i = -1 \end{cases} \quad i = 1, 2, ...N.$$

This can be rewritten as

$$y_i(x_i^T w + w_0) \geq b, \quad i = 1, 2, ...N.$$

The margin is equal to $2b/|w|$. Since w, w_0 and b are defined up to multiplication constant, we can set $b = 1$.

Problem statement

Problem statement:

$$\begin{cases} \frac{1}{2}w^T w \to \min\limits_{w,w_0} \\ y_i(x_i^T w + w_0) \geq 1, \quad i = 1, 2, ...N. \end{cases}$$

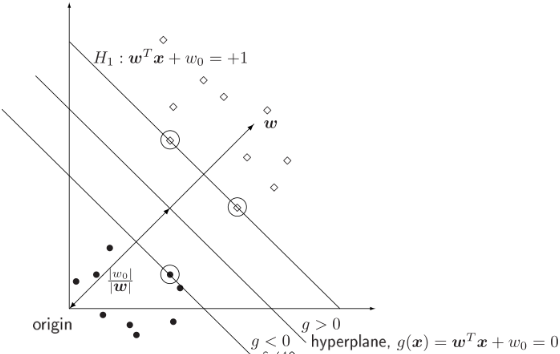
Support vectors

non-informative observations: $y_i(x_i^T w + w_0) > 1$

- do not affect the solution

support vectors: $y_i(x_i^T w + w_0) = 1$

- lie at distance $1/|w|$ to separating hyperplane
- affect the the solution.



Solution

Denote $\mathcal{SV}$ - the set of indexes of support vectors.
For some α_i (which stand for dual variables) weights are equal to:

$$w = \sum_{i \in \mathcal{SV}} \alpha_i y_i x_i$$

w_0 can be found from any edge equality for support vectors:

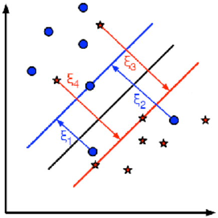
$$y_i(x_i^T w + w_0) = 1, \; i \in \mathcal{SV}$$

Solution from summation over n_{SV} equation provides a more robust estimate of w_0:

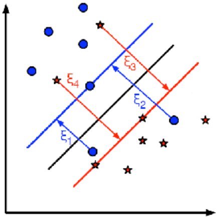
$$n_{\mathcal{SV}} w_0 + \sum_{i \in \mathcal{SV}} x_i^T w = \sum_{i \in \mathcal{SV}} y_i$$

1. Support vector machines
   - Linearly separable case
   - Linearly non-separable case

Linearly non-separable case

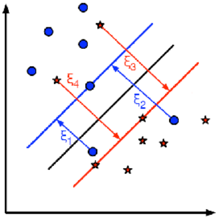


Linearly non-separable case



$$\begin{cases} \frac{1}{2} w^T w \to \min_{w, w_0} \\ y_i(x_i^T w + w_0) \geq 1, \quad i = 1, 2, \ldots N. \end{cases}$$

Linearly non-separable case



$$\begin{cases} \frac{1}{2} w^T w \to \min\limits_{w, w_0} \\ y_i(x_i^T w + w_0) \geq 1, \quad i = 1, 2, \ldots N. \end{cases}$$

Problem

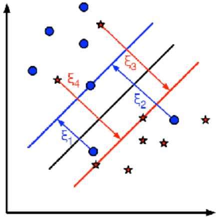
Constraints become incompatible and give empty set!

Linearly non-separable case

No separating hyperplane exists. Errors are permitted by including slack variables ξ_i:

$$\begin{cases} \frac{1}{2}w^T w + C\sum_{i=1}^N \xi_i \to \min_{w,\xi} \\ y_i(w^T x_i + w_0) \ge 1 - \xi_i,\ i = 1, 2, \dots N \\ \xi_i \ge 0,\ i = 1, 2, \dots N \end{cases}$$

- Parameter C is the cost for misclassification and controls the bias-variance trade-off.
- It is chosen on validation set.
- *Other penalties are possible, e.g.* $C\sum_i \xi_i^2$.



Classification of training objects

- **Non-informative objects:**
  - $y_i(w^T x_i + w_0) > 1$
- **Support vectors SV:**
  - $y_i(w^T x_i + w_0) \leq 1$
  - **boundary support vectors $\widetilde{SV}$:**
    - $y_i(w^T x_i + w_0) = 1$
  - **violating support vectors:**
    - $y_i(w^T x_i + w_0)] > 0$: violating support vector is correctly classified.
    - $y_i(w^T x_i + w_0)] < 0$: violating support vector is misclassified.

