SVM – the original papers

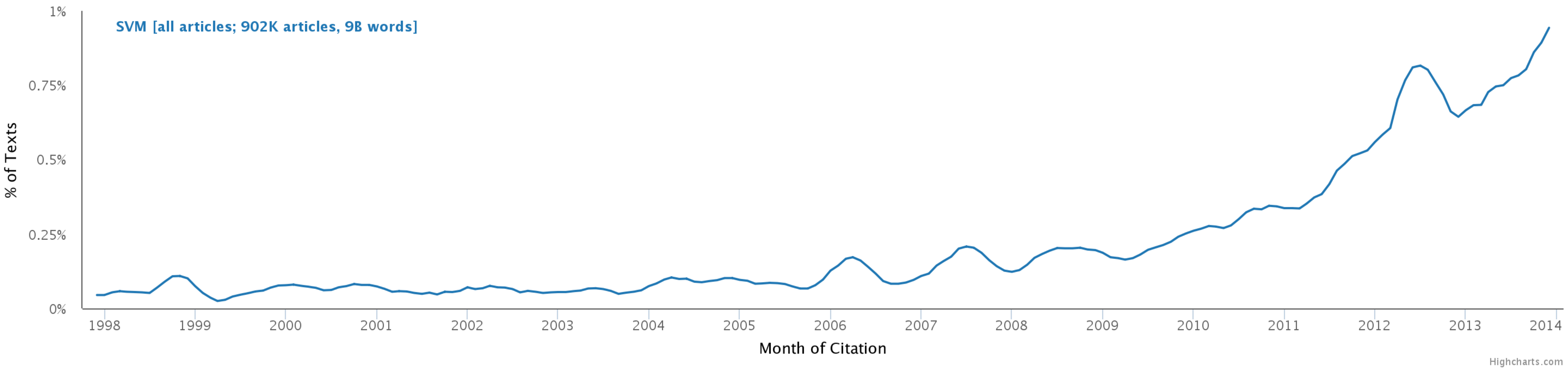
Adrian Florea

Papers We Love - Bucharest Chapter

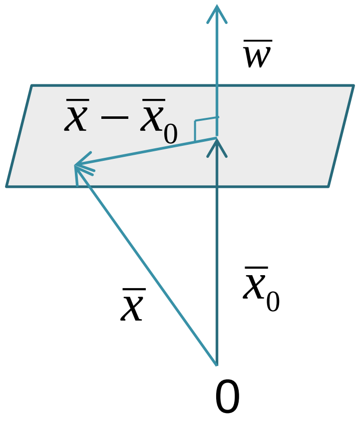
November 3rd, 2016

SVM-related papers trend on arXiv

data from Nov 1997 to Nov 2013
http://bookworm.culturomics.org/arxiv/



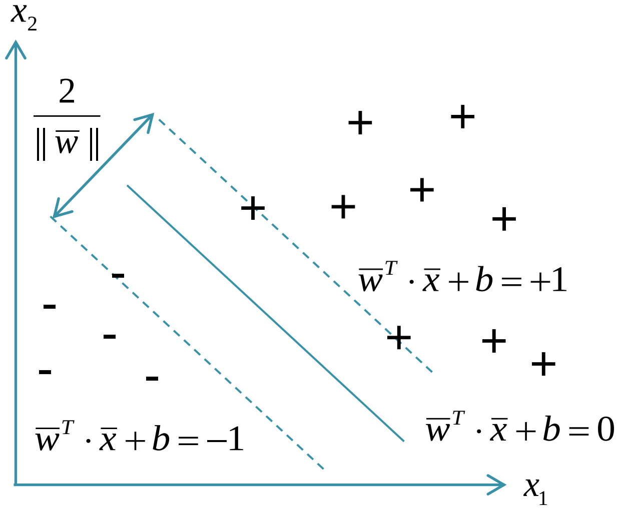
Introduction



training set: $\left\{(\overline{x}_i, y_i) \mid i=\overline{1,p}, \overline{x}_i \in R^N, y_i = \begin{cases} +1, \overline{x}_i \in A \\ -1, \overline{x}_i \in B \end{cases}\right\}$

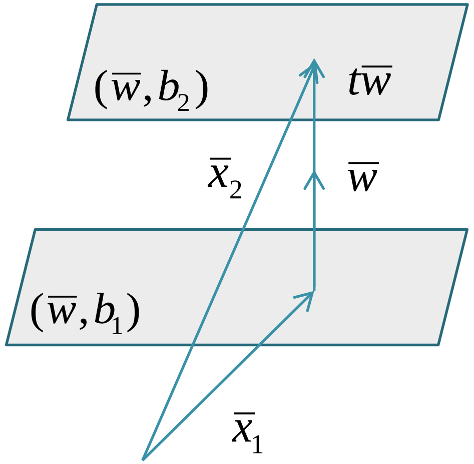
$$d((\overline{w}, b+1),(\overline{w}, b-1)) = \frac{|b+1-b+1|}{\|\overline{w}\|} = \frac{2}{\|\overline{w}\|}$$

$$\begin{cases} \overline{w}\cdot\overline{x}_i + b \geq +1, y_i = +1 \\ \overline{w}\cdot\overline{x}_i + b \leq -1, y_i = -1 \end{cases} \Leftrightarrow y_i(\overline{w}\cdot\overline{x}_i + b) \geq 1$$



$$\overline{w} \perp \overline{x}-\overline{x}_0 \Rightarrow \overline{w}\cdot(\overline{x}-\overline{x}_0) = 0 \Rightarrow \overline{w}\cdot\overline{x} + b = 0$$

$$\max_{\overline{w},b} \frac{1}{\|\overline{w}\|}, \text{s.t. } y_i(\overline{w}\cdot\overline{x}_i + b) \geq 1$$



$$d((\overline{w}, b_1),(\overline{w}, b_2)) = \|t\overline{w}\| = |t|\,\|\overline{w}\|$$

$$\overline{w}\cdot\overline{x}_2 + b_2 = 0 \Rightarrow \overline{w}\cdot(\overline{x}_1 + t\overline{w}) + b_2 = 0 \Rightarrow \overline{w}\cdot\overline{x}_1 + t\|\overline{w}\|^2 + b_2 = 0$$

