“Machine Learning estadístico para interfaces Cerebro-Computadora”

MODULO III - parte II

Dra. Victoria Peterson
vpeterson@santafe-conicet.gov.ar

Nov 2023

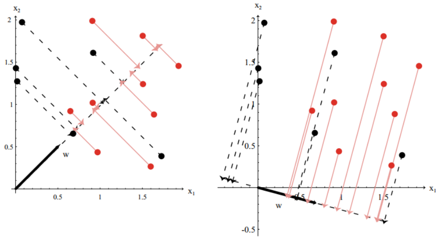
NiCALab

CONICET

I M A L

Problema

Hallar $\mathbf{w}$ tal que $J_{\mathbf{F}} = \dfrac{\mathbf{w}^T \boldsymbol{\Sigma}_b \mathbf{w}}{\mathbf{w}^T \hat{\boldsymbol{\Sigma}}_w \mathbf{w}}$ se maximice.



Fuente: [Duda et al., 2012]

1 Análisis Discriminante Lineal
- Método de transformación lineal
- Regularización
- Regresión

2 Máquinas de soporte vectorial
- Datos linealmente separables
- Datos no-linealmente separables
- Kernel trick

LDA regularizado

⚠

LDA se basa en estimaciones de las matrices de covarianza muestrales $\Rightarrow$ cuando hay pocos datos $\boldsymbol{\Sigma}$ es *mal-condicionada*, y el cálculo de su inversa es *inestable*.

LDA regularizado

⚠

LDA se basa en estimaciones de las matrices de covarianza muestrales ⇒ cuando hay pocos datos $\boldsymbol{\Sigma}$ es *mal-condicionada*, y el cálculo de su inversa es *inestable*.

Regularización, ¿qué es?

Las técnicas de regularización buscan otorgar *estabilidad* a un proceso intrínsecamente inestable.

LDA regularizado

⚠

LDA se basa en estimaciones de las matrices de covarianza muestrales $\Rightarrow$ cuando hay pocos datos $\boldsymbol{\Sigma}$ es *mal-condicionada*, y el cálculo de su inversa es *inestable*.

Regularización, ¿qué es?

Las técnicas de regularización buscan otorgar *estabilidad* a un proceso intrínsecamente inestable.
Permite inducir en la solución ciertas características (deseables) que, por ejemplo, mejorarán el proceso de aprendizaje, evitarán el sobre-entrenamiento y/o generarán soluciones más robustas a outliers.

Shrinkage LDA

Cuando la cantidad de variables es mucho mayor que la cantidad de observaciones ($p >> N_t$) la estimación de $\mathbf{\Sigma}_c = \frac{1}{N_t} \sum_{i \in I_c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^T$ es pobre.

Shrinkage LDA

Cuando la cantidad de variables es mucho mayor que la cantidad de observaciones ($p >> N_t$) la estimación de $\mathbf{\Sigma}_c = \frac{1}{N_t} \sum_{i \in I_c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^T$ es pobre.
Se define entonces

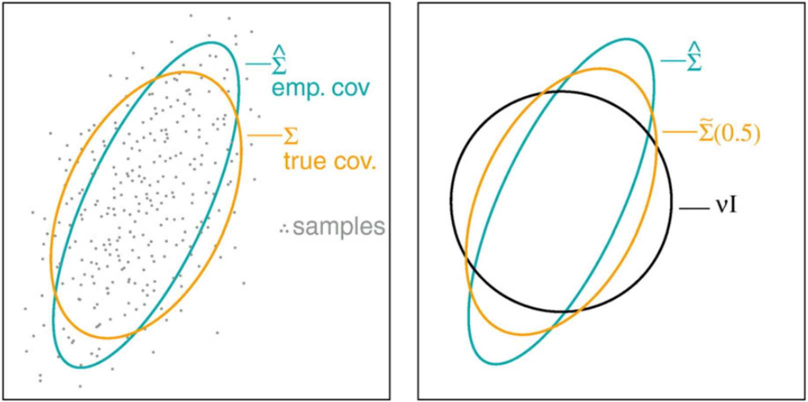
$$\tilde{\mathbf{\Sigma}}_c = (1-\gamma)\mathbf{\Sigma}_c + \gamma \mathbf{I},$$

donde $\gamma \in [0,1]$ es el denominado parámetro de regularización

Shrinkage LDA

Formulación

[Blankertz et al., 2011]



