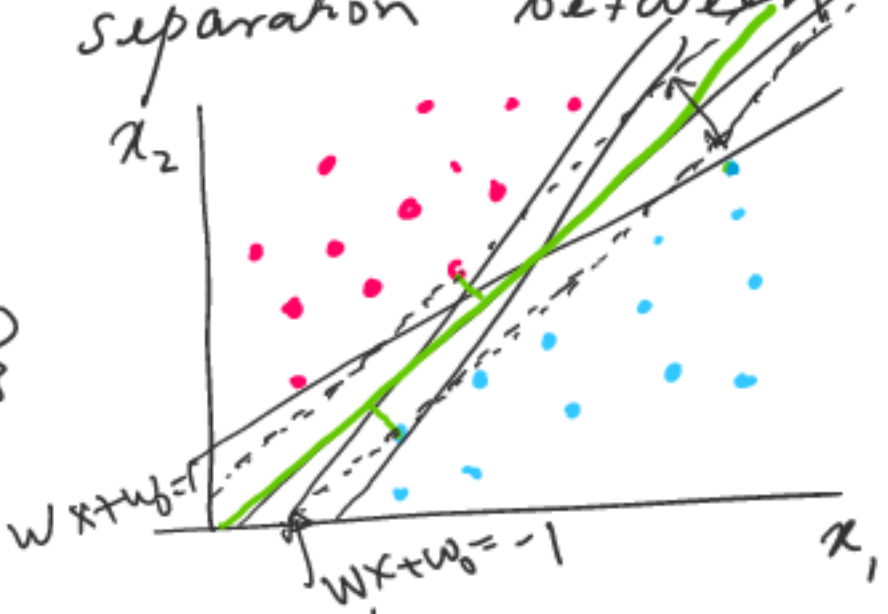
Support Vector Machines

26 October 2018 11:46

Classifiers that attempt to separate instances of two classes by hyperplanes in the original linear or a special high-dimensional space while maximizing the margin of separation between the classes.



$y = \{+1, -1\}$

$D = \{(x^i, y^i) : i = 1 \cdots N\}$

Hyperplanes in linear space:

$wx+w_0 \equiv w_1 x_1 + w_2 x_2 + \cdots w_d x_d + w_0 = 0 \quad d = 2, \; k = 2$

1) Separable case

... boundary ... points

w s.t. $\forall$ positive points

$$w^T x^i + w_0 \geq 1 \quad \forall i : y_i = +1$$

$$w^T x^i + w_0 \leq -1 \quad \forall i : y_i = -1$$

$\Big\}\ i = 1 \cdots N$

What is the geometric distance between

$w^T x^i + w_0 - 1 = 0$ & $w^T x^i + w_0 + 1 = 0$ is $\frac{2}{\|w\|}$

$$\max_{w, w_0} \frac{2}{\|w\|} \quad \text{s.t}$$

$$y_i (w^T x^i + w_0) \geq 1 \quad \forall i$$

$$= \min_{w, w_0} \frac{1}{2} \|w\|^2$$

$$\text{s.t}$$

$$y_i (w^T x^i + w_0) \geq 1$$

Case II Non-separable data

$$\min_{w, \xi_i, w_0} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i$$

s.t

ξ_i ← slack variable

$$\begin{cases} y_i(w^T x^i + w_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

C is a hyper-parameter

$\xi_i \equiv$ Hinge loss with

$$f(x) = wx^i + w_0$$

$$\text{Hinge loss}(yf(x)) = \max(1 - yf(x),\ 0)$$

Proof:

$$\xi_i \geq 0 \quad \& \quad \xi_i \geq 1 - y_i(w^T x^i + w_0)$$

$$\xi_i \geq \max(0,\ 1 - y_i f(x^i))$$

$$\xi_i \geq \text{Hinge loss}(y^i f(x^i))$$

since the objective minimizes w.r.t ξ_i, ξ_i appears as a standalone positive linear term in objective.

as a standard [illegible]

the optimal $\xi_i = \text{Hingeloss}(y^i f(x^i))$

Linear SVM for the non-separable case

$\equiv$ C. Hinge loss + square regularizer.

Non-linear SVMs

$\phi \rightarrow (x_1, x_2, x_1^2, x_2^2, x_1 x_2)$ then they can be easily separated linearly in 5-D space.