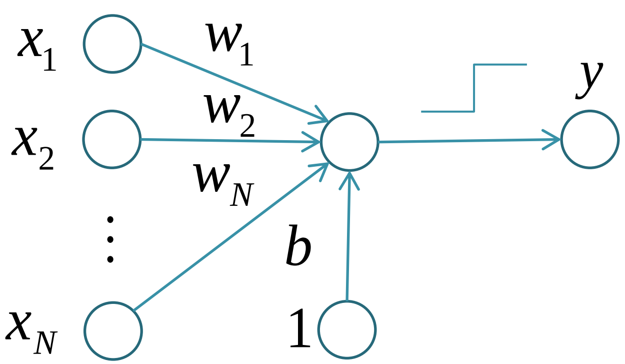
$$(\overline{w}\cdot\overline{x}_1 + b_1) - b_1 + t\|\overline{w}\|^2 + b_2 = 0 \Rightarrow t = \frac{b_1 - b_2}{\|\overline{w}\|^2}$$

$$d((\overline{w}, b_1),(\overline{w}, b_2)) = |t|\,\|\overline{w}\| = \frac{|b_1 - b_2|}{\|\overline{w}\|}$$

$$\max \frac{1}{\|\overline{w}\|} \Rightarrow \min \|\overline{w}\| => \boxed{\min \frac{\|\overline{w}\|^2}{2}}$$

$$y_i(\overline{w}\cdot\overline{x}_i + b) \geq 1 \Rightarrow \boxed{y_i = \text{sign}(\overline{w}\cdot\overline{x}_i + b)}$$

Perceptron



decision function

$$\hat{f}(\bar{x}) = \mathrm{sign}(\bar{w} \cdot \bar{x} + b)$$

$\bar{x} = (x_1, x_2, \ldots, x_N)$ features of the customer $\bar{x}$

approve credit if $\sum_{i=1}^{N} w_i x_i > \text{threshold}$

deny credit if $\sum_{i=1}^{N} w_i x_i < \text{threshold}$

Frank Rosenblatt

input: $\mathrm{T} = \{(\bar{x}_i, y_i) \mid i = \overline{1, p}\} \subset \mathbf{R}^N \times \{+1, -1\}$

$\bar{w} \leftarrow \bar{0}, b \leftarrow 0$

repeat

 for i=1 **to** p

  if $\mathrm{sign}(\bar{w} \cdot \bar{x}_i + b) \neq y_i$ **then**

   $\bar{w} \leftarrow \bar{w} + y_i \bar{x}_i$

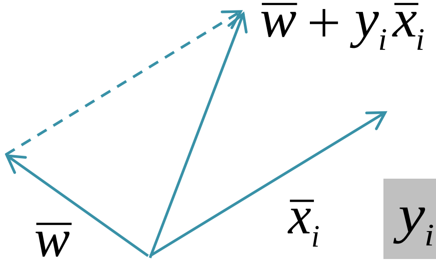
   $b \leftarrow b + y_i$

  end if

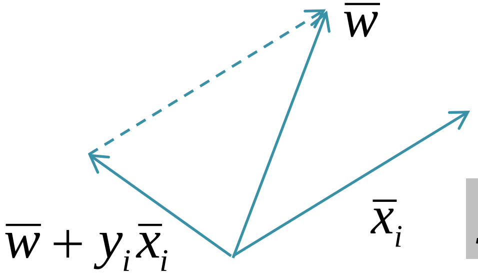
 end for

until $\mathrm{sign}(\bar{w} \cdot \bar{x}_j + b) = y_j, \forall j = \overline{1, p}$

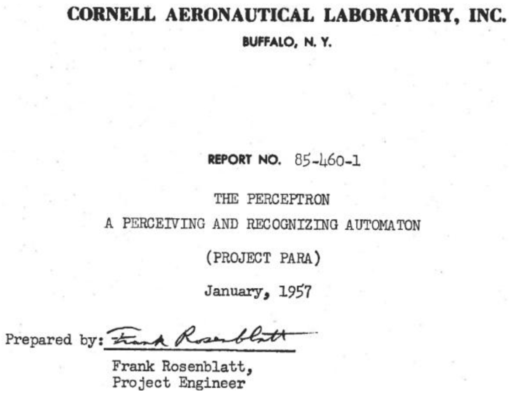
return $\bar{w}, b$



$y_i = +1$



$y_i = -1$



CORNELL AERONAUTICAL LABORATORY, INC.
BUFFALO, N. Y.

REPORT NO. 85-460-1

THE PERCEPTRON
A PERCEIVING AND RECOGNIZING AUTOMATON
(PROJECT PARA)

January, 1957

Prepared by:
Frank Rosenblatt,
Project Engineer

Novikoff's Theorem of Convergence

Assume $\exists \overline{w}^*, \| \overline{w}^* \| = 1, \rho > 0$ s.t. $y_i \overline{w}^* \cdot \overline{x}_i \geq \rho, \forall i = \overline{1, p}$

$$R \overset{\text{def}}{=} \max_{i=1,p} \| \overline{x}_i \|$$

Then the perceptron algorithm makes at most $\dfrac{R^2}{\rho^2}$ errors.