Solution

Denote $\mathcal{SV}$ - the set of indexes of support vectors with $\alpha_i > 0$ ($\Leftrightarrow y(w^T x_i + w_0) = 1 - \xi_i$) and $\widetilde{\mathcal{SV}}$ - the set of indexes of support vectors with $\alpha_i \in (0, C)$ ($\Leftrightarrow \xi_i = 0$, $y(w^T x_i + w_0) = 1$)

Optimal α_i determine weights directly:

$$w = \sum_{i \in \mathcal{SV}} \alpha_i y_i x_i$$

w_0 can be found from any edge equality for support vectors:

$$y_i(x_i^T w + w_0) = 1,\ i \in \widetilde{\mathcal{SV}}$$

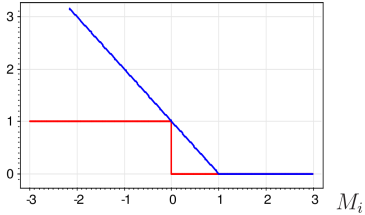
Solution from summation of equations for each $i \in \widetilde{SV}$ provides a more robust estimate of w_0:

$$n_{\widetilde{\mathcal{SV}}} w_0 + \sum_{i \in \widetilde{\mathcal{SV}}} x_i^T w = \sum_{i \in \widetilde{\mathcal{SV}}} y_i$$

Another view on SVM

Optimization problem:

$$\begin{cases} \frac{1}{2}w^T w + C\sum_{i=1}^N \xi_i \to \min_{w,\xi} \\ y_i(w^T x_i + w_0) = M_i(w, w_0) \ge 1 - \xi_i, \\ \xi_i \ge 0,\ i = 1, 2, ...N \end{cases}$$



can be rewritten as

$$\frac{1}{2C}|w|^2 + \sum_{i=1}^N [1 - M_i(w, w_0)]_+ \to \min_{w,\xi}$$

Thus SVM is linear discriminant function with cost approximated with $\mathcal{L}(M) = [1 - M]_+$ and L_2 regularization.

Table of contents

Linear SVM reminder

- Solution for weights:

$$w = \sum_{i \in \mathcal{SV}} \alpha_i y_i x_i$$

Discriminant function

$$g(x) = \sum_{i \in \mathcal{SV}} \alpha_i y_i \langle x_i, x \rangle + w_0$$

$$w_0 = \frac{1}{n_{\widetilde{\mathcal{SV}}}} \left(\sum_{j \in \widetilde{\mathcal{SV}}} y_j - \sum_{j \in \widetilde{\mathcal{SV}}} \sum_{i \in \mathcal{SV}} \alpha_i y_i \langle x_i, x_j \rangle \right)$$

where $SV = \{i : y_i(x_i^T w + w_0) \leq 1)\}$ are indexes of all support vectors and $\tilde{SV} = \{i : y_i(x_i^T w + w_0) = 1\}$ are boundary support vectors.

Kernel SVM

Discriminant function

$$g(x) = \sum_{i \in \mathcal{SV}} \alpha_i y_i K(x_i, x) + w_0$$

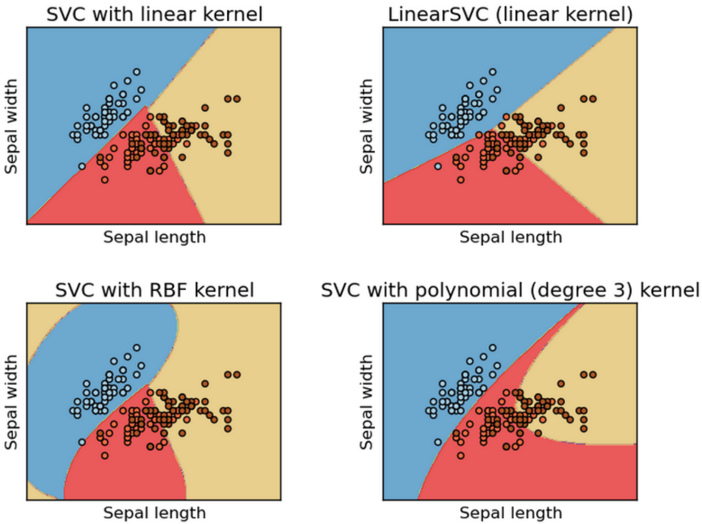
$$w_0 = \frac{1}{n_{\widetilde{\mathcal{SV}}}} \left(\sum_{j \in \widetilde{\mathcal{SV}}} y_j - \sum_{j \in \widetilde{\mathcal{SV}}} \sum_{i \in \mathcal{SV}} \alpha_i y_i K(x_i, x_j) \right)$$

