Nonlinear classification: Support Vector Machines

COW&MP, Dolní lomná

Tomáš Kalvoda

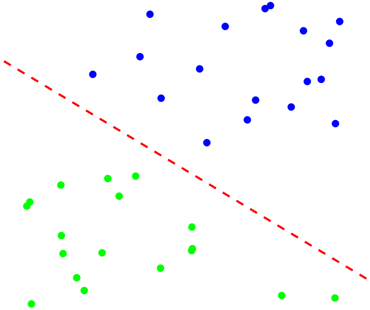
tomas.kalvoda@fit.cvut.cz

KAM FIT ČVUT

May 24, 2013

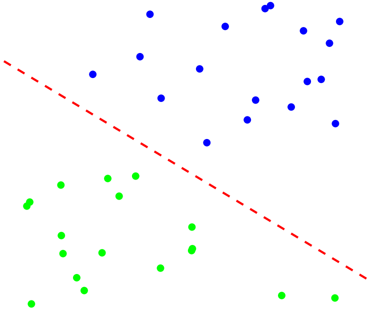
Classification problem (2 classes)

Linear classification

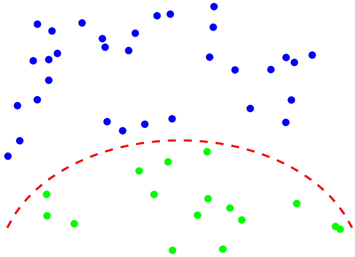


Classification problem (2 classes)

Linear classification



Nonlinear classification



Applications of SVM (90s, [Vapnik, 1979])

- Isolated handwritten digit recognition.
- Object recognition.
- Speaker identification.
- Face detection in images.

Setup

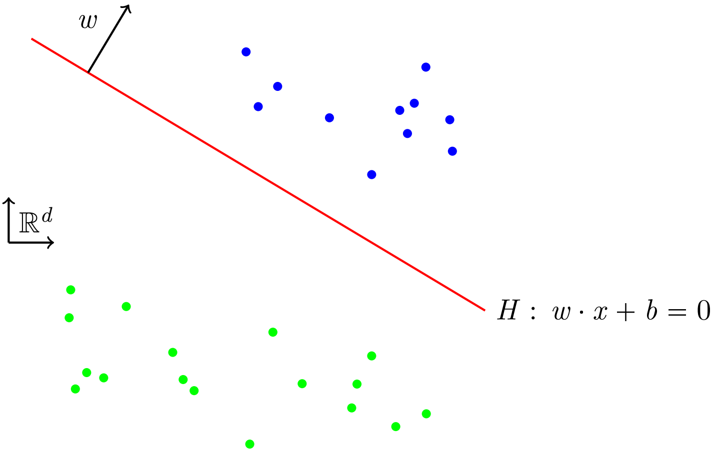
- We are given some initial data and their classes:

$$\{(x_i, y_i) \,:\, i = 1, 2, \ldots, \ell\},$$

where $x_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$ for any $i = 1, 2, \ldots, \ell$.

- Our task: For $x \in \mathbb{R}^d$ decide whether it belongs to $y = 1$ or $y = -1$ class.

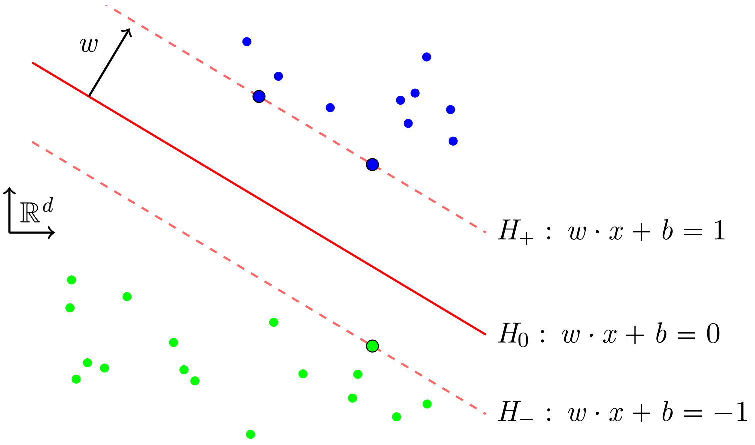
Linear SVM



Assumptions: $\exists w \in \mathbb{R}^d, \ b \in \mathbb{R}: \quad w \cdot x_i + b > 0, \ y_i = 1$