A.B.J. Novikoff, "On convergence proofs on perceptrons", Proc. of the *Symposium on the Mathematical Theory of Automata*, **12**, pp 615–622 (1962)

Assume k^{th} error $\overline{w}^{(k)}$ is made on $\overline{x}_t$ and $\overline{w}^{(1)} = \overline{0}$.

$$\overline{w}^{(k+1)} \cdot \overline{w}^* = (\overline{w}^{(k)} + y_t \overline{x}_t) \cdot \overline{w}^* = \overline{w}^{(k)} \cdot \overline{w}^* + y_t \overline{w}^* \cdot \overline{x}_t \geq \overline{w}^{(k)} \cdot \overline{w}^* + \rho$$

By induction on k : $\overline{w}^{(k+1)} \cdot \overline{w}^* \geq k\rho$

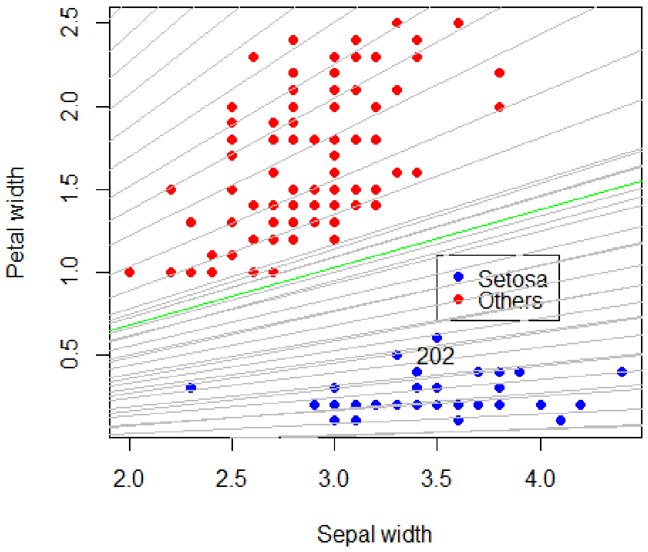
Cauchy – Schwartz : $\| \overline{w}^{(k+1)} \| \cdot \| \overline{w}^* \| \geq \overline{w}^{(k+1)} \cdot \overline{w}^*$

$$\| \overline{w}^* \| = 1 \Rightarrow \| \overline{w}^{(k+1)} \| \geq k\rho$$

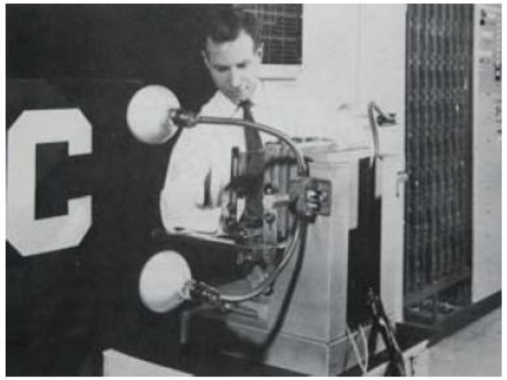
$$\| \overline{w}^{(k+1)} \|^2 = \| \overline{w}^{(k)} + y_t \overline{x}_t \|^2 = \| \overline{w}^{(k)} \|^2 + \underbrace{y_t^2}_{=1} \underbrace{\| \overline{x}_t \|^2}_{\leq R^2} + \underbrace{2 y_t \overline{w}^{(k)} \cdot \overline{x}_t}_{\leq 0\,(k^{\text{th}}\text{ error})}$$

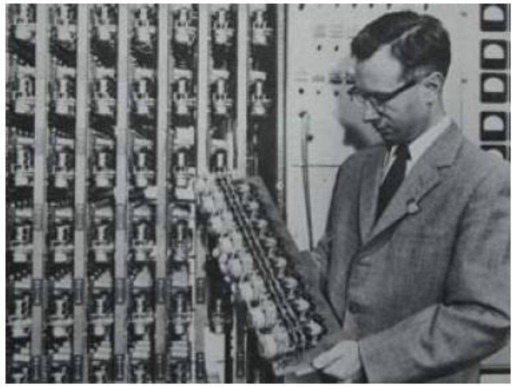
By induction on k : $\| \overline{w}^{(k+1)} \|^2 \leq kR^2 \Rightarrow k^2\rho^2 \leq \| \overline{w}^{(k+1)} \|^2 \leq kR^2 \Rightarrow k \leq \dfrac{R^2}{\rho^2}$

Demo



Mark I Perceptron hardware





N = 400
p = 512

input: $\mathrm{T} = \{(\overline{x}_i, y_i) \mid i = \overline{1, p}\} \subset \mathbf{R}^N \times \{+1, -1\}$

$w \leftarrow 0, b \leftarrow 0$

repeat

  for i=1 **to** p

    if $\mathrm{sign}(\overline{w} \cdot \overline{x}_i + b) \neq y_i$ **then**

      $\overline{w} \leftarrow \overline{w} + y_i \overline{x}_i$

      $b \leftarrow b + y_i$

    end if

  end for

until $\mathrm{sign}(\overline{w} \cdot \overline{x}_j + b) = y_j, \forall j = \overline{1, p}$

return $\overline{w}, b$

- inputs captured by a 20 × 20 array of cadmium sulphide photocells
- a patch board allowed different configurations of input features to be tried
- each adaptive weight was implemented using a rotary variable resistor/potentiometer, driven by an electric motor

C. Bishop, "Pattern Recognition and Machine Learning", Springer (2006), p. 196

Linear separability as an LP problem

A necessary and sufficient condition for the **linear separability** of the pattern sets A and B is that:

$$\varphi(A, B) > 0$$

where φ(A, B) is the solution of the linear programming problem:

LINEAR AND NONLINEAR SEPARATION OF PATTERNS BY LINEAR PROGRAMMING

O. L. Mangasarian

Shell Development Company, Emeryville, California

(Received September, 1964)

$$\varphi(A,B) = \min_{\bar{u},\bar{v},\bar{p}} \{ \bar{1}_n^T \bar{p} \mid \bar{1}_m^T \bar{u} = 1, \bar{1}_k^T \bar{v} = 1, -A^T\bar{u} + B^T\bar{v} + \bar{p} \geq \bar{0}, A^T\bar{u} - B^T\bar{v} + \bar{p} \geq \bar{0}, \bar{u} \geq \bar{0}, \bar{v} \geq \bar{0} \}$$

where u, v, and p are m-, k-, and n-dimensional column vectors.