Commonly used kernels

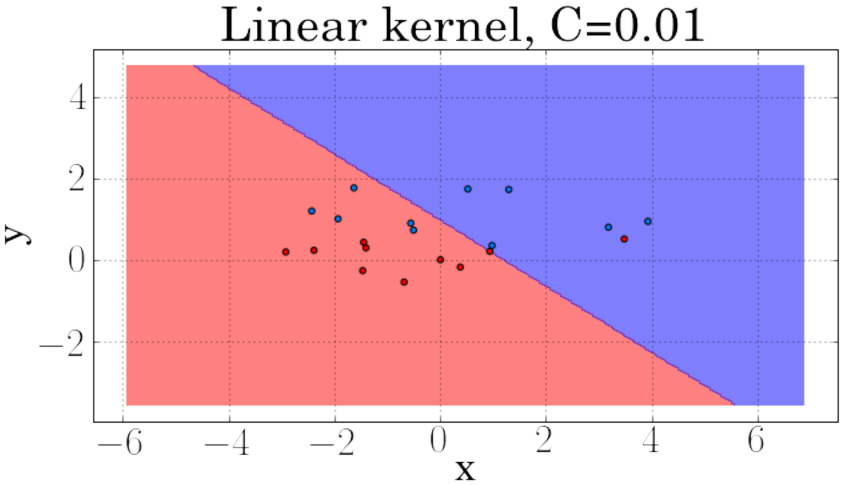
Let x and x' be two objects.

Kernel	Mathematical form
linear	$\langle x, x' \rangle$
polynomial	$(\gamma\langle x, x' \rangle + r)^d$
RBF	$\exp(-\gamma \left\Vert x - x' \right\Vert^2)$
sigmoid	$\text{tangh}(\gamma\langle x, y \rangle + r)$