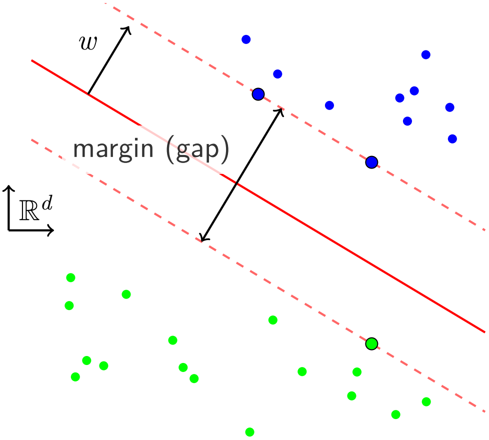
$$w \cdot x_i + b < 0, \ y_i = -1.$$

Linear SVM



Equivalently: $\exists w \in \mathbb{R}^d, \; b \in \mathbb{R} : \quad \begin{aligned} w \cdot x_i + b &\geqslant 1, \; y_i = 1 \\ w \cdot x_i + b &\leqslant -1, \; y_i = -1. \end{aligned}$

Linear SVM



$$\mathrm{dist}(H_+, H_-) = \frac{2}{\|w\|}$$

$$H_+ : \; w \cdot x + b = 1$$

$$H_0 : \; w \cdot x + b = 0$$

$$H_- : \; w \cdot x + b = -1$$

Equivalently: $\exists w \in \mathbb{R}^d, \; b \in \mathbb{R} : \quad \begin{array}{l} w \cdot x_i + b \geqslant 1, \; y_i = 1 \\ w \cdot x_i + b \leqslant -1, \; y_i = -1. \end{array}$

Basic idea of the SVM

Support vectors...

...are those training points x_i that lie on hyperplanes H_+ or H_-.

Goal:

Maximize the margin (gap) between H_+ and H_-.

Primal problem: Summary for the linear separable case

Training data

We are given $\ell \in \mathbb{N}$ samples

$$\{(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\} \,:\, i = 1, 2, \ldots, \ell\}.$$

Primal problem: Summary for the linear separable case

Training data

We are given $\ell \in \mathbb{N}$ samples

$$\left\{(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\} \,:\, i = 1, 2, \ldots, \ell\right\}.$$

Our task

Minimize

$$f : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}, \quad f(w, b) := \frac{1}{2}\|w\|^2$$

subject to ℓ linear inequality constraints

$$g_i(w, b) := y_i(w \cdot x_i + b) - 1 \geqslant 0, \quad i = 1, 2, \ldots, \ell.$$

Primal problem: Summary for the linear separable case

Training data

We are given $\ell \in \mathbb{N}$ samples

$$\{(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\} \,:\, i = 1, 2, \ldots, \ell\}.$$

Our task

Minimize

$$f : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}, \quad f(w, b) := \frac{1}{2}\|w\|^2$$

subject to ℓ linear inequality constraints

$$g_i(w, b) := y_i(w \cdot x_i + b) - 1 \geqslant 0, \quad i = 1, 2, \ldots, \ell.$$

Classification

If w_* and b_* solve the problem above then the class of $x \in \mathbb{R}^d$ is given by

$$\operatorname{sign}(w_* \cdot x + b_*).$$

Lagrange formulation

Minimize $f : \mathbb{R}^n \to \mathbb{R}$ subject to $g_i(z) \geqslant 0$, $g_i : \mathbb{R}^n \to \mathbb{R}$, $1 \leqslant i \leqslant \ell$.

$\downarrow \quad \uparrow$

Necessary

$f, g_i \in C^1$, LICQ

Sufficient

$f \in C^2$, $g \in C^2$ and

$$v^T \nabla_{zz} \mathscr{L}(x; \lambda) v \geqslant 0$$

for all suitable v.