A necessary and sufficient condition that the sets of patterns A and B be **linearly inseparable** is that the system has a solution:

$$\begin{cases} A^T\bar{u} - B^T\bar{v} = \bar{0} \\ \bar{1}^T\bar{u} = 1 \\ \bar{1}^T\bar{v} = 1 \\ \bar{u} \geq \bar{0} \\ \bar{v} \geq \bar{0} \end{cases}$$

card(A)=m, card(B)=k, R^n

(<u>m + k + n + 2</u> + n) x (<u>m + k</u> + n)

Linear separability as an LP problem

$$
\begin{bmatrix}
1 & \dots & 1 & 0 & \dots & 0 & 0 & \dots & 0 \\
0 & \dots & 0 & 1 & \dots & 1 & 0 & \dots & 0 \\
-a_{11} & \dots & -a_{m1} & b_{11} & \dots & b_{k1} & 1 & \dots & 0 \\
\vdots & \dots & \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\
-a_{1n} & \dots & -a_{mn} & b_{1n} & \dots & b_{kn} & 0 & \dots & 1 \\
a_{11} & \dots & a_{m1} & -b_{11} & \dots & -b_{k1} & 1 & \dots & 0 \\
\vdots & \dots & \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\
a_{1n} & \dots & a_{mn} & -b_{1n} & \dots & -b_{kn} & 0 & \dots & 1 \\
1 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\
\vdots & \dots & \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\
0 & \dots & 1 & 0 & \dots & 0 & 0 & \dots & 0 \\
0 & \dots & 0 & 1 & \dots & 0 & 0 & \dots & 0 \\
\vdots & \dots & \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\
0 & \dots & 0 & 0 & \dots & 1 & 0 & \dots & 0
\end{bmatrix}
\begin{Bmatrix} u_1 \\ \vdots \\ u_m \\ v_1 \\ \vdots \\ v_k \\ p_1 \\ \vdots \\ p_n \end{Bmatrix}
\begin{matrix} = \\ = \\ \geq \\ \vdots \\ \geq \\ \geq \\ \vdots \\ \geq \\ \geq \\ \vdots \\ \geq \\ \geq \\ \vdots \\ \geq \end{matrix}
\begin{Bmatrix} 1 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{Bmatrix}
$$

A and B are **linear separable** iff:

$$\min_{\bar{u},\bar{v},\bar{p}} \sum_{i=1}^{n} p_i > 0$$

A and B are **linear inseparable** iff this system has a solution

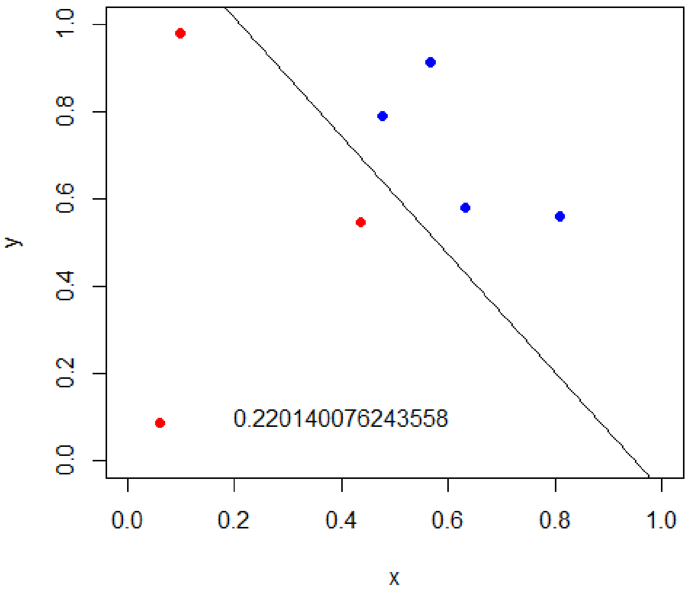
$$
\begin{bmatrix}
a_{11} & \dots & a_{m1} & -b_{11} & \dots & -b_{k1} \\
\dots & \dots & \dots & \dots & \dots & \dots \\
a_{1n} & \dots & a_{mn} & -b_{1n} & \dots & -b_{kn} \\
1 & \dots & 1 & 0 & \dots & 0 \\
0 & \dots & 0 & 1 & \dots & 1 \\
1 & \dots & 0 & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots \\
0 & \dots & 1 & 0 & \dots & 0 \\
0 & \dots & 0 & 1 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots \\
0 & \dots & 0 & 0 & \dots & 1
\end{bmatrix}
\begin{Bmatrix} u_1 \\ \dots \\ u_m \\ v_1 \\ \dots \\ v_k \end{Bmatrix}
\begin{matrix} = \\ \dots \\ = \\ = \\ = \\ \geq \\ \dots \\ \geq \\ \geq \\ \dots \\ \geq \end{matrix}
\begin{Bmatrix} 0 \\ \dots \\ 0 \\ 1 \\ 1 \\ 0 \\ \dots \\ 0 \\ 0 \\ \dots \\ 0 \end{Bmatrix}
$$