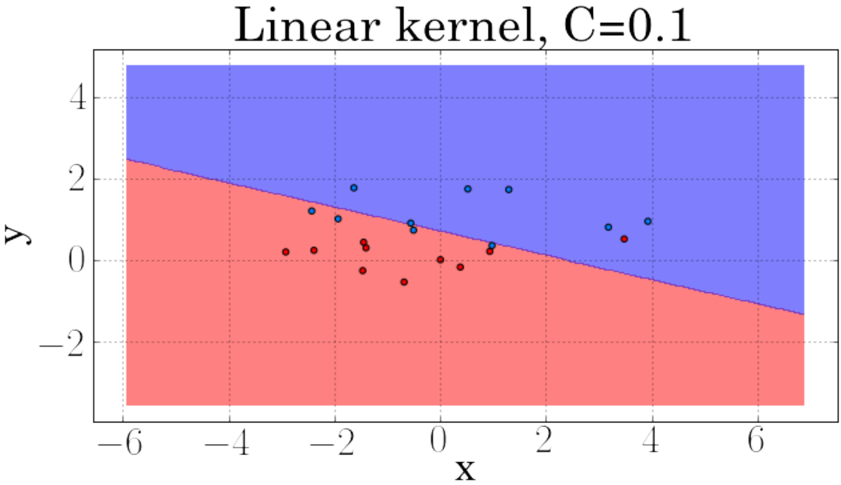
Kernel results



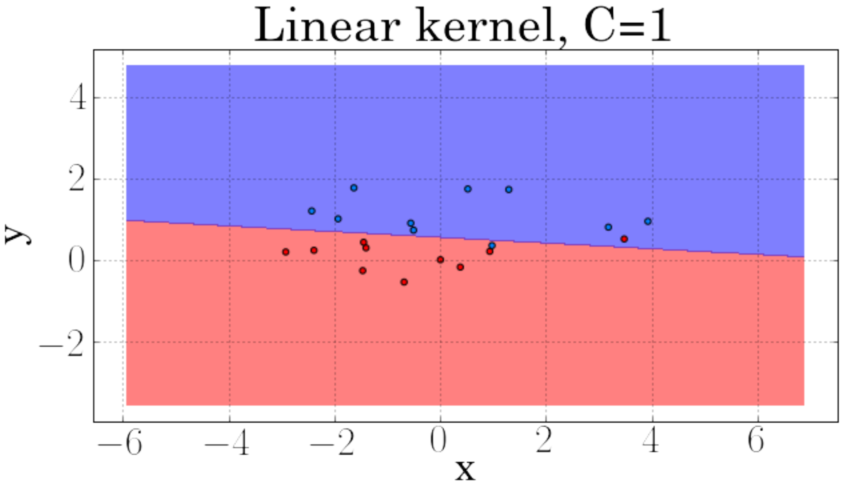
Linear kernel - variable C



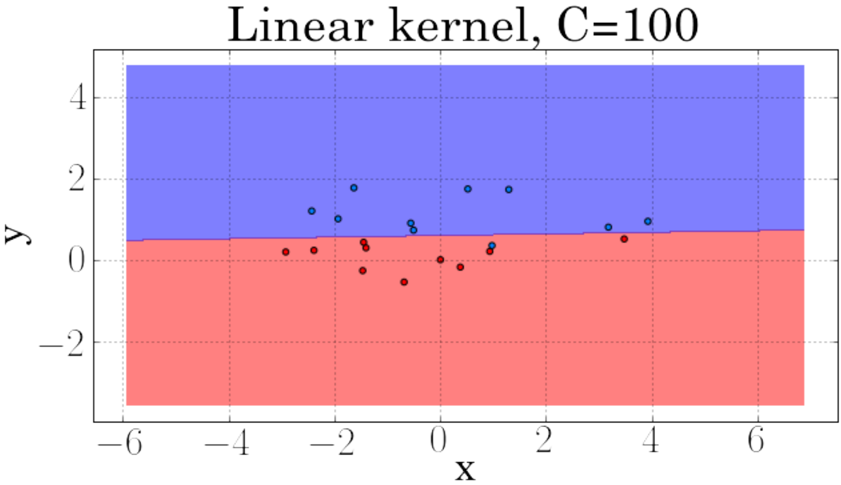
Linear kernel - variable C



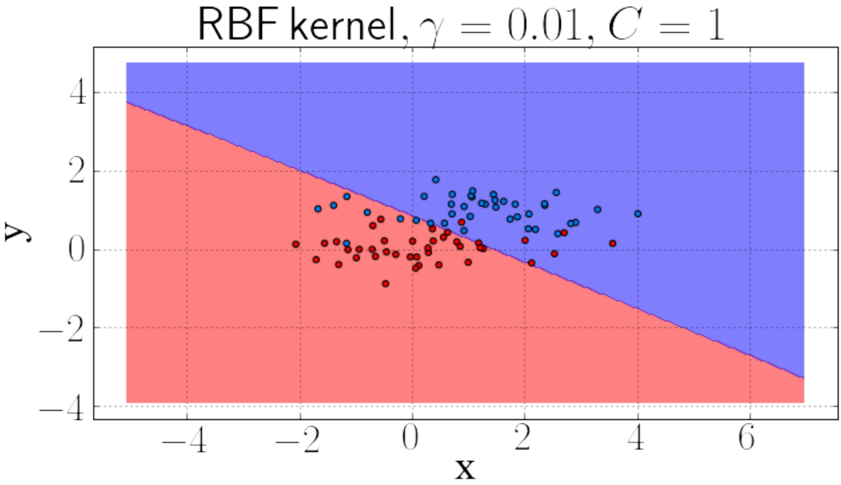
Linear kernel - variable C



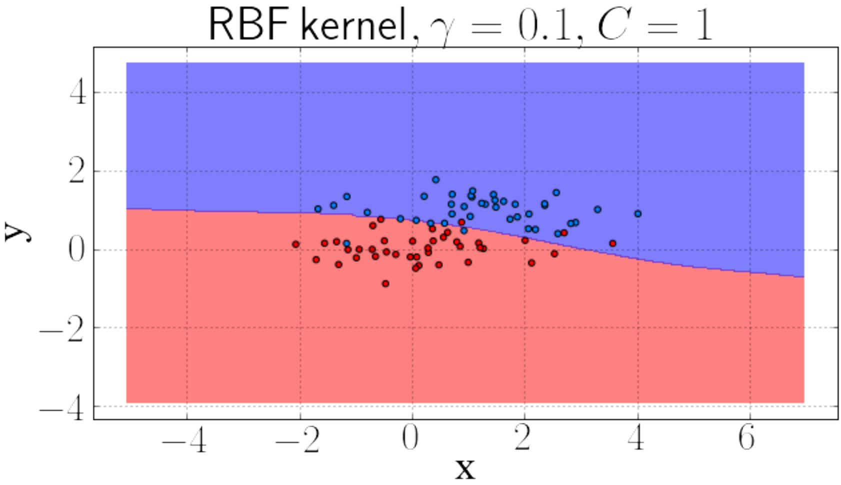