Instead explicitly representing points in a transformed space, in SVMs you work with [illegible] original space.

transformed space ... original space.
kernels to stay in the original space.

Kernel: $K(x, x') = \varphi(x) \cdot \varphi(x')$

$\varphi(x) \equiv$ embedding vector of x

Example kernel function: $K(x, x') \equiv (x \cdot x' + 1)^2$

To show: when $\varphi(x) = (x_1, x_2, x_1^2, x_2^2, x_1 x_2)$, then $(x \cdot x' + 1)^2 = \varphi(x) \cdot \varphi(x')$

$x \equiv (x_1, x_2)$

$x' \equiv (x_1', x_2')$

$\varphi(x) \cdot \varphi(x') \equiv (x_1, x_2, x_1^2, x_2^2, x_1 x_2) \cdot (x_1', x_2', x_1'^2, x_2'^2, x_1' \cdot x_2')$

$= x_1 x_1' + x_2 x_2' + x_1^2 x_1'^2 + x_2^2 x_2'^2 + x_1 x_2 x_1' x_2'$

$K(x, x') \equiv (x \cdot x' + 1)^2$

$\equiv ((x_1, x_2) \cdot (x_1', x_2') + 1)^2$

$\equiv (x \cdot x' + x_2 x_2' + 1)^2$

the same set of terms.

$$= (x_1 x_1' + x_2 x_2' + 1) \quad //$$

$$= x_1^2 x_1'^2 + x_2^2 x_2'^2 + 1 + 2 x_1 x_1' x_2 x_2' + 2 x_1 x_1' + 2 x_2 x_2'$$

Types of kernels

$$K(x, x') = x \cdot x' \leftarrow \text{linear kernel}$$

$$= (x \cdot x' + 1)^d \leftarrow \text{polynomial of degree } d \text{ kernel}$$

$$= e^{-\sigma (x - x')^2} \leftarrow \text{Gaussian kernel or the RBF kerned.}$$

the SVM classifier is defined in terms of the kernel function as

$$\sum_{i=1}^{N} \quad \ldots$$

$$f(x) = \sum_{i=1}^{N} \alpha_i k(x^i, x) y_i + w_0$$

Say $k(x^i, x) = (x^i \cdot x)$

similarity among points

$$f(x) = \sum_{i=1}^{N} \alpha_i (x^i \cdot x) y_i + w_0$$

$$= \left(\sum_{i=1}^{N} \alpha_i x^i y_i \right) \cdot x + w_0$$

$$w^T x + w_0$$

Training SVMs

31 October 2018 11:37

Linear SVMs Constrained QP $\Leftrightarrow$ Hinge-loss formulation

$$\min_{w, w_0, \xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i$$

s.t

$$y_i(w^T x^i + w_0) \geq 1 - \xi_i \quad \forall i = 1 \ldots N$$

$$\xi_i \geq 0$$

$$\min_{w, w_0} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} \max(0, 1 - y_i(w^T x^i + w_0))$$

$F(w, w_0)$

Two methods

1) Hinge-loss formulation

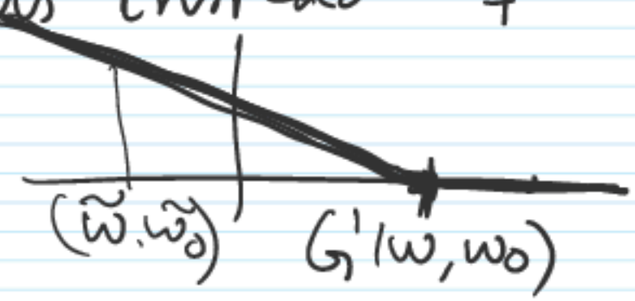
Apply stochastic gradient descent type algorithms

But Hinge loss is not differentiable in w.

So, we will work with sub-gradients instead of gradient.

$$\min_{w, w_0} G(w, w_0) \equiv \max(0, G'(w, w_0))$$

non-differentiable, differentiable



sub-gradient at $(\tilde{w}, \tilde{w}_0)$ — 0 if $G'(\tilde{w}, \tilde{w}_0) \leq 0$

sub gradient at at $(\tilde{w}, \tilde{w}_0) \equiv 0$ if $G(w, w_0) \leq 0$

$\equiv \nabla_{w, w_0} G'(\tilde{w}, \tilde{w}_0)$ otherwise

Run gradient descent algorithm with sub-gradients instead of gradient.