Fuente: [Blankertz et al., 2011]

LDA como modelo de regresión

Sea $\mathbf{t} \in \mathbb{R}^{N_t}$ tal que $t_i = N_2/n \quad \forall\, i/y_i = 1;\ t_i = N_1/n \quad \forall\, i/y_i = 2.$

$$\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha} \in \mathbb{R}^p}{\arg\min} \|\mathbf{t} - X\boldsymbol{\alpha}\|^2$$

LDA como modelo de regresión

Sea $\mathbf{t} \in \mathbb{R}^{N_t}$ tal que $t_i = N_2/n \quad \forall i/y_i = 1$; $t_i = N_1/n \quad \forall i/y_i = 2$.

$$\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha} \in \mathbb{R}^p}{\arg\min} \|\mathbf{t} - X\boldsymbol{\alpha}\|^2$$

Solución:

$$\boldsymbol{\alpha}^* = \frac{N_2 N_1}{N_t} (\boldsymbol{\Sigma}_t)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

LDA como modelo de regresión

Sea $\mathbf{t} \in \mathbb{R}^{N_t}$ tal que $t_i = N_2/n \quad \forall i/y_i = 1;\ t_i = N_1/n \quad \forall i/y_i = 2.$

$$\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha} \in \mathbb{R}^p}{\arg\min} \|\mathbf{t} - X\boldsymbol{\alpha}\|^2$$

Solución:

$$\boldsymbol{\alpha}^* = \frac{N_2 N_1}{N_t} (\boldsymbol{\Sigma}_t)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

$$\boldsymbol{\alpha}^* \propto \mathbf{w}^*$$

LDA como modelo de regresión

Sea $\mathbf{t} \in \mathbb{R}^{N_t}$ tal que $t_i = N_2/n \quad \forall i/y_i = 1;\ t_i = N_1/n \quad \forall i/y_i = 2.$

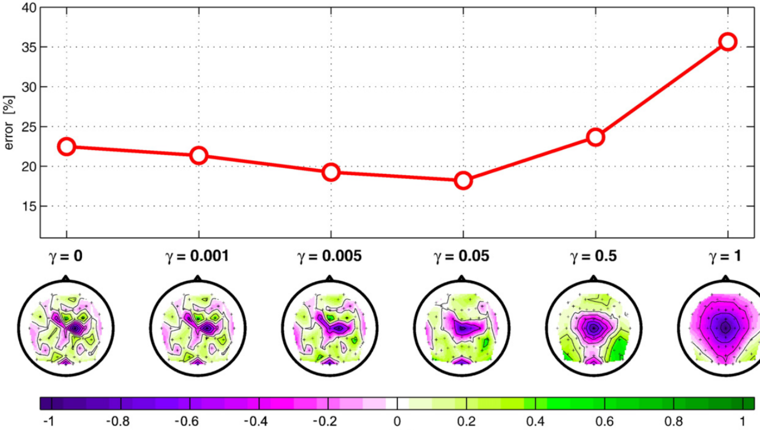
$$\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha} \in \mathbb{R}^p}{\arg\min} \|\mathbf{t} - X\boldsymbol{\alpha}\|^2$$

$$\boldsymbol{\alpha}^* \propto \mathbf{w}^*$$

LDA regularizado

$$\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha} \in \mathbb{R}^p}{\arg\min} \|\mathbf{t} - X\boldsymbol{\alpha} + \|^2 + \gamma \|\boldsymbol{\alpha}\|^2$$

Ejemplo BCI



Fuente: [Blankertz et al., 2011]

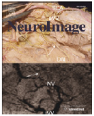
A tutorial

NeuroImage 56 (2011) 814–825

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



ELSEVIER

Single-trial analysis and classification of ERP components — A tutorial

Benjamin Blankertz [a,b,*], Steven Lemm [b], Matthias Treder [a], Stefan Haufe [a], Klaus-Robert Müller [a]

[a] *Berlin Institute of Technology, Machine Learning Laboratory, Berlin, Germany*
[b] *Fraunhofer FIRST, Intelligent Data Analysis Group, Berlin, Germany*

