Freie Universität Berlin

CCNB
Center for Cognitive Neuroscience Berlin

An Introduction to Support Vector Classification

Pattern Recognition in Neuroimaging

Outline

Constrained nonlinear optimization

Support Vector Classification

Discussion

Outline

Constrained nonlinear optimization

Support Vector Classification

Discussion

Constrained Nonlinear Optimization

$$\min_{x \in \mathbb{R}^d} f(x) \text{ subject to } c_i(x) = 0 \ (i \in E) \text{ and } c_i(x) \geq 0 \ (i \in I)$$

Objective function $f: \mathbb{R}^d \to \mathbb{R}$

Equality constraint functions $c_i: \mathbb{R}^d \to \mathbb{R} \quad (i \in E)$

Inequality constraint functions $c_i: \mathbb{R}^d \to \mathbb{R} \quad (i \in I)$

$$E \cup I = \mathbb{N}_k, E \cap I = \emptyset$$

Nocedal J and Wright SJ (2000) *Numerical Optimization 2nd Ed.*

Constrained Nonlinear Optimization

Solution strategy for constrained nonlinear optimization problems

(1) Formulate Lagrange function

$$L : \mathbb{R}^d \times \mathbb{R}^k, (x, \lambda_1, \dots, \lambda_k) \mapsto L(x, \lambda_1, \dots, \lambda_k) \coloneqq f(x) - \sum_{i \in E \cup I} \lambda_i c_i(x)$$

(2) First-order necessary conditions (Karush-Kuhn-Tucker conditions)

Let $f(x^*) \leq f(x) \left(\forall\, x \in N(x)\right), c_i(x^*) = 0\ (i \in E)$ and $c_i(x^*) \geq 0\ (i \in I)$.

Then there exists an Lagrange multiplier $\lambda^* \coloneqq (\lambda_1^*, \dots, \lambda_k^*)^T \in \mathbb{R}^k$ such that

$\nabla_x L(x^*, \lambda^*) = 0$, $\lambda_i^* \geq 0\ (i \in I)$ and $\lambda_i^* c_i(x^*) = 0\ (i \in E \cup I)$

Many numerical algorithms for finding (x^*, λ^*) exist!

Nocedal J and Wright SJ (2000) *Numerical Optimization 2nd Ed.*

Outline

Constrained nonlinear optimization

Support Vector Classification

Discussion

Support Vector Classification

Training Data Set

$$\left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{n} \coloneqq \left\{\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)\right\} \left(x^{(i)} \in \mathbb{R}^m, y^{(i)} \in \{-1,1\}\right)$$

Discriminant Function

$$f\colon \mathbb{R}^m \to \mathbb{R}, x \mapsto f(x) \coloneqq w^T x + w_0 \quad (w \in \mathbb{R}^m, w_0 \in \mathbb{R})$$

Decision Function

$$g\colon \mathbb{R} \to \{-1,1\}, f(x) \mapsto g(f(x)) \coloneqq \begin{cases} -1, & f(x) < 0 \\ 1, & f(x) \geq 0 \end{cases}$$