Karush-Kuhn-Tucker (KKT) Conditions

Let $\mathscr{L}(z; \lambda) := f(z) - \lambda^T g(z), \quad z \in \mathbb{R}^n, \lambda \in \mathbb{R}^n$.

$$\begin{aligned} \nabla_z \mathscr{L}(z; \lambda) &= 0, & g(z) &\geqslant 0, \\ \lambda &\geqslant 0, & \lambda_i g_i(z) &= 0 \quad i = 1, 2, \ldots, \ell. \end{aligned}$$

In our case

Lagrangian for our particular case is given by

$$\mathscr{L}(w, b; \lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{\ell} \lambda_i y_i (w \cdot x_i + b) + \sum_{i=1}^{\ell} \lambda_i.$$

and KKT conditions are

$$\begin{aligned} w - \sum_{i=1}^{\ell} \lambda_i y_i x_i &= 0, & \sum_{i=1}^{\ell} \lambda_i y_i &= 0, \\ y_i (w \cdot x_i + b) - 1 &\geqslant 0, & \lambda &\geqslant 0, \\ \lambda_i \big(y_i (w \cdot x_i + b) - 1 \big) &= 0. \end{aligned}$$

Note: Quadratic programming problem

Objective function is quadratic and convex, constraints are linear.

Wolfe dual problem

The problem (P)

Minimize $f : \mathbb{R}^n \to \mathbb{R}$ subject to $g_i(z) \geqslant 0$, where f and $-g_i$ are convex functions.

Wolfe dual problem

The problem (P)

Minimize $f : \mathbb{R}^n \to \mathbb{R}$ subject to $g_i(z) \geqslant 0$, where f and $-g_i$ are convex functions.

Wolfe dual

Maximize $\mathscr{L}(z; \lambda)$ with respect to $z, \lambda \in \mathbb{R}^n$ subject to conditions

$$\nabla_z \mathscr{L}(z; \lambda) = 0 \quad \text{and} \quad \lambda \geqslant 0.$$

Wolfe dual problem

The problem (P)

Minimize $f : \mathbb{R}^n \to \mathbb{R}$ subject to $g_i(z) \geqslant 0$, where f and $-g_i$ are convex functions.

Wolfe dual

Maximize $\mathscr{L}(z; \lambda)$ with respect to $z, \lambda \in \mathbb{R}^n$ subject to conditions

$$\nabla_z \mathscr{L}(z; \lambda) = 0 \quad \text{and} \quad \lambda \geqslant 0.$$

- If $z_* \in \mathbb{R}^n$ solves (P), then it solves the Wolfe dual problem with some $\lambda_* \in \mathbb{R}^n$.
- Every local solution x_* to a convex programming problem is a global solution.

Wolfe dual: Our problem

Wolfe dual

Maximize

$$\mathscr{L}(w, b; \lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{\ell} \lambda_i y_i (w \cdot x_i + b) + \sum_{i=1}^{\ell} \lambda_i$$

with respect to $w \in \mathbb{R}^d$, $b \in \mathbb{R}$, and $\lambda \in \mathbb{R}^d$ subject to

$$w = \sum_{i=1}^{\ell} \lambda_i y_i x_i, \quad \sum_{i=1}^{\ell} \lambda_i y_i = 0, \quad \text{and} \quad \lambda \geqslant 0.$$

Wolfe dual: Our problem

Wolfe dual

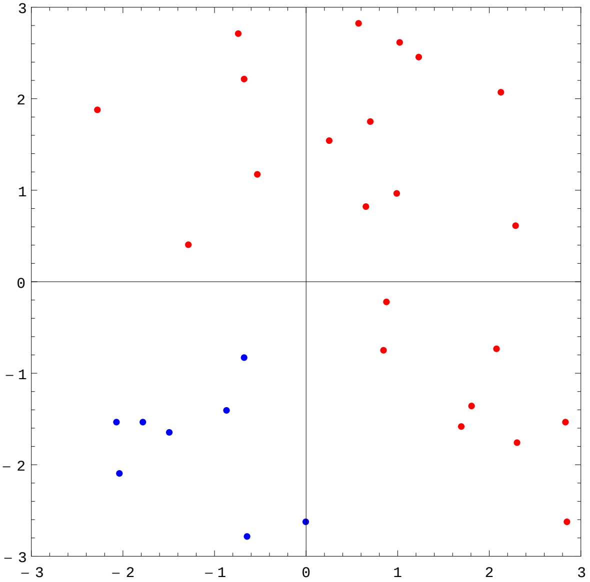
Maximize

$$\mathscr{L}(w, b; \lambda) = \sum_{i=1}^{\ell} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \lambda_i \lambda_j y_i y_j \, x_i \cdot x_j$$