Linear separability as an LP problem

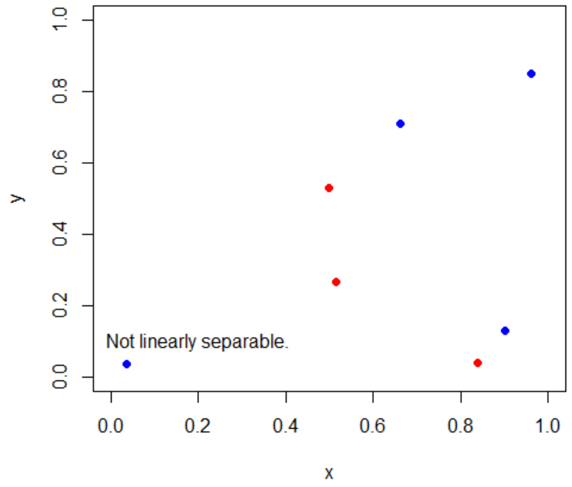
$$\exists \beta \in R, \bar{h} \in R^n, s.t. \forall \bar{a} \in A, \bar{b} \in B:$$

$$\begin{cases} \bar{h}^T \bar{a} > \beta \\ \bar{h}^T \bar{b} < \beta \end{cases} \Rightarrow \begin{cases} \bar{h}^T \bar{a} \ge \beta + \varepsilon \\ \bar{h}^T \bar{b} \le \beta - \varepsilon \end{cases} \Bigg| \left(1/\varepsilon \right) \Rightarrow \begin{cases} -\bar{h}^T \bar{a} + \beta \le -1 \\ \bar{h}^T \bar{b} - \beta \le -1 \end{cases}$$

$$\begin{bmatrix} -a_{11} & \dots & -a_{1n} & 1 \\ \dots & \dots & \dots & \dots \\ -a_{m1} & \dots & -a_{mn} & 1 \\ b_{11} & \dots & b_{1n} & -1 \\ \dots & \dots & \dots & \dots \\ b_{k1} & \dots & b_{kn} & -1 \end{bmatrix} \begin{Bmatrix} h_1 \\ \dots \\ h_n \\ \beta \end{Bmatrix} \le \begin{Bmatrix} -1 \\ \dots \\ -1 \\ -1 \end{Bmatrix}$$



Demo



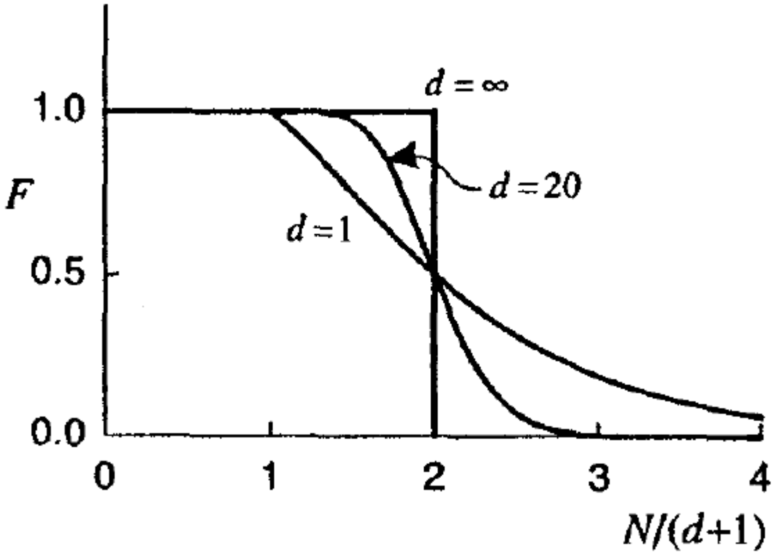
Cover’s Theorem

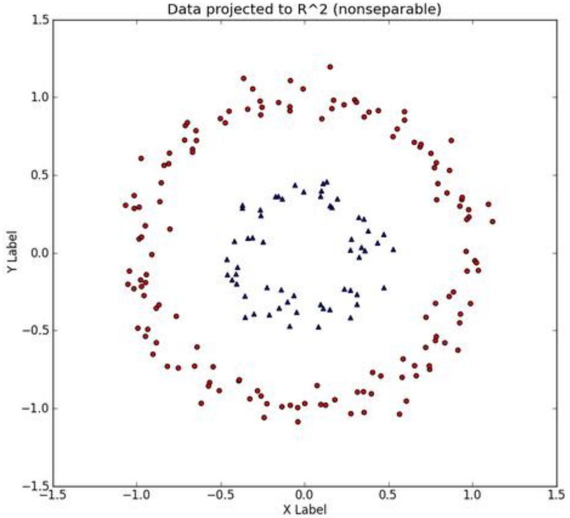
dichotomy: $\{(\overline{x}_i, y_i) \mid i = \overline{1, N}\} \subset R^d \times \{-1,1\}$

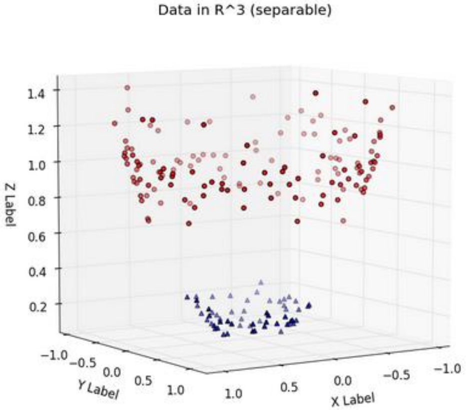
Assume there is no subset of d or fewer points linearly dependent.