Linear kernel - variable C



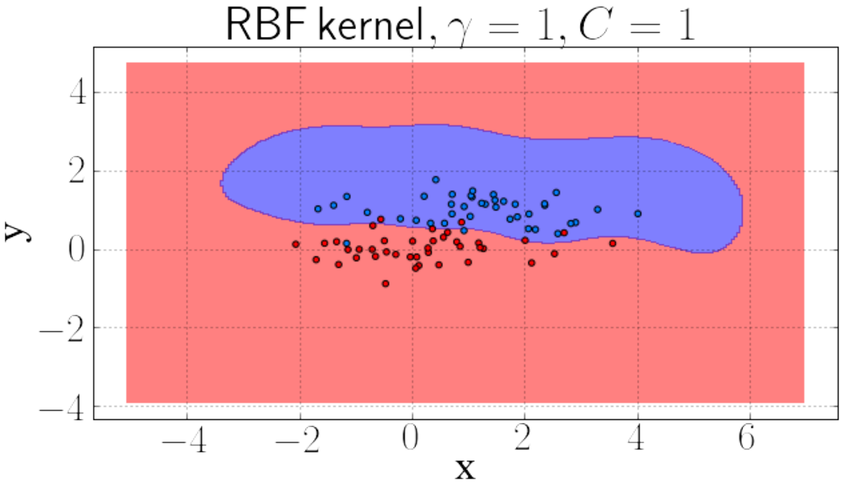
RBF kernel - variable γ



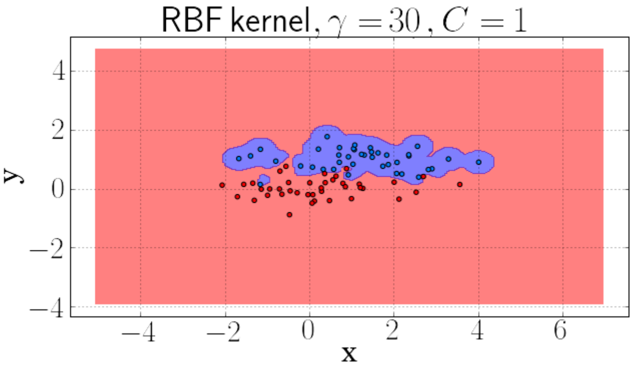
RBF kernel - variable γ



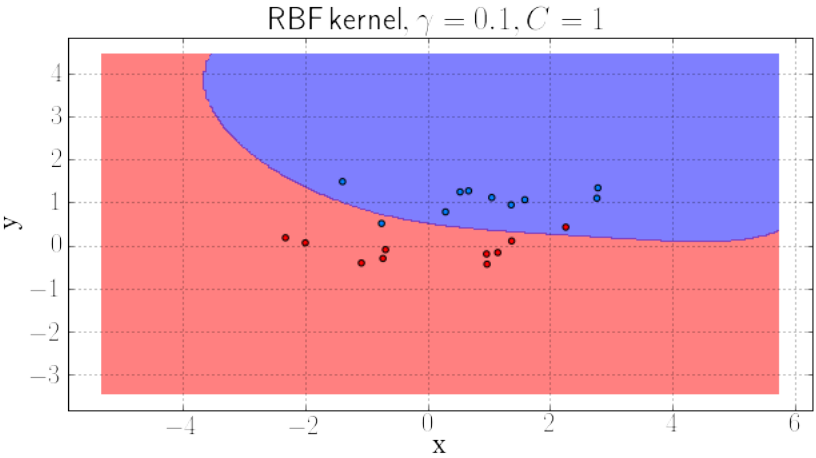
RBF kernel - variable γ



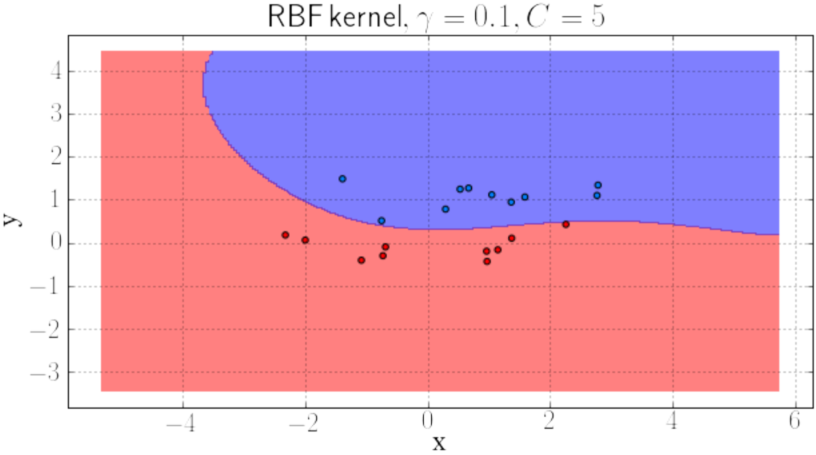
RBF kernel - variable γ