Hyperplane

$$H \coloneqq \{x \in \mathbb{R}^m | f(x) = 0\} \subset \mathbb{R}^m$$

Support Vector Classification

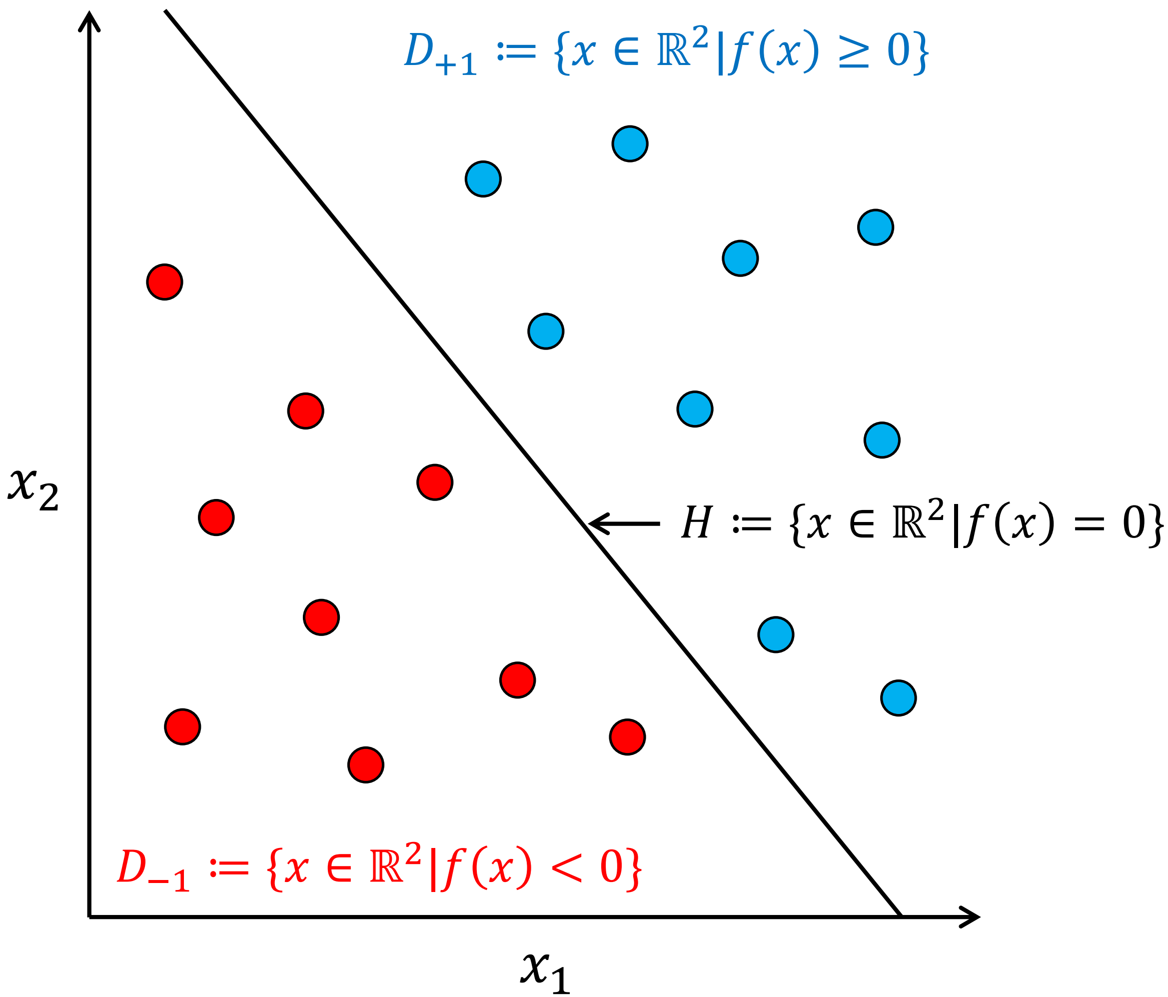
A special case

$m := 2$

$\Rightarrow x^{(i)} = \left(x_1^{(i)}, x_2^{(i)}\right)^T \in \mathbb{R}^2$

● $x^{(i)}$ with $y^{(i)} = -1$ (red)

● $x^{(i)}$ with $y^{(i)} = +1$ (blue)



Support Vector Classification

Fundamental aim of classifier training

Determine $w^* \in \mathbb{R}^m, w_0^* \in \mathbb{R}$ such that training set classification is "optimal"

$\rightarrow$ Parameter point estimation for model $h: \mathbb{R}^m \rightarrow \{-1,1\}, x \mapsto h(x) := g(f(x))$

Geometric properties of discriminant functions and hyperplanes

(1) w is orthogonal to any vector pointing in the direction of the hyperplane

(2) $w_0/\|w\|_2$ is the (direct) distance from the origin to a point on the hyperplane