II) Primal-dual method

Starting with the constained QP formulation

Primal (P)

$\min_w F(w)$

s.t. $g_k(w) \leq 0 \; ; k = 1 \cdots K$

$g_k(w)$ is convex

$F(w)$ is convex

Lagrangian dual of (P)

$\theta(\alpha_1, \alpha_2 \cdots \alpha_K)$

$= \min_w F(w) + \sum_{k=1}^{K} \alpha_k g_k(w)$

s.t. $\alpha_k \geq 0 \quad \forall k = 1 \cdots K$

Thm (Weak duality)

Thm (Weak duality)

Let $\hat{w}$ be a feasible solution of the primal $\hat{P}$

Then, $\forall \alpha_k \geq 0$, the dual $\theta(\alpha_1 \ldots \alpha_k) \equiv \theta(\alpha) \leq F(\hat{w})$

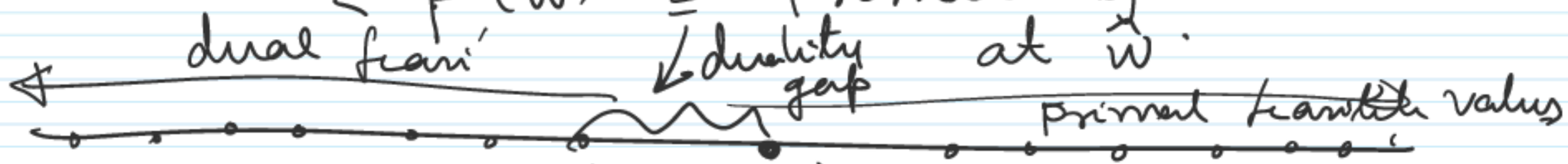
$$\theta(\alpha) = \min_{w} F(w) + \sum_k \alpha_k g_k(w)$$

$$\leq F(\hat{w}) + \sum_k \alpha_k g_k(\hat{w})$$

$\hat{w}$ is feasible $g_k(\hat{w}) \leq 0$

$\alpha_k \geq 0$

$$\leq F(\hat{w}) \equiv \text{Primal objective value}$$



$\forall \alpha \quad \theta(\alpha) \leq F(\hat{w}) \quad \forall$ feasible $\hat{w}$

s.t $g_k(\hat{w}) \leq 0$ primal.

$g_k(\hat{w}) \le 0$ primal

$$\max_{\alpha_1 \dots \alpha_k \ge 0} \theta(\alpha) \equiv \max_{\alpha_1 \dots \alpha_k} \min_{W} F(W) + \sum_{k=1}^{K} \alpha_k g_k(W)$$

$$\alpha_k \ge 0 \quad \forall k = 1 \dots K$$

Lagrangian Dual

Thm (Strong duality)

If $F(W)$, $g_k(W)$ are convex functions, $\exists\, W$ s.t $g_k(W) < 0 \ \forall k$, then there exists (α^*, W^*) s.t.

$F(W^*)$ is primal optimal and W^* is primal feasible and $\theta(\alpha^*)$ is dual feasible and optimum and $\theta(\alpha^*) = F(W^*)$

$\Rightarrow$ duality gap is 0.

Primal

$$\min_{W} F(W) \quad \text{s.t}$$

Dual

$$\max_{\alpha_1 \dots \alpha_k} \min_{W} F(W) + \sum_{k=1}^{K} \alpha_k g_k(W)$$

$$\alpha_k \ge 0 \quad \forall k = 1 \dots K.$$

$\min_{w} f(w)$ s.t
$g_k(w) \le 0 \quad \forall k = 1 \cdots K$

$\Leftrightarrow$ $\max_{\alpha_k \ge 0} \min_{w} L(w, \alpha_k)$ $\quad \alpha_k \ge 0 \quad \forall k = 1 \cdots K$.

when $f(w)$ & $g_k(w)$ are convex.