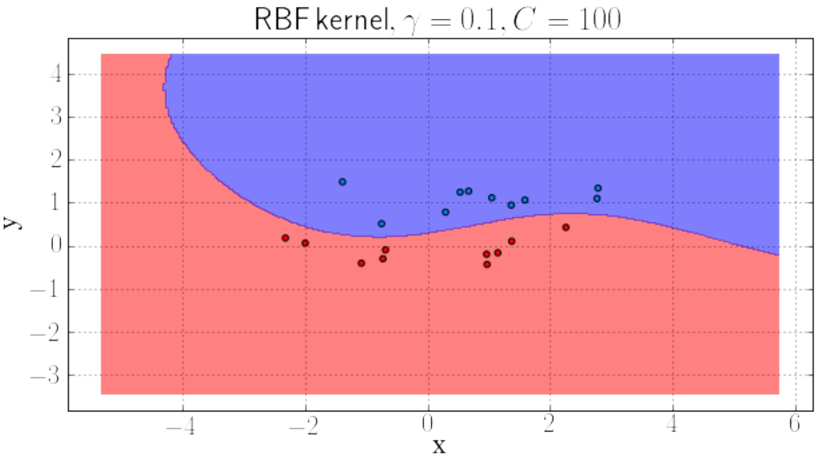
RBF kernel - variable C



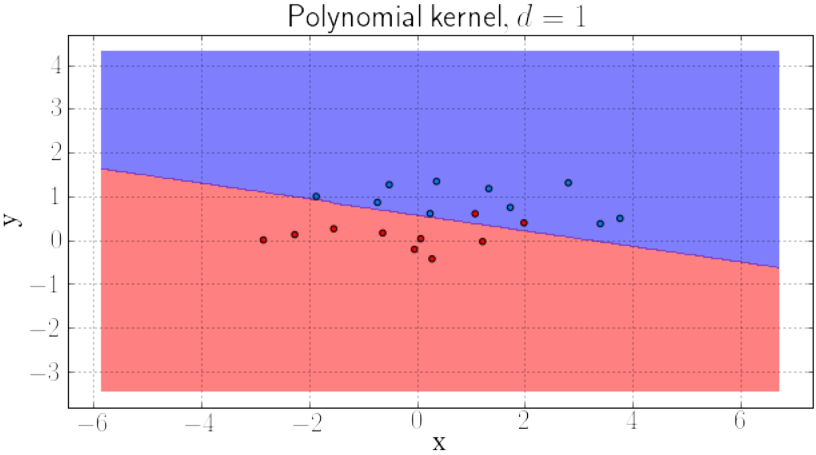
RBF kernel - variable C



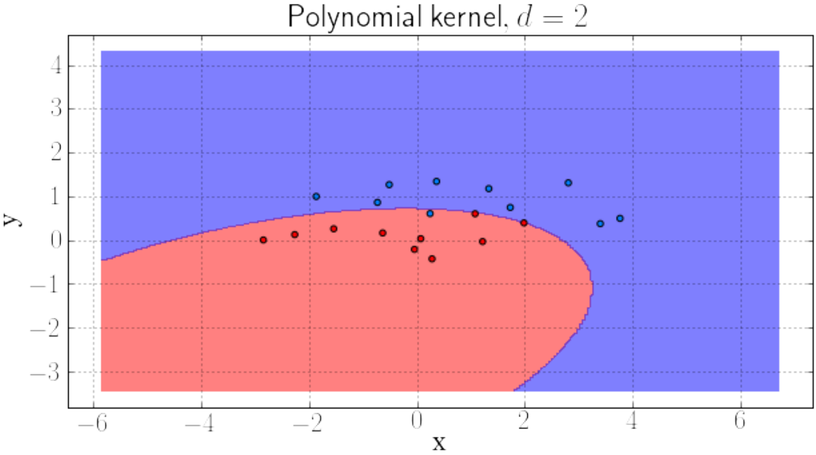
RBF kernel - variable C



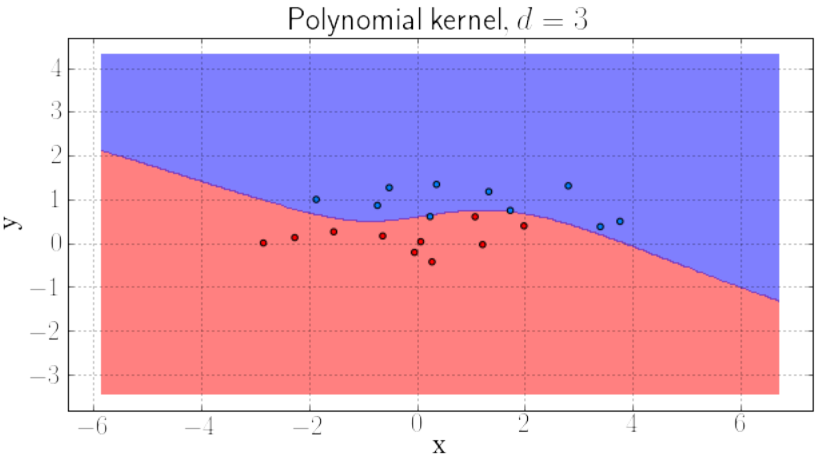
Polynomial kernel - variable d