A R T I C L E I N F O

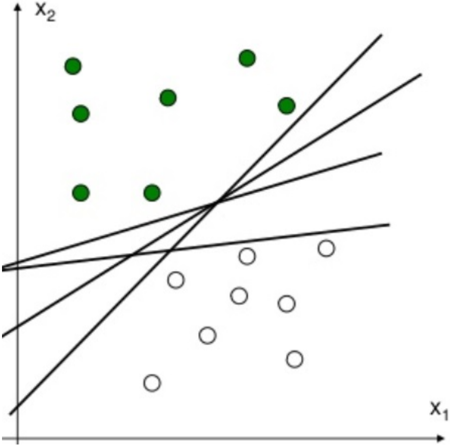
Article history:
Received 15 December 2009
Revised 14 June 2010
Accepted 18 June 2010
Available online 28 June 2010

Keywords:
EEG
ERP
BCI
Decoding
Machine learning
Shrinkage

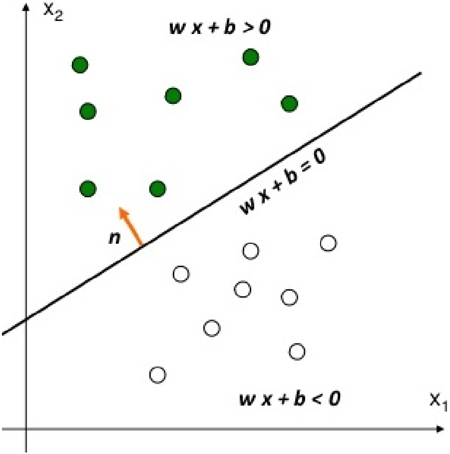
A B S T R A C T

Analyzing brain states that correspond to event related potentials (ERPs) on a single trial basis is a hard problem due to the high trial-to-trial variability and the unfavorable ratio between signal (ERP) and noise (artifacts and neural background activity). In this tutorial, we provide a comprehensive framework for decoding ERPs, elaborating on linear concepts, namely spatio-temporal patterns and filters as well as linear ERP classification. However, the bottleneck of these techniques is that they require an accurate covariance matrix estimation in high dimensional sensor spaces which is a highly intricate problem. As a remedy, we propose to use shrinkage estimators and show that appropriate regularization of linear discriminant analysis (LDA) by shrinkage yields excellent results for single-trial ERP classification that are far superior to classical LDA classification. Furthermore, we give practical hints on the interpretation of what classifiers learned from the data and demonstrate in particular that the trade-off between goodness-of-fit and model complexity in regularized LDA relates to a morphing between a difference *pattern* of ERPs and a spatial *filter* which cancels non task-related brain activity.

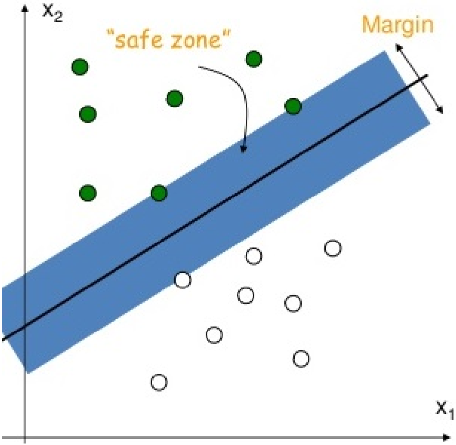
¿Cuál hiperplano es el mejor?



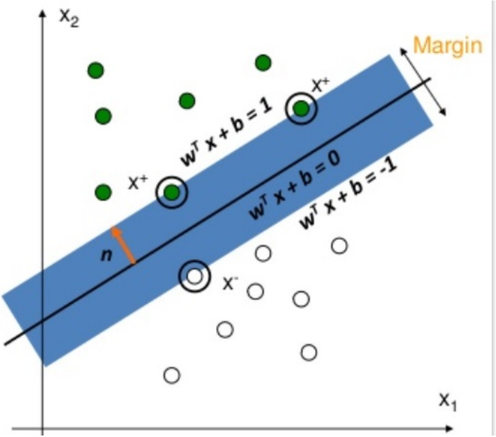
Definición de márgenes



Definición de márgenes



Maximización del margen



Datos linealmente separables

$$\mathbf{w}^T\mathbf{x}+b\begin{cases}>0\\<0\end{cases}\quad\Rightarrow\mathbf{x}\in\begin{cases}\mathscr{X}_1\ (y_i=+1)\\\mathscr{X}_2\ (y_i=-1)\end{cases}$$