Applying to SVMs

$$\min_{w, w_0, \xi_i} \; \underbrace{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i}_{F(w, w_0, \xi_i)}$$

$$\text{s.t} \quad \begin{cases} 1 - \xi_i - y_i(w^T x^i + w_0) \le 0 & \forall i = 1 \cdots N \quad (\alpha_i) \\ -\xi_i \le 0 & (\mu_i) \end{cases}$$

$$\Leftrightarrow \max_{\substack{\alpha_1 \cdots \alpha_N \\ \mu_1 \cdots \mu_N}} \min_{w, w_0, \xi_i} \underbrace{\left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \alpha_i \left(1 - \xi_i - y_i(w^T x^i + w_0)\right) + \sum_{i=1}^{N} \mu_i(-\xi_i) \right\}}_{L(w, w_0, \xi_i, \alpha_i, \mu_i)}$$

$$\text{s.t} \quad \alpha_i \ge 0 \; ; \; \mu_i \ge 0 \quad \forall i = 1 \cdots N$$

Solving

$$\max_{\alpha_i, \mu_i} \min_{w, w_0, \xi_i} L(w, w_0, \xi_i, \alpha_i, \mu_i)$$

convex in w, w_0, ξ_i

Step 1

$$\nabla_w L \equiv w^* + \sum_{i=1}^{N} \alpha_i^* (-y_i x^i) = 0 \Rightarrow \boxed{w^* = \sum_{i=1}^{N} \alpha_i^* y_i x^i}$$

$$\nabla_{w_0} L \equiv -\sum_{i=1}^{N} \alpha_i^* y_i = 0 \Rightarrow \boxed{\sum_{i=1}^{N} \alpha_i^* y_i = 0}$$

$$= \sum_{i=1}^{N} \alpha_i^* = \sum_{\substack{i=1 \\ y_i=+1}}^{N} \alpha_i^* = \sum_{\substack{i=1 \\ y_i=-1}}^{N} \alpha_i^*$$

$$\nabla_{\xi_i} L = C - \alpha_i^* - \mu_i^* = 0$$

$$\Rightarrow \boxed{\alpha_i^* + \mu_i^* = C}$$

Step-2

substitute $w = \sum_{i=1}^{N} \alpha_i y_i x^i$ in $L(w, w_0, \xi_i, \alpha_i, \mu_i)$ and simplify and that will give us finally

$$\min_{w, w_0, \xi_i} L(w, w_0, \xi_i, \alpha_i, \mu_i) \quad \Rightarrow -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \, x^i \cdot x^j + \sum_{i=1}^{N} \alpha_i$$

$$\text{s.t } \alpha_i \geq 0, \ \mu_i \geq 0$$

$$\text{s.t} \sum_{i=1}^{N} \alpha_i y_i = 0 \ ; \ \alpha_i + \mu_i = C \ \forall i$$

$$\alpha_i \geq 0 \ ; \ \mu_i \geq 0$$

Step-3

Eliminate μ_i

$$\max_{\alpha_1 \cdots \alpha_N} -\frac{1}{2} \sum_{i=} \sum_{j} \alpha_i \alpha_j \, y_i y_j \, x^i \cdot x^j + \sum_{i=1}^{N} \alpha_i$$

$$\text{s.t} \sum_{i=1}^{N} \alpha_i y_i = 0 \ , \quad \alpha_i \geq 0 \ ; \ \alpha_i \leq C$$

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

Last class Primal ✓

$$\min_{w, w_0, \xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i$$

s.t $y_i(wx^i + w_0) \geq 1 - \xi_i$ $\forall i = 1 \ldots N$

Dual ✓

$$\max_{\alpha_1 \ldots \alpha_N} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boxed{x^i \cdot x^j} + \sum_{i=1}^{N} \alpha_i$$

s.t

$0 \leq \alpha_i \leq C$ ← box constraint

$\sum_{i=1}^{N} \alpha_i y_i = 0$

Extending the dual to arbitrary kernels using "kernel trick"

1) In the dual, x^i & x^j only appear as $x^i \cdot x^j$

2) For a non-linear embedding $\varphi(x)$, we can use $\varphi_K(x^i) \cdot \varphi_K(x^j) = K(x^i, x^j)$