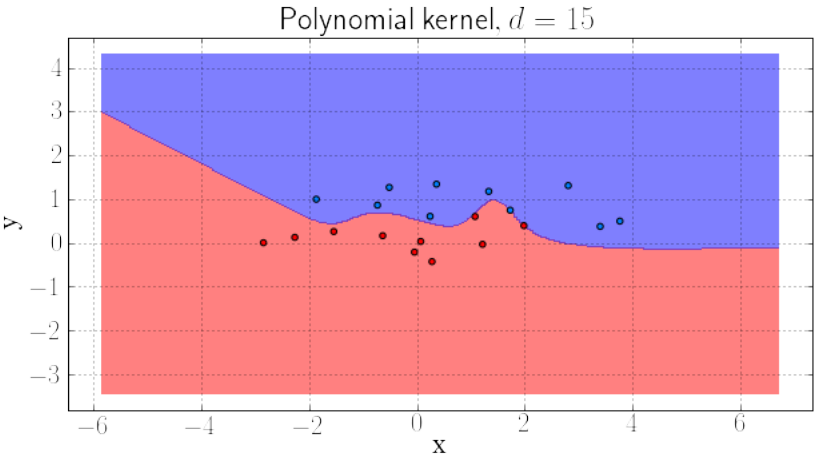
Polynomial kernel - variable d



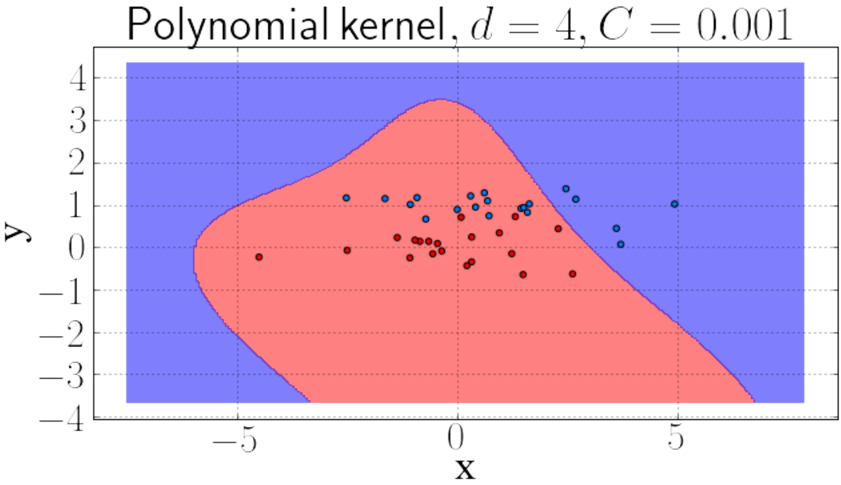
Polynomial kernel - variable d



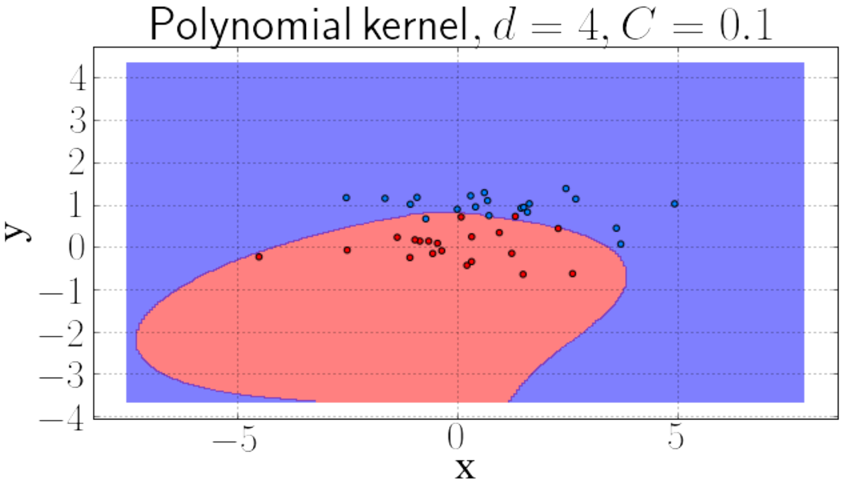
Polynomial kernel - variable d



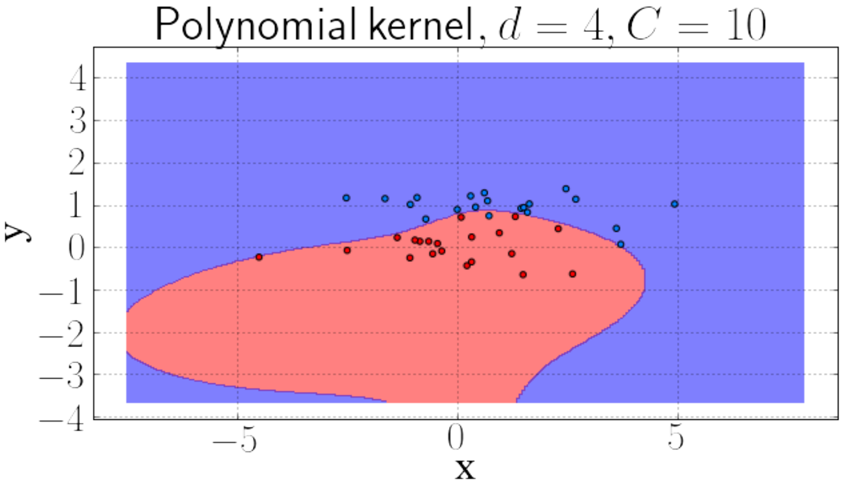
Polynomial kernel - variable C