Datos linealmente separables

$$\mathbf{w}^T\mathbf{x}+b\begin{cases}>0\\<0\end{cases}\Rightarrow \mathbf{x}\in\begin{cases}\mathscr{X}_1\ (y_i=+1)\\ \mathscr{X}_2\ (y_i=-1)\end{cases}$$

$$y_i(\mathbf{w}^T\mathbf{x}+b)>0\,\forall\, i$$

Datos linealmente separables

$$\mathbf{w}^T\mathbf{x}+b\begin{cases} >0 \\ <0 \end{cases} \Rightarrow \mathbf{x}\in \begin{cases} \mathscr{X}_1 \; (y_i=+1) \\ \mathscr{X}_2 \; (y_i=-1) \end{cases}$$

$$y_i(\mathbf{w}^T\mathbf{x}+b)>0\,\forall i$$

Hiperplanos canónicos

$$H_1:\mathbf{w}^T\mathbf{x}+b=+1; \quad H_2:\mathbf{w}^T\mathbf{x}+b=-1$$

Datos linealmente separables

$$\mathbf{w}^T\mathbf{x} + b \begin{cases} > 0 \\ < 0 \end{cases} \Rightarrow \mathbf{x} \in \begin{cases} \mathscr{X}_1 \ (y_i = +1) \\ \mathscr{X}_2 \ (y_i = -1) \end{cases}$$

$$y_i(\mathbf{w}^T\mathbf{x} + b) > 0 \, \forall \, i$$

Hiperplanos canónicos

$$H_1 : \mathbf{w}^T\mathbf{x} + b = +1; \quad H_2 : \mathbf{w}^T\mathbf{x} + b = -1$$

$$\mathbf{w}^T\mathbf{x} + b \geq +1; \text{ para } y_i = +1$$

$$\mathbf{w}^T\mathbf{x} + b \leq -1; \text{ para } y_i = -1$$

Datos linealmente separables

$$\mathbf{w}^T\mathbf{x}+b\begin{cases}>0\\<0\end{cases}\Rightarrow \mathbf{x}\in\begin{cases}\mathscr{X}_1\ (y_i=+1)\\ \mathscr{X}_2\ (y_i=-1)\end{cases}$$

$$y_i(\mathbf{w}^T\mathbf{x}+b)>0\,\forall i$$

Hiperplanos canónicos

$$H_1:\mathbf{w}^T\mathbf{x}+b=+1;\quad H_2:\mathbf{w}^T\mathbf{x}+b=-1$$

$$\mathbf{w}^T\mathbf{x}+b\geq+1;\text{ para } y_i=+1$$

$$\mathbf{w}^T\mathbf{x}+b\leq-1;\text{ para } y_i=-1$$

La distancia entre estos dos hiperplanos y el plano de separación $\mathbf{w}^T\mathbf{x}+b=0$ es el vector normal $1/|\mathbf{w}|$, denominado **margen**.

Datos linealmente separables

$$\mathbf{w}^T\mathbf{x}+b \begin{cases} >0 \\ <0 \end{cases} \Rightarrow \mathbf{x}\in \begin{cases} \mathscr{X}_1 \ (y_i=+1) \\ \mathscr{X}_2 \ (y_i=-1) \end{cases}$$

$$y_i(\mathbf{w}^T\mathbf{x}+b)>0\,\forall\, i$$

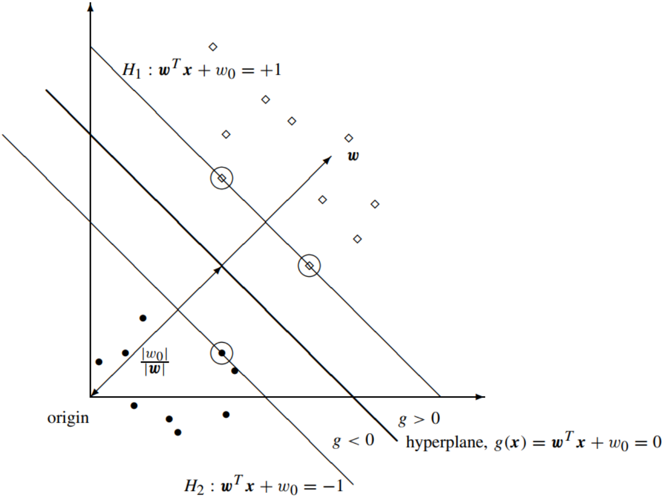
Hiperplanos canónicos

$$H_1 : \mathbf{w}^T\mathbf{x}+b=+1; \quad H_2 : \mathbf{w}^T\mathbf{x}+b=-1$$

$$\mathbf{w}^T\mathbf{x}+b \geq +1; \text{ para } y_i=+1$$

$$\mathbf{w}^T\mathbf{x}+b \leq -1; \text{ para } y_i=-1$$

La distancia entre estos dos hiperplanos y el plano de separación $\mathbf{w}^T\mathbf{x}+b=0$ es el vector normal $1/|\mathbf{w}|$, denominado **margen**. Los puntos que caen sobre los hiperplanos canónicos se denominan vectores soporte.