Number of dichotomies linearly separable: $F(N,d) = \begin{cases} 1, N \leq d+1 \\ \frac{1}{2^{N-1}} \sum_{i=0}^{d} C_{N-1}^{i}, N \geq d+1 \end{cases}$

Thomas Cover



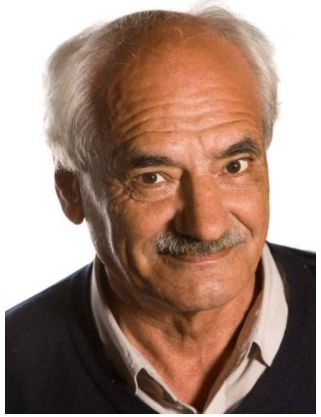




$$F(N, d+1) = F(N, d) + \frac{C_{N-1}^{d+1}}{2^{N-1}}$$

A complex pattern-classification problem, cast in a **high-dimensional space** nonlinearly, is more likely to be linearly separable than in a **low-dimensional space**

V. Vapnik, A. Chervonenkis,
"A note on one class of perceptrons",
Automation and remote control, **25**, 1 (1964)


Alexey Chervonenkis


Vladimir Vapnik


Bernhard Boser


Isabelle Guyon


Vladimir Vapnik

A Training Algorithm for Optimal Margin Classifiers

Bernhard E. Boser*
EECS Department
University of California
Berkeley, CA 94720
boser@eecs.berkeley.edu

Isabelle M. Guyon
AT&T Bell Laboratories
50 Fremont Street, 6th Floor
San Francisco, CA 94105
isabelle@neural.att.com

Vladimir N. Vapnik
AT&T Bell Laboratories
Crawford Corner Road
Holmdel, NJ 07733
vlad@neural.att.com

Machine Learning, 20, 273–297 (1995)

Support-Vector Networks

CORINNA CORTES
VLADIMIR VAPNIK
AT&T Bell Labs., Holmdel, NJ 07733, USA

corinna@neural.att.com
vlad@neural.att.com


Corinna Cortes


Vladimir Vapnik

Lagrangian optimization problem

Primal formulation: $\min\limits_{\overline{x}} \Phi(\overline{x}) \text{ s.t. } g_i(\overline{x}) \geq 0$ $\quad i = \overline{1,p},\ \overline{x} \in \mathrm{R}^N,\ \Phi$ convex, g_i linear

Dual formulation: $\max\limits_{\overline{\alpha}} \min\limits_{\overline{x}} L(\overline{\alpha}, \overline{x}) = \max\limits_{\overline{\alpha}} \min\limits_{\overline{x}} \left(\Phi(\overline{x}) - \sum_{i=1}^{p} \alpha_i g_i(\overline{x}) \right) \text{ s.t. } \overline{\alpha}_i \geq \overline{0}$ $\quad i = \overline{1,p},\ \overline{x} \in \mathrm{R}^N$

$$\max_{\overline{\alpha}} L(\overline{\alpha}, \overline{x}^*) = \max_{\overline{\alpha}} \left(\Phi(\overline{x}^*) - \sum_{i=1}^{p} \overline{\alpha}_i g_i(\overline{x}^*) \right) \qquad \min_{\overline{x}} L(\overline{\alpha}^*, \overline{x}) = \min_{\overline{x}} \left(\Phi(\overline{x}) - \sum_{i=1}^{p} \overline{\alpha}_i^* g_i(\overline{x}) \right)$$

$$\max_{\overline{\alpha}} \min_{\overline{x}} L(\overline{\alpha}, \overline{x}) = L(\overline{\alpha}^*, \overline{x}^*) = \Phi(\overline{x}^*) - \sum_{i=1}^{p} \overline{\alpha}_i^* g_i(\overline{x}^*)$$

$\exists! (\overline{\alpha}^*, \overline{x}^*) \in R^p \times R^N$ saddle point

Karush-**K**uhn-**T**ucker KKT conditions

$$\begin{cases} \dfrac{\partial L}{\partial \overline{x}}(\overline{\alpha}^*, \overline{x}^*) = \overline{0} \\ \overline{\alpha}_i^* g_i(\overline{x}_i^*) = 0,\ i = \overline{1,p} \\ g_i(\overline{x}_i^*) \geq 0,\ i = \overline{1,p} \\ \overline{\alpha}_i^* \geq 0,\ i = \overline{1,p} \end{cases}$$

KKT complementarity condition

Generalized Portrait Method

Primal formulation for linear SVM

training set: $\{(\bar{x}_i, y_i) \mid i = \overline{1, p}\} \subset R^N \times \{1, -1\}$

$$y_i = \begin{cases} +1, \bar{x}_i \in A \\ -1, \bar{x}_i \in B \end{cases}$$

$$\min \frac{\|\bar{w}\|^2}{2}$$

$$y_i(\bar{w} \cdot \bar{x}_i + b) \geq 1, \overline{i = 1, p}$$

convex QP with N variables $w_i, i = \overline{1,N}$

Classifier for new instances: $f(\bar{x}) = \text{sign}(\bar{w} \cdot \bar{x} + b)$

Dual formulation for linear SVM

Apply the method of Lagrange multipliers:

$$L(\bar{w}, b, \bar{\alpha}) = \frac{1}{2}\sum_{i=1}^{N} w_i^2 - \sum_{i=1}^{p} \alpha_i [y_i(\bar{w} \cdot \bar{x} + b) - 1]$$