Now the kernelized dual objective becomes:

$$\max_{\alpha_1 \cdots \alpha_N} -\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j k(x^i, x^j) + \sum_i \alpha_i \quad \leftarrow \theta(\alpha)$$

s.t. $0 \le \alpha_i \le C$; $\sum_i \alpha_i y_i = 0$

$$w = \sum_{i=1}^{N} \alpha_i y_i \underline{x^i}$$

Your classifier: $w^T x + w_0$

$$\left(\sum_{i=1}^{N} \alpha_i y_i x^i\right) \cdot (x) + w_0$$

$$= \sum_{i=1}^{N} \alpha_i y_i k(x^i, x) + w_0$$

$\theta(\alpha) \equiv \theta(\alpha_1 \cdots \alpha_N)$ is concave in $\alpha_1 \cdots \alpha_N$

1) $\max_{\alpha_1 \cdots \alpha_N} \theta(\alpha_1 \cdots \alpha_N)$ can be solved using gradient ascent subject to the simple constants

Cannot use direct gradient ascent since the objective is quadratic in the number of training examples.

in one [illegible]

2) Single co-ordinate ascent will not work because of the constraint $\sum \alpha_i y_i = 0$

2) SMO Algorithm — change two α-s in each step.

Initially: $t = 0$

→ $\alpha_i^t = 0 \quad \forall \; i = 1 \ldots N$

while (change) {

→ (α_i^t, α_j^t = pick two α-s to change) ← based on heuristics

choose a change amount: λ^t

→ $\alpha_i^{t+1} \leftarrow \alpha_i^t + \lambda^t$

$\alpha_j^{t+1} \leftarrow \alpha_j^t \pm \lambda^t$ ← depend on whether $y_i = y_j$ or not.

Pick λ so as to maximize

$\lambda^* = \arg\max_{\lambda} \Theta(\alpha_1^t \ldots \alpha_i^t + \lambda, \ldots \alpha_j^t \mp \lambda, \ldots \alpha_N^t)$ ← can be solved in closed form by computing gradient

λ^t so as to be closest

$\underline{\lambda}^t \leftarrow$ so as to be closest to λ^* which satisfying the constraints

closest equating λ to of θ w.r.t be 0.

$$\left| \begin{array}{l} 0 \le \alpha_i^t + \lambda^t \le C \\ \& \\ 0 \le \alpha_j^t \pm \lambda^t \le C \end{array} \right.$$

$t = t+1$

Example

$t=0 \; \alpha_1 \ldots \alpha_N \equiv 0$; $\alpha_1, \alpha_2 \equiv \alpha_i, \alpha_j$; $\boxed{y_1 \neq y_2}$

$$\theta(\lambda) = -\frac{2}{2}(\lambda)(\lambda) k(x_1,x_2) y_1 y_2 + 2\lambda - \tfrac{1}{2}\lambda^2 k(x_1,x_1) y_1^2 - \tfrac{1}{2}\lambda^2 k(x_2,x_2) y_2^2$$

$$\frac{\partial \theta}{\partial \lambda} \equiv -2\lambda k(x_1,x_2)(-1) - \lambda k(x_1,x_1) - \lambda k(x_2,x_2) + 2$$

$$\Rightarrow \lambda^* = \frac{-2}{2k(x_1,x_2) - k(x_1,x_1) - k(x_2,x_2)}$$

Say $\rightarrow k(x_1,x_2) = -1$; $C = 1/3$.