Fuente: [Webb, 2003]

Problema

$$\text{máx}\ \|\mathbf{w}\|_2^2 \quad s.a \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1\ \forall i$$

Problema

$$\max \|\mathbf{w}\|_2^2 \quad s.a \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \; \forall i$$

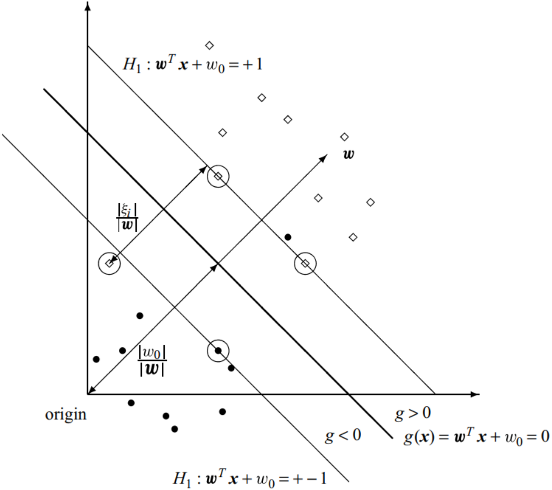
$$\mathcal{L}_p = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{n} \lambda_i (y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1)$$

Problema

$$\text{máx}\ \|\mathbf{w}\|_2^2 \quad s.a \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1\ \forall i$$

$$\mathscr{L}_p = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{n} \lambda_i(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1)$$

Problema de optimización cuadrático con restricciones lineales



Fuente: [Webb, 2003]

Relajamos las restricciones

$$H_1 : \mathbf{w}^T\mathbf{x} + b = +1; \quad H_2 : \mathbf{w}^T\mathbf{x} + b = -1$$

$$\begin{aligned} \mathbf{w}^T\mathbf{x} + b &\geq +1 - \zeta_i; \quad \text{para } y_i = +1, \\ \mathbf{w}^T\mathbf{x} + b &\leq -1 + \zeta_i; \quad \text{para } y_i = -1, \\ \zeta_i &\geq 0. \end{aligned}$$

Problema regularizado

$$\text{máx}\, \frac{1}{2}\|\mathbf{w}\|_2^2 + \alpha \sum_i \zeta_i \quad s.a \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \zeta_i \, \forall i$$

Relajamos las restricciones

$$H_1 : \mathbf{w}^T\mathbf{x} + b = +1; \quad H_2 : \mathbf{w}^T\mathbf{x} + b = -1$$

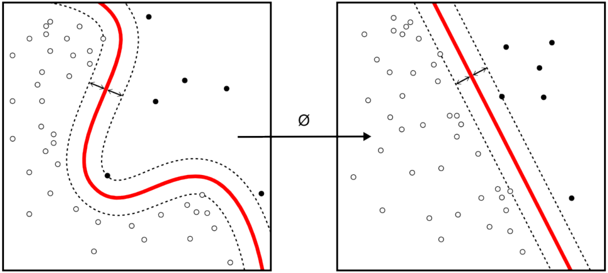
$$\begin{aligned} \mathbf{w}^T\mathbf{x} + b &\geq +1 - \zeta_i; \quad \text{para } y_i = +1, \\ \mathbf{w}^T\mathbf{x} + b &\leq -1 + \zeta_i; \quad \text{para } y_i = -1, \\ \zeta_i &\geq 0. \end{aligned}$$

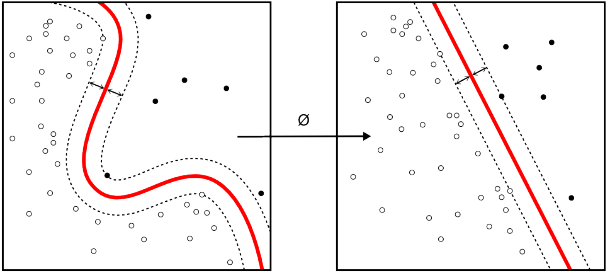
Problema regularizado

$$\text{máx } \frac{1}{2}\|\mathbf{w}\|_2^2 + \alpha \sum_i \zeta_i \quad s.a \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \zeta_i \; \forall i$$

$$\mathcal{L}_p = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \alpha \sum_i \zeta_i - \sum_{i=1}^{n} \lambda_i(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \zeta_i) - \sum_{i=1}^{n} \delta_i \zeta_i,$$

siendo λ y δ los multiplicadores de Lagrange.