The saddle point is $\min_{\bar{w}} L$ ($\bar{w}$ solves primal)

and $\max_{\bar{\alpha}} L$ ($\bar{\alpha}$ solves dual), $\bar{\alpha} \geq \bar{0}$

$$\frac{\partial L(\bar{w}, b, \bar{\alpha})}{\partial \bar{w}} = 0 \Rightarrow \bar{w} = \sum_{i=1}^{p} \alpha_i y_i \bar{x}_i$$

$$\frac{\partial L(\bar{w}, b, \bar{\alpha})}{\partial b} = 0 \Rightarrow \sum_{i=1}^{p} \alpha_i y_i = 0$$

$$L(\bar{\alpha}) = \frac{1}{2}\Big(\sum_{i=1}^{p} \alpha_i y_i \bar{x}_i\Big)\Big(\sum_{i=1}^{p} \alpha_i y_i \bar{x}_i\Big) - \sum_{i=1}^{p} \alpha_i y_i \bar{x}_i \Big(\sum_{j=1}^{p} \alpha_j y_j \bar{x}_j\Big) - \underbrace{\sum_{i=1}^{p} \alpha_i y_i}_{=0} b + \sum_{i=1}^{p} \alpha_i$$

convex QP with p variables $\alpha_i, i = \overline{1,p}$

$$L(\bar{\alpha}) = \sum_{i=1}^{p} \alpha_i - \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} \alpha_i \alpha_j y_i y_j \bar{x}_i \cdot \bar{x}_j$$

$$\max_{\bar{\alpha}} L(\bar{\alpha})$$

$$\alpha_i \geq 0, \mathrm{i} = \overline{1, \mathrm{p}}, \sum_{i=1}^{p} \alpha_i y_i = 0$$

Support Vectors

KKT complementarity condition: $\overline{\alpha}_i^{*}[y_i^{*}(\overline{w}\cdot\overline{x}_i^{*}+b)-1]=0,\ i=\overline{1,p}$

Primal formulation conditions for linear SVM: $y_i(\overline{w}\cdot\overline{x}_i+b)\geq 1,\ \overline{i=1,p}$

$\alpha_i \geq 0, i=\overline{1,p}$

training set linear separability

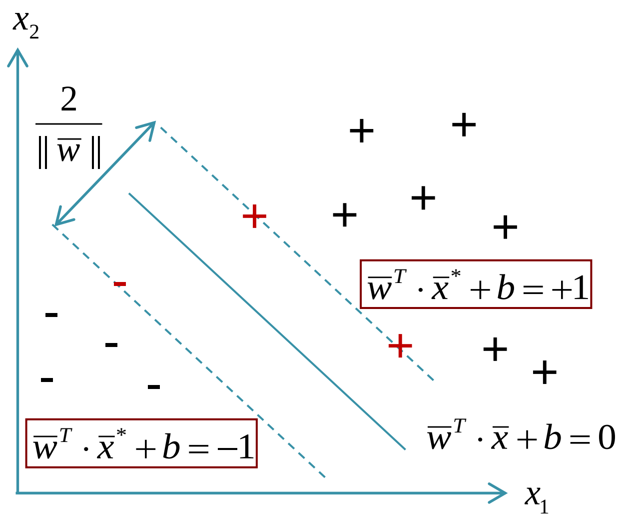
$$\left.\begin{array}{l}\alpha_i \geq 0, i=\overline{1,p} \\ \text{training set linear separability}\end{array}\right\} \exists S\subset\{1,2,\cdots,p\}, \overline{\alpha}_s^{*}>0,\ \forall s\in S$$

$$\boxed{\begin{array}{l} y_s^{*}(\overline{w}\cdot\overline{x}_s^{*}+b)=1 \\ \forall s\in S\end{array}}$$

Support vectors = points with non-zero Lagrangian multipliers

$$\text{card}(S) << p$$

Primal	Supporting hyperplanes
Dual	Support vectors



Kernel functions

$$\Phi : R^2 \to R^3, \Phi(\bar{x}) = \Phi(x_1, x_2) = (x_1{}^2, x_2{}^2, \sqrt{2}x_1x_2)$$

$$\Phi(\bar{x}) \cdot \Phi(\bar{y}) = (x_1{}^2, x_2{}^2, \sqrt{2}x_1x_2) \cdot (y_1{}^2, y_2{}^2, \sqrt{2}y_1y_2) =$$

$$= x_1{}^2 y_1{}^2 + x_2{}^2 y_2{}^2 + 2x_1x_2y_1y_2 =$$

$$= (x_1y_1 + x_2y_2)(x_1y_1 + x_2y_2) =$$

$$= (\bar{x} \cdot \bar{y})(\bar{x} \cdot \bar{y}) = (\bar{x} \cdot \bar{y})^2$$

$$K(\bar{x}, \bar{y}) = \Phi(\bar{x}) \cdot \Phi(\bar{y})$$ kernel function

Examples

$$K(\bar{x}, \bar{y}) = \bar{x} \cdot \bar{y}$$

$$K(\bar{x}, \bar{y}) = (p + \bar{x} \cdot \bar{y})^q$$

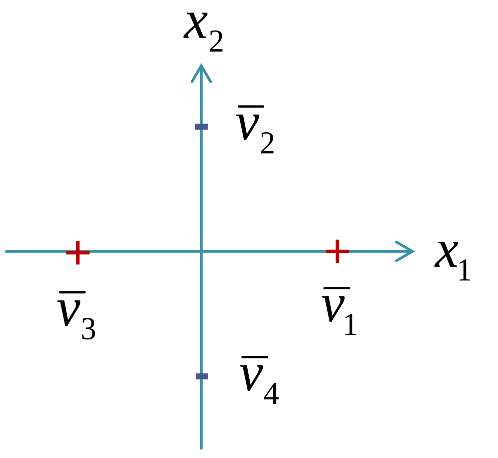
$$K(\bar{x}, \bar{y}) = \exp(-\gamma \| \bar{x} - \bar{y} \|)$$