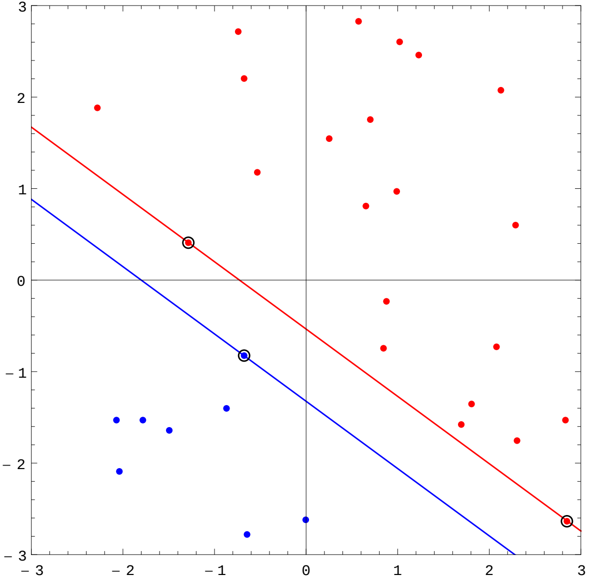
with respect to $w \in \mathbb{R}^d$, $b \in \mathbb{R}$, and $\lambda \in \mathbb{R}^d$ subject to

$$\left(w = \sum_{i=1}^{\ell} \lambda_i y_i x_i \right), \quad \sum_{i=1}^{\ell} \lambda_i y_i = 0, \quad \text{and} \quad \lambda \geqslant 0.$$

Illustration



Illustration



What to do if linear separation is not possible?

- We wish to map our data $x_i \in \mathbb{R}^d$ to some – *real* – Hilbert space $\mathscr{H}$ with inner product $\langle \cdot, \cdot \rangle$,

$$\Phi : \mathbb{R}^d \to \mathscr{H} .$$

- Wolfe dual problem involves only inner products $x_i \cdot x_j$ in $\mathbb{R}^d$.

What to do if linear separation is not possible?

- We wish to map our data $x_i \in \mathbb{R}^d$ to some – *real* – Hilbert space $\mathscr{H}$ with inner product $\langle \cdot, \cdot \rangle$,

$$\Phi : \mathbb{R}^d \to \mathscr{H} .$$

- Wolfe dual problem involves only inner products $x_i \cdot x_j$ in $\mathbb{R}^d$.
- Instead of those products we will have to compute expressions of the form

$$K(x_i, x_j) := \langle \Phi(x_i), \Phi(x_j) \rangle, \quad K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}.$$

What to do if linear separation is not possible?

- We wish to map our data $x_i \in \mathbb{R}^d$ to some – *real* – Hilbert space $\mathscr{H}$ with inner product $\langle \cdot, \cdot \rangle$,

$$\Phi : \mathbb{R}^d \to \mathscr{H}.$$

- Wolfe dual problem involves only inner products $x_i \cdot x_j$ in $\mathbb{R}^d$.
- Instead of those products we will have to compute expressions of the form

$$K(x_i, x_j) := \langle \Phi(x_i), \Phi(x_j) \rangle, \quad K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}.$$

- N.B.:

$$\begin{aligned} K(u, v) &= K(v, u), & u, v \in \mathbb{R}^d \\ \sum_{i,j=1}^{n} c_i c_j K(u_i, u_j) &= \left\| \sum_{i=1}^{n} c_i \Phi(u_i) \right\|^2 \geqslant 0, & n \in \mathbb{N},\ u_i \in \mathbb{R}^d,\ c \in \mathbb{R}^n. \end{aligned}$$

Kernel

Definition

Any $K : X \times X \to \mathbb{R}$ such that

- $K(x,y) = K(y,x)$ for any $x, y \in X$,
- K is positive definite, i.e. for any $n \geqslant 1$, $x_1, \ldots, x_n \in X$ and any $c_1, \ldots, c_n \in \mathbb{R}$

$$\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) \geqslant 0.$$

is called *kernel* on $X \times X$.