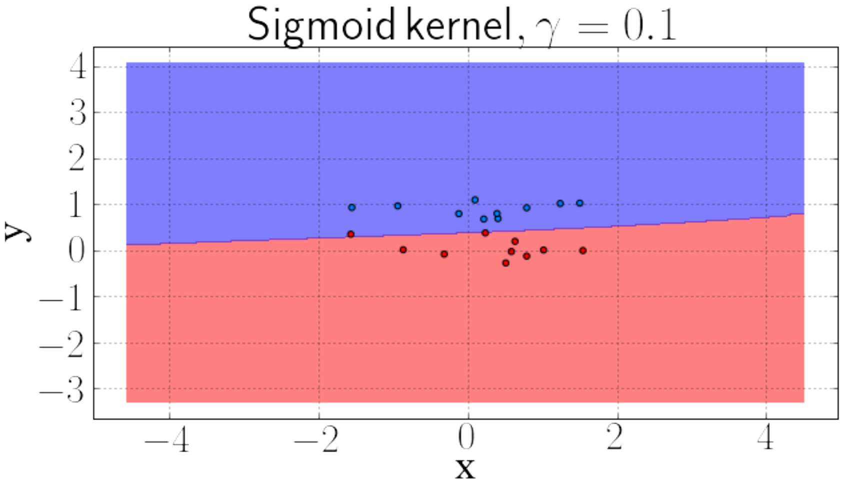
Polynomial kernel - variable C



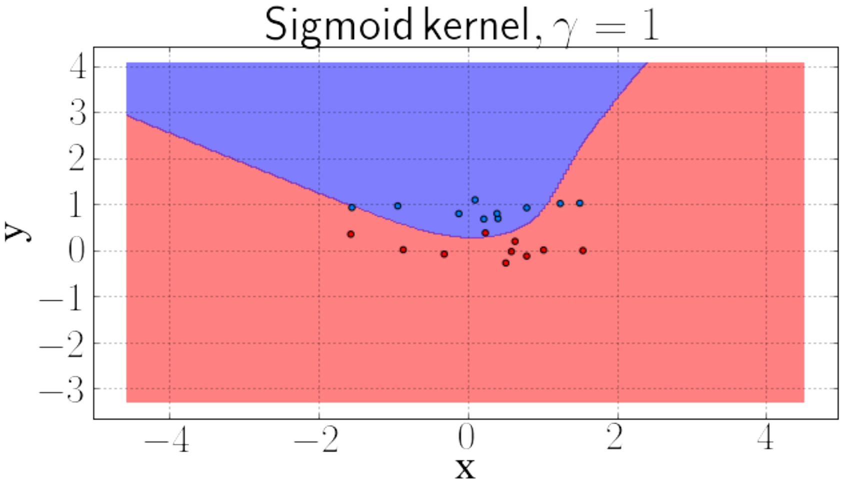
Polynomial kernel - variable C



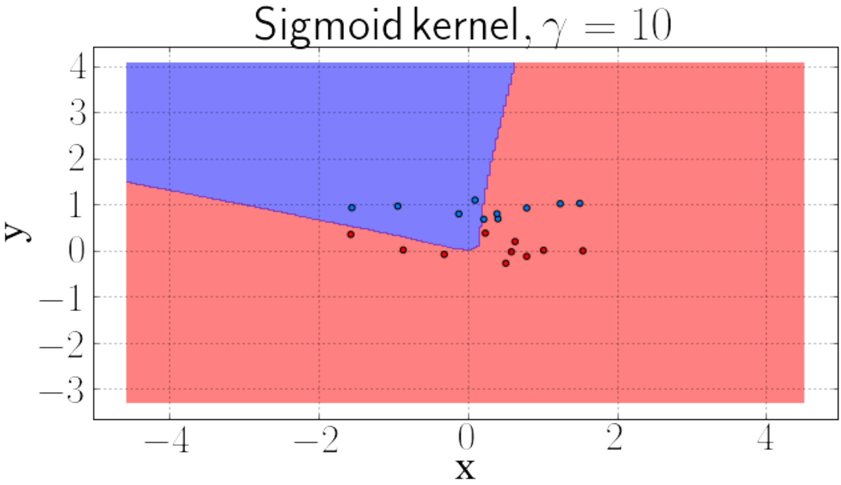
Sigmoid kernel - variable γ



Sigmoid kernel - variable γ



Sigmoid kernel - variable γ



Sigmoid kernel - variable C