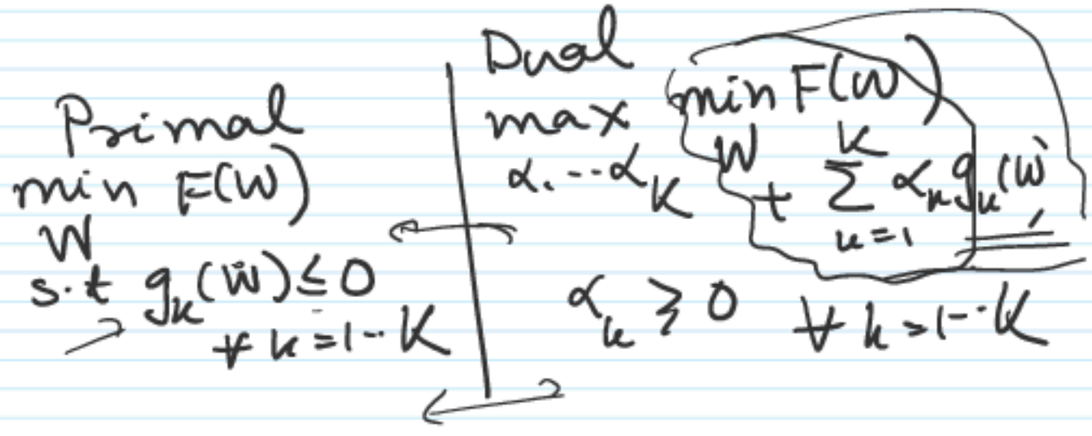
$k(x_1,x_1) = k(x_2,x_2) = 1$

$$\Rightarrow \lambda^* = \frac{-2}{-2-1-1} = \frac{1}{2}$$

$\lambda^t \equiv 1/2$ to ensure than $\alpha_1 + \lambda^t \le C$

$\lambda \equiv 1/3$ to ensure than α_1 [illegible]

$\Rightarrow \alpha_1 = \frac{1}{3}, \quad \alpha_2 = \frac{1}{3}$



Thm (KKT conditions)

When $F(w)$, $g_k(w)$ are convex then $a\ (\alpha^*, w^*)$ are optimal solutions for the dual & primal respectively if and only if

1) $\nabla_w F(w^*) + \sum_{k=1}^{K} \alpha_k^* \nabla_w g_k(w^*) = 0$

2) $\alpha_k g_k(w) = 0 \quad \forall k = 1 \cdots K$ — $\sum_k \alpha_k^* g_k(w^*) = 0$

3) $\alpha_k^* \ge 0$

4) $g_k(w^*) \le 0$

Applying KKT conditions to SVMs

$w^* = \sum_{i=1}^{N} \alpha_i^* y_i x^i$) 1st KKT condition

$$\sum_{i=1}^{N} \alpha_i^* y_i = 0$$

$$\alpha_i^* + \mu_i^* = C$$

} KKT condition

$$\mu_i^* \geq 0, \quad \alpha_i^* \geq 0, \quad \xi_i^* \geq 0$$

} primal dual feasibility

$$\alpha_i^* \left(1 - y_i(w^{*T} x^i + w_0^*) - \xi_i^*\right) = 0 \quad \forall i$$

$$\mu_i^* \xi_i^* = 0$$

} 2nd KKT condition.

Case 1: consider i : $y_i(w^{*T} x^i + w_0) > 1$, $\xi_i^* = 0$

$\Rightarrow \alpha_i^* = 0$ from

case 2: consider i: $\xi_i^* > 0 \Rightarrow \mu_i^* = 0$ cause of

$\Rightarrow \alpha_i^* = C$

case 3: i s.t

$0 < \alpha_i^* < C$

$$0 < \alpha_i^* < C$$

$$\Rightarrow \mu_i^* > 0 \; ; \; \xi_i^* = 0$$

$$\Rightarrow y_i(w^* x^i + w_0) = 1$$

$$w^* \equiv \sum_{i=1}^{N} \alpha_i^* y_i x^i$$

$$\equiv \sum_{i \in \text{support vectors}} \alpha_i^* y_i x^i$$

↓ those examples for which $\alpha_i > 0$

$$f(x) = w^* x + w_0$$

$$\equiv \sum_{i \in \text{support vectors}} \alpha_i^* y_i K(x^i, x) + w_0$$

How to compute w_0

Find an example i for which α_i is $< C$ & > 0.

$y_i(w^* x^i + w_0) = 1$ → solve this to find w_0

$y_i(w \cdot x + w_0)$ — find w_0

Extending SVMs to multi-class classification.

Suppose we have K classes.

Like in logistic regression classifiers we allocate parameters w^y, w_0^y for each class y.

$$w^y \in \mathbb{R}^d, \quad w_0^y \in \mathbb{R}.$$

score for an example being in class y as

$$s(x, y) \equiv w^y x + w_0^y$$

$$\min_{\{w^y, w_0^y\}, \xi_i} \; -\frac{1}{2}\sum_{y=1}^{K} \|w^y\| + C \sum_{i=1}^{N} \xi_i$$

s.t

$$(x^i, y_i) \quad w^{y_i} x^i + w_0^{y_i} \geq w^y x^i + w_0^y + 1 - \xi_i \quad \forall i = 1 \cdots N \;\; \forall y \neq y_i$$

$$\xi_i \geq 0$$

Hinge loss

Hinge loss

$$\max_{y \neq y_i} \left(0, 1 - [s(x^i, y_i) - s(x^i, y)]\right)$$