- Application to SVM [Boser, Guyon and Vapnik, 1992].

On the other hand. . .

- . . . one can start with a kernel $K : X \times X \to \mathbb{R}$ and ask whether there is a Hilbert space $\mathscr{H}$ and $\Phi : X \to \mathscr{H}$ such that
$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle, \quad x, y \in X.$$
In this case it would be possible to use K only.

- The answer to that question is positive.

RKHS

Theorem

Let X be a separable metric space and $K : X \times X \to \mathbb{R}$ a continuous kernel on $X \times X$. Then there is a separable Hilbert space $\mathscr{H}$ of functions on X and mapping $\Phi : X \to \ell^2 \simeq \mathscr{H}$ such that

$$K(u, v) = \langle \Phi(u), \Phi(v) \rangle_{\ell^2}, \quad u, v \in X.$$

Note

$\mathscr{H}$ is the *Reproducing kernel Hilbert space* (RKHS) associated with K. This terminology is due to the property

$$f(x) = \langle K_x, f \rangle, \quad x \in X, f \in \mathscr{H},$$

where $K_x = K(x, \cdot)$.

Proof

- $V := \mathrm{span}_{\mathbb{R}}\{K_x : x \in X\}$, recall $K_x(y) = K(x, y)$.

Proof

- $V := \mathrm{span}_{\mathbb{R}}\{K_x \,:\, x \in X\}$, recall $K_x(y) = K(x, y)$.
- For $f, g \in V$, $f = \sum_{i=1}^{n} c_i K_{x_i}$, $g = \sum_{j=1}^{m} d_j K_{y_j}$ set

$$\langle f, g\rangle := \sum_{i,j} c_i d_j K(x_i, x_j) = \sum_i c_i g(x_i) = \sum_j d_j f(x_j).$$

$\langle \cdot, \cdot \rangle$ does not depend on the representation of f and g, is bilinear, symmetric and $\langle f, f\rangle \geqslant 0$ for any $f \in V$.

Proof

- $V := \mathrm{span}_{\mathbb{R}}\{K_x : x \in X\}$, recall $K_x(y) = K(x, y)$.
- For $f, g \in V$, $f = \sum_{i=1}^{n} c_i K_{x_i}$, $g = \sum_{j=1}^{m} d_j K_{y_j}$ set

$$\langle f, g \rangle := \sum_{i,j} c_i d_j K(x_i, x_j) = \sum_i c_i g(x_i) = \sum_j d_j f(x_j).$$

  $\langle \cdot, \cdot \rangle$ does not depend on the representation of f and g, is bilinear, symmetric and $\langle f, f \rangle \geqslant 0$ for any $f \in V$.
- $\langle \cdot, \cdot \rangle$ has the reproducing property,

$$\langle f, K_x \rangle = \sum_{j=1}^{1} 1 \cdot f(x) = f(x), \quad f \in V, x \in X.$$

Proof

- $V := \mathrm{span}_{\mathbb{R}}\{K_x \,:\, x \in X\}$, recall $K_x(y) = K(x, y)$.
- For $f, g \in V$, $f = \sum_{i=1}^{n} c_i K_{x_i}$, $g = \sum_{j=1}^{m} d_j K_{y_j}$ set

$$\langle f, g\rangle := \sum_{i,j} c_i d_j K(x_i, x_j) = \sum_i c_i g(x_i) = \sum_j d_j f(x_j).$$

  $\langle \cdot, \cdot \rangle$ does not depend on the representation of f and g, is bilinear, symmetric and $\langle f, f\rangle \geqslant 0$ for any $f \in V$.
- $\langle \cdot, \cdot \rangle$ has the reproducing property,

$$\langle f, K_x\rangle = \sum_{j=1}^{1} 1 \cdot f(x) = f(x), \quad f \in V, x \in X.$$

- Since

$$f(x)^2 = \langle f, K_x\rangle^2 \leqslant \langle f, f\rangle \cdot \langle K_x, K_x\rangle = \|f\|^2 \cdot K(x, x), \quad x \in X,$$

  we conclude that if $\|f\| = 0$ then $f(x) = 0$ for any $x \in X$.