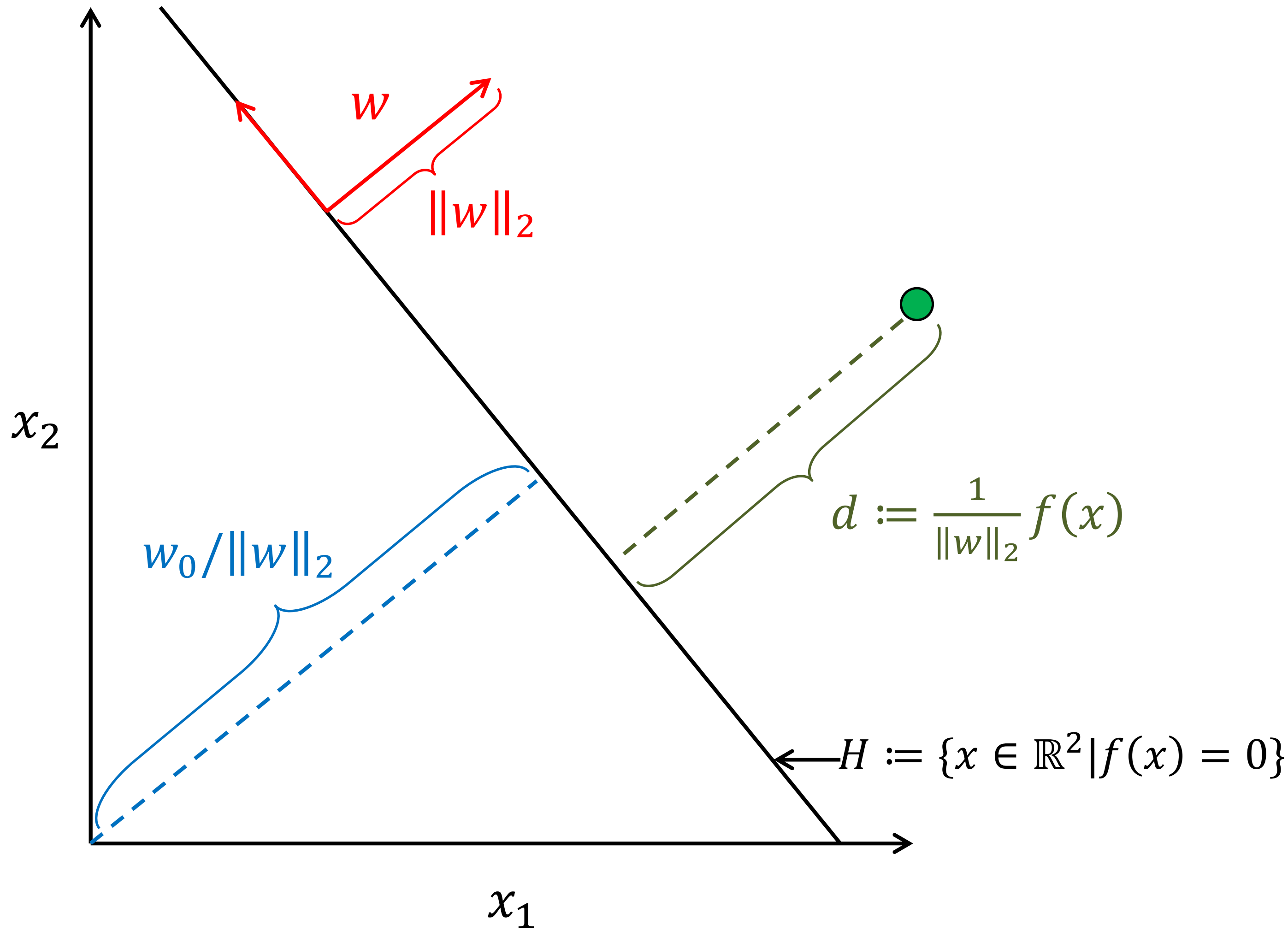
$\Rightarrow$ w and w_0 determine the location and orientation of the hyperplane

(3) The (direct) distance of a point x to the hyperplane is $d = \frac{1}{\|w\|_2} f(x)$

$\Rightarrow$ The discriminant function describes the distance from the hyperplane

(for proofs, see e.g. Ostwald *Probabilistic Models for Functional Neuroimaging 2015*)

Support Vector Classification



(for proofs, see e.g. Ostwald *Probabilistic Models for Functional Neuroimaging 2015*)

Support Vector Classification

Absolute (direct) distance of $x^{(i)}$ from hyperplane

$$\delta^{(i)} \coloneqq \left|d^{(i)}\right| = \left|\frac{1}{\|w\|_2} f\left(x^{(i)}\right)\right| = \frac{y^{(i)}}{\|w\|_2} f\left(x^{(i)}\right) \; (i = 1, \dots, n)$$