$$K(\bar{x}, \bar{y}) = \exp(-\gamma \| \bar{x} - \bar{y} \|^2)$$

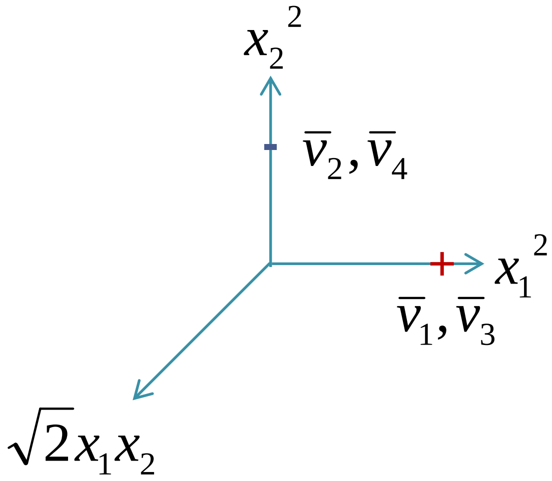
$$K(\bar{x}, \bar{y}) = (p + \bar{x} \cdot \bar{y})^q \exp(-\gamma \| \bar{x} - \bar{y} \|^2)$$

$$K(\bar{x}, \bar{y}) = \tanh(k\bar{x} \cdot \bar{y} - \delta)$$

a kernel expresses a measure of “similarity” between vectors



linear inseparable



linear separable

From Cover’s theorem:

A complex pattern-classification problem, cast in a **high-dimensional space** nonlinearly, is more likely to be linearly separable than in a **low-dimensional space**

Kernel trick

input space	image space
$\hat{f}(\overline{x}) = \mathrm{sign}(\overline{w}^* \cdot \overline{x} + b^*) \Leftrightarrow$	$\hat{f}(\overline{x}) = \mathrm{sign}(\overline{w}^* \cdot \Phi(\overline{x}) + b^*) =$
$\overline{w}^* = \sum_{i=1}^{p} \alpha_i^* y_i \overline{x}_i$	$\overline{w}^* = \sum_{i=1}^{p} \alpha_i^* y_i \Phi(\overline{x}_i)$

$$\hat{f}(\overline{x}) = \mathrm{sign}\left(\sum_{i=1}^{p} \alpha_i^* y_i \Phi(\overline{x}_i) \cdot \Phi(\overline{x}) + b\right) = \mathrm{sign}\left(\sum_{i=1}^{p} \alpha_i^* y_i K(\overline{x}_i, \overline{x}) + b\right)$$

No need to know Φ explicitly!

$\alpha_i^* = 0, i \notin S \Rightarrow$ instead of obtaining a function with complexity proportional to the image space dimension, we obtained an expression with complexity proportional to the number of support vectors

$$\max_{\overline{\alpha}}\left(\sum_{i=1}^{p} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{p} \alpha_i \alpha_j y_i y_j K(\overline{x}_i, \overline{x}_j)\right)$$

$$\text{s.t.} \sum_{i=1}^{p} \alpha_i y_i = 0, \alpha_i \ge 0, i = \overline{1, p}$$

Long history of development

- "You look at stuff like this in a book and you think, well, Vladimir Vapnik just figured this out one Saturday afternoon when the weather was too bad to go outside. That's not how it happened.“ - Patrick Winston

- "The invention of SVMs happened when Bernhard decided to implement Vladimir's algorithm in the three months we had left before we moved to Berkeley. After some initial success of the linear algorithm, Vladimir suggested introducing products of features. I proposed to rather use the **kernel trick** of the 'potential function' algorithm. Vladimir initially resisted the idea because the inventors of the 'potential functions' algorithm (Aizerman, Braverman, and Rozonoer) were from a competing team of his institute back in the 1960’s in Russia! But Bernhard tried it anyways, and the SVMs were born!“ – Isabel Guyon

Bibliography

- C. Bishop, "Pattern Recognition and Machine Learning", Springer (2006)
- B. Boser, I. Guyon, V. Vapnik, "A Training Algorithm for Optimal Margin Classifiers", Proc. of the 5th annual workshop on Computational Learning Theory (COLT '92), pp. 144-152 (1992)
- T. Cover, "Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition", IEEE Transactions on Electronic Computers, EC-14, No. 3, pp. 326-334 (1965)
- I. Guyon - https://www.facebook.com/WomenInMachineLearning/posts/1314864275195877
- L. Hamel, "Knowledge Discovery with Support Vector Machines", Wiley (2009)
- E. Kim, "Everything You Wanted to Know about the Kernel Trick (But Were Too Afraid to Ask)" - http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html
- O. Mangasarian, "Linear and Nonlinear Separation of Patterns by Linear Programming", *Operations Research*, **13**, pp. 444-452 (1965)
- H. Murrell, "Data Mining with R" - http://www.cs.ukzn.ac.za/~hughm/dm/
- F. Rosenblatt, "The Perceptron. A Perceiving and Recognizing Automaton (Project PARA)", Cornell Aeronautical Laboratory, Report No. 85-460-1 (1957)
- A. Smola, "Introduction to Machine Learning, Class 10-701/15-781", Carnegie Mellon University - http://alex.smola.org/teaching/10-701-15/
- A. Statnikov, D. Hardin, I. Guyon, C. Aliferis, "A Gentle Introduction to Support Vector Machines in Biomedicine", The American Medical Informatics Association (AMIA) 2009 Annual Symposium, San Francisco (2009)
- V. Vapnik, "The Nature of Statistical Learning Theory", 2nd Ed., Springer (2000)
- V. Vapnik, "Estimation of Dependences Based on Empirical Data", 2nd ed., Springer (2006)
- R. Vogler, "Testing for Linear Separability with Linear Programming in R" - http://www.joyofdata.de/blog/testing-linear-separability-linear-programming-r-glpk/
- P. Winston, "6.034 Artificial Intelligence. Lecture 16: Learning: Support Vector Machines" - https://www.youtube.com/watch?v=_PwhiWxHK8o