Proof (cont'd)

- The pair $(V, \langle\cdot,\cdot\rangle)$ forms a real pre-Hilbert space.

Proof (cont'd)

- The pair $(V, \langle \cdot, \cdot \rangle)$ forms a real pre-Hilbert space.

- Let $\mathscr{H}$ be the completion of $(V, \langle \cdot, \cdot \rangle)$.

Proof (cont'd)

- The pair $(V, \langle \cdot, \cdot \rangle)$ forms a real pre-Hilbert space.

- Let $\mathscr{H}$ be the completion of $(V, \langle \cdot, \cdot \rangle)$.

- Note that if $\{f_n\}_{n=1}^{\infty}$ is a Cauchy sequence in V then $\{f_n(x)\}_{n=1}^{\infty}$ is a Cauchy sequence in $\mathbb{R}$ for any $x \in \mathbb{R}$. Indeed, recall the inequality

$$(f_n(x) - f_m(x))^2 \leqslant \|f_n - f_m\|^2 \cdot K(x, x).$$

So $\mathscr{H}$ consists of a larger class of real valued functions on X.

Proof (cont'd)

- The pair $(V, \langle \cdot, \cdot \rangle)$ forms a real pre-Hilbert space.

- Let $\mathscr{H}$ be the completion of $(V, \langle \cdot, \cdot \rangle)$.

- Note that if $\{f_n\}_{n=1}^{\infty}$ is a Cauchy sequence in V then $\{f_n(x)\}_{n=1}^{\infty}$ is a Cauchy sequence in $\mathbb{R}$ for any $x \in \mathbb{R}$. Indeed, recall the inequality

$$(f_n(x) - f_m(x))^2 \leqslant \|f_n - f_m\|^2 \cdot K(x, x).$$

  So $\mathscr{H}$ consists of a larger class of real valued functions on X.

- $\mathscr{H}$ has the reproducing property and is separable.

Proof (cont'd)

- Let $\{\phi_i\}_i$ be an orthonormal basis of $\mathscr{H}$. For any $x \in X$ we have the Fourier expansion of $K_x \in V \subset \mathscr{H}$

$$K_x = \sum_i \langle \phi_i, K_x \rangle \phi_i.$$

For any $y \in X$ we obtain

$$K(x, y) = K_x(y) = \sum_i \phi_i(x) \phi_i(y).$$

So $\Phi : X \to l^2(\mathbb{N}, \mathbb{R})$, $\Phi(x) := \{\phi_i(x)\}_i$ $\square$

Problem to be maximized

Wolfe dual with the Kernel

Maximize

$$\sum_{i=1}^{\ell} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \lambda_i \lambda_j y_i y_j K(x_i, x_j)$$

with respect to $\lambda \geqslant 0$ subject to $\sum_{i=1}^{\ell} \lambda_i y_i = 0$.

Classificator

Compute the sign of

$$w \cdot x + b = \sum_{x_i \text{ is s.v.}} \lambda_i y_i x_i \cdot x + b \leftrightarrow \sum_{x_i \text{ is s.v.}} \lambda_i y_i K(x_i, x) + b,$$

where

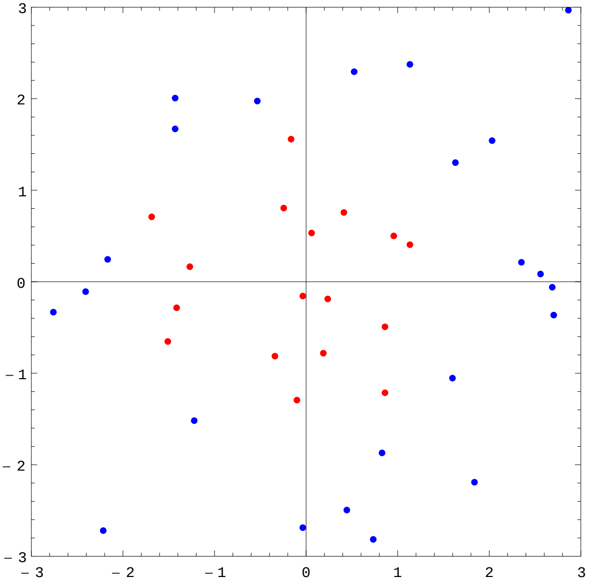
$$b = \frac{1}{y_i} - w \cdot x_i = y_i - \sum_{x_j \text{ is s.v.}} \lambda_j y_j x_j \cdot x_i \leftrightarrow y_i - \sum_{x_j \text{ is s.v.}} \lambda_j y_j K(x_j, x_i).$$