Hyperplane Margin

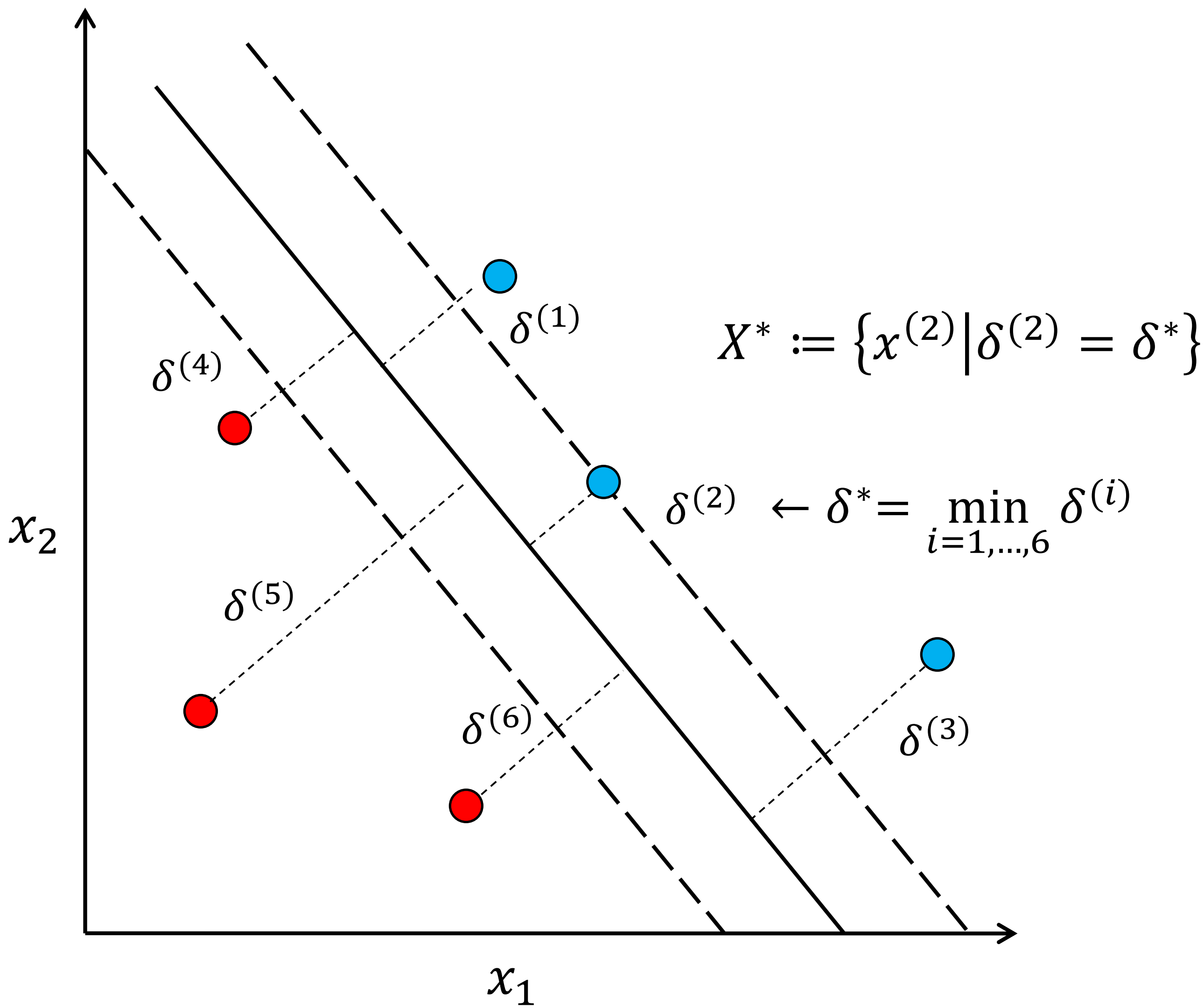
$$\delta^* \coloneqq \min_{i=1,\dots,n} \delta^{(i)}$$

Support Vector Set

$$X^* \coloneqq \left\{x^{(i)} \in \{x^{(1)}, \dots, x^{(n)}\} \middle| \delta^{(i)} = \delta^*\right\}$$

Zaki M and Meira JR W (2014) *Data Mining and Analysis*

Support Vector Classification



Support Vector Classification

Discriminant functions characterize hyperplanes up to scalar multiplication

$$f(x) = 0 \Leftrightarrow a(w^T x + w_0) = a \cdot 0 \Leftrightarrow a w^T x + a w_0 = 0 \Leftrightarrow a f(x) = 0$$

