Capítulo 3

Redes de Regularización y Máquinas de Soporte Vectorial

"The real act of discovery consists not in finding new lands but seeing with new eyes"
Marcel Proust.

Para comenzar

Las redes de regularización, fruto del principio inductivo de penalización, son una herramienta válida de trabajo que poseen un profundo fundamento teórico. Sin embargo, sus propiedades de aproximación exclusivamente asintóticas y la expansión de la función solución sobre un elevado número de vectores convierten en poco prácticas este tipo de máquinas en su definición original. La búsqueda de una reducción en la expansión de la solución hizo fijar los ojos de algunos investigadores en el buen comportamiento de las máquinas de soporte vectorial. Este tipo de máquinas poseen además la característica de estar basadas en un principio inductivo, el SRM, que tiene en cuenta la finitud del conjunto de entrenamiento y su construcción se desarrolla sobre espacios de aproximación anidados.

Tras realizar una breve introducción a los planteamientos de las máquinas de soporte vectorial para los problemas de clasificación y su extensión a problemas de regresión, el presente Capítulo está especialmente dedicado a justificar el cambio de visión del presente estudio hacia la teoría del aprendizaje estadístico. Como se verá, aunque el campo de investigación aún está abierto, resulta factible la existencia de un camino hacia el establecimiento de un marco unificado que permita derivar tanto las redes de regularización como las máquinas de soporte vectorial, por lo que el tratamiento del problema general, aceptando algunos condicionantes sobre la forma

de las funciones de Green[1] , es equivalente.

3.1 Máquinas de Soporte Vectorial

Utilizando el principio de inducción de Minimización del Riesgo Estructural, SRM, como proceso de inferencia, se desea construir un método de aprendizaje que permita dar respuesta al Problema General de Aprendizaje a partir de Ejemplos, PGAE, definido en la página 18. Debido a que el principio SRM se basa en un proceso de construcción de la solución sobre espacios anidados cuya amplitud o capacidad está determinada por funciones indicador, el tipo de problema de aprendizaje para el que resulta más sencillo hallar un método ofreciendo respuesta es el de clasificación binaria mediante hiperplanos — método SVMC, del inglés *Support Vector Machine for Classification* — .

3.1.1 SVM para Clasificación

Se desea construir un hiperplano que separe las dos clases, etiquetadas $y \in \{-1, +1\}$, de forma que la distancia entre el hiperplano óptimo y el patrón de entrenamiento más cercano — *margen* — sea máxima, con la intención de forzar la generalización de la máquina de aprendizaje [Burges, 1998], [Smola, 1998], [Vapnik, 1995].

La expansión del método SVMC a funciones de decisión no lineales se realiza introduciendo el espacio de entrada $\mathcal{X} \subseteq \mathbb{R}^d$ en otro espacio de mayor dimensión $\mathcal{F}$, denominado *espacio de características*, dotado de producto interno, vía una inyección no lineal, $\iota : \mathcal{X} \subseteq \mathbb{R}^d \rightarrow \mathcal{F}_{\langle \cdot,\cdot \rangle}$, de forma que el *hiperplano óptimo*

$$f(\mathbf{x}, \omega) = \langle \omega, \mathbf{x} \rangle_{\mathcal{F}} + b = k(\omega, \mathbf{x}) + b\,, \tag{3.1}$$

hallado linealmente en el espacio de características $\mathcal{F}$, con

$$k(\mathbf{a}, \mathbf{b}) = \langle \mathbf{a}, \mathbf{b} \rangle_{\mathcal{F}} = (\iota(\mathbf{a}) \cdot \iota(\mathbf{b}))\ , \tag{3.2}$$

correspondiendo a