Ø

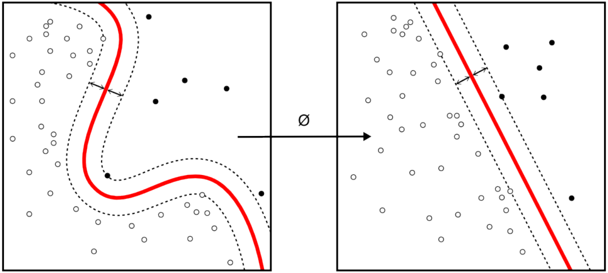


Función discriminativa

Clasificadores lineales

$$f(\mathbf{x}) = \sum_{i=1}^{p} w_i x_i + b = \mathbf{w}^T\mathbf{x} + b,$$

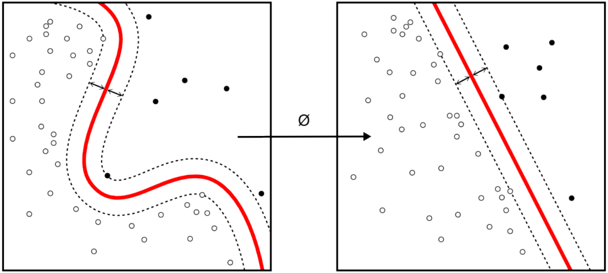
tal que si $sgn(\mathbf{w}^T\mathbf{x} + b) > 0 \Rightarrow \mathbf{x} \in \mathscr{X}_1$



Clasificadores basados en Kernels

$$f(\mathbf{x}) = \sum_{i=1}^{p} w_i \phi(x)_i + b = \mathbf{w}^T \phi(\mathbf{x}) + b,$$

tal que si $sgn(\mathbf{w}^T \phi(\mathbf{x}) + b) > 0 \Rightarrow \mathbf{x} \in \mathscr{W}_1$



Clasificadores basados en Kernels

$$f(\mathbf{x}) = \sum_{i=1}^{p} w_i \phi(x)_i + b = \mathbf{w}^T \phi(\mathbf{x}) + b,$$

tal que si $sgn(\mathbf{w}^T \phi(\mathbf{x}) + b) > 0 \Rightarrow \mathbf{x} \in \mathscr{W}_1$

Kernel: $K(\mathbf{x}, \mathbf{y}) = \phi^T(\mathbf{x})\phi(\mathbf{y})$

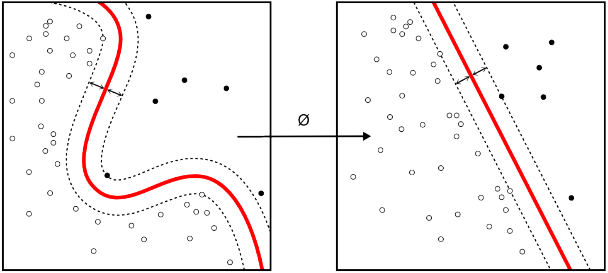


Table 5.2 Support vector machine kernels

Nonlinearity	Mathematical form $K(\boldsymbol{x}, \boldsymbol{y})$
Polynomial	$(1 + \boldsymbol{x}^T \boldsymbol{y})^d$
Gaussian	$\exp(-\lvert\boldsymbol{x} - \boldsymbol{y}\rvert^2/\sigma^2)$
Sigmoid	$\tanh(k\boldsymbol{x}^T \boldsymbol{y} - \delta)$

Fuente: [Webb, 2003]

Bibliografía utilizada I

Blankertz, B., Lemm, S., Treder, M., Hauf, S., and Müler, K. R. (2011). Single-trial analysis and classification of ERP component- a tutorial. *Neuroimage*, 56:814–825.

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Webb, A. R. (2003). *Statistical pattern recognition*. John Wiley & Sons.