$\Rightarrow$ Joint variation of the length of weight vector and bias yields the same hyperplane

Canonical discriminant function

Fix length of w and w_0 such that $y^* f(x^*) = y^*(w^T x^* + w_0) = 1 \; (x^* \in X^*)$

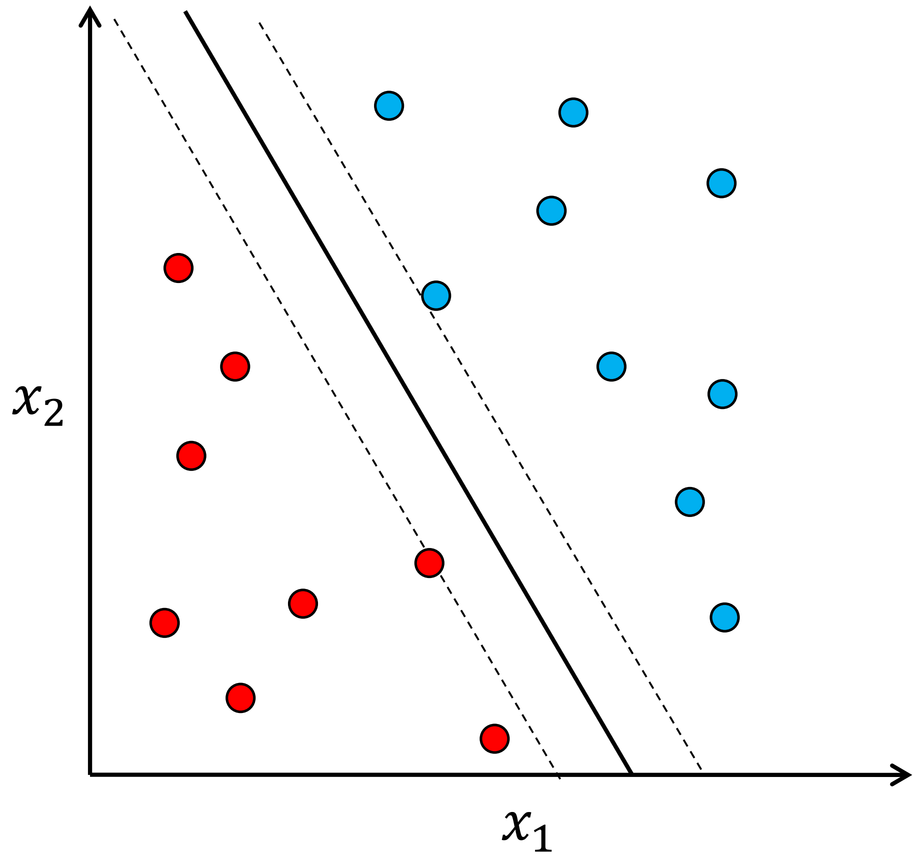
$\Rightarrow$ For support vectors: $\delta^* = \frac{y^*(w^T x^* + w_0)}{\|w\|_2} = \frac{1}{\|w\|_2}$

$\Rightarrow$ For non-support vectors: $\delta^{(i)} > \frac{1}{\|w\|_2} \Rightarrow y^{(i)}\left(w^T x^{(i)} + w_0\right) > 1$