Example

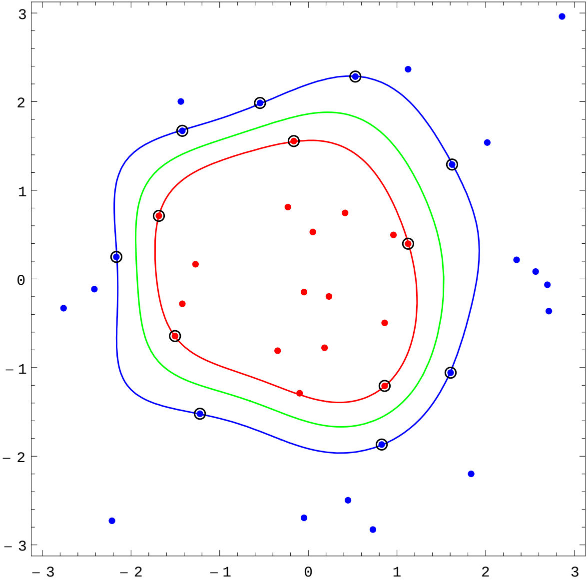
Results for "radial basis kernel"

$$K(u, v) = \exp\left(-\|u - v\|^2 / (2\sigma^2)\right), \quad u, v \in \mathbb{R}^d.$$

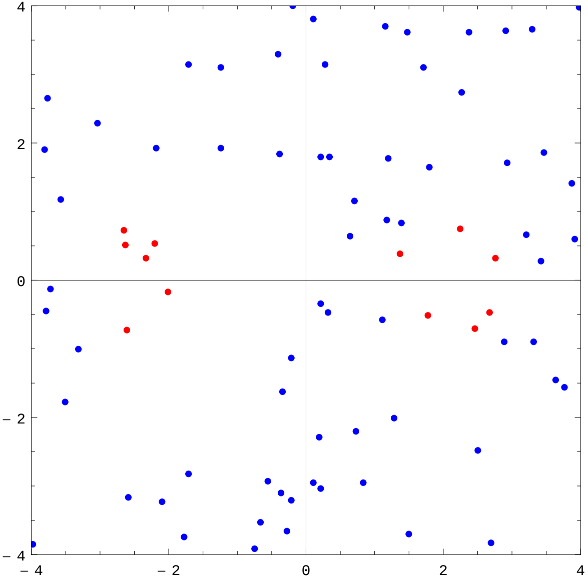
Example



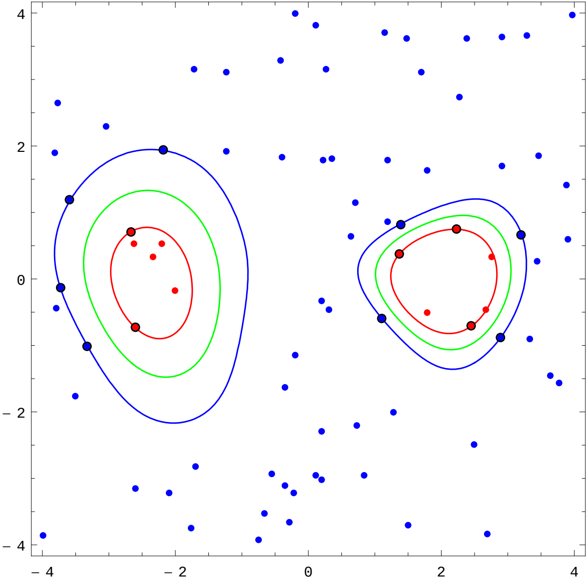
Example



Example



Example



Example

Thank you for your attention.