Zaki M and Meira JR W (2014) *Data Mining and Analysis*

Support Vector Classification

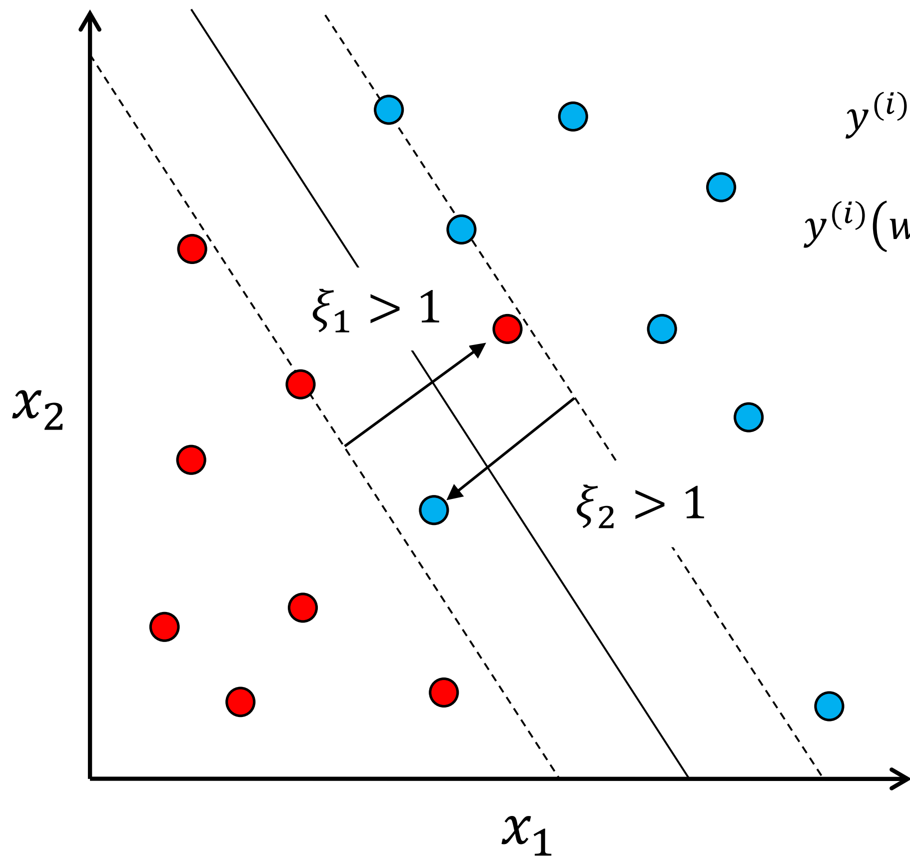
Maximum margin classification for linearly separable training sets



$$w^* = \arg\max_w \frac{1}{\|w\|_2} \Leftrightarrow w^* = \arg\min_w \frac{1}{2} w^T w \text{ subject to } y^{(i)}\left(w^T x^{(i)} + w_0\right) \geq 1$$

Support Vector Classification

Soft margin classification for not linearly separable training sets



$$y^{(i)}\left(w^T x^{(i)} + w_0\right) \geq 1$$
$$\downarrow$$
$$y^{(i)}\left(w^T x^{(i)} + w_0\right) \geq 1 - \xi_i$$

$$(w^*, \xi^*) = \arg\min_{w,\xi} \left(\frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i\right) \text{ subject to } y^{(i)}\left(w^T x^{(i)} + w_0\right) \geq 1 - \xi_i, \xi_i \geq 0$$

Support vector classification and constrained nonlinear optimization

$$(w^*, \xi^*) = \arg\min_{w,\xi} \left(\frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i\right) \text{ subject to } y^{(i)}\left(w^T x^{(i)} + w_0\right) \geq 1 - \xi_i, \xi_i \geq 0$$

$$\Leftrightarrow \min_{x \in \mathbb{R}^d} f(x) \text{ subject to } c_i(x) = 0 \; (i \in E) \text{ and } c_i(x) \geq 0 \; (i \in I)$$

Variable

$$x \coloneqq (w, w_0, \xi) \in \mathbb{R}^{2m+1}$$

Objective function

$$f: \mathbb{R}^{2m+1} \to \mathbb{R}, (w, w_0, \xi) \mapsto f(w, w_0, \xi) \coloneqq \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$$

Inequality constraint functions

$$c_i : \mathbb{R}^{2m+1} \to \mathbb{R}, (w, w_0, \xi) \mapsto c_i(w, w_0, \xi) \coloneqq y^{(i)}\left(w^T x^{(i)} + w_0\right) + \xi_i - 1 \; (i = 1, \ldots, n)$$

$$c_i : \mathbb{R}^{2m+1} \to \mathbb{R}, (w, w_0, \xi) \mapsto c_i(w, w_0, \xi) \coloneqq \xi_{i-n} \; (i = n+1, \ldots, 2n)$$

Many algorithms for solving constrained nonlinear optimization problems exist!

Outline

Constrained nonlinear optimization

Support Vector Classification

Discussion

Discussion

Support vector classification parameter estimation

→ A constrained nonlinear optimization problem

Neuroscientific value of the model?

Scientific value of the model‘s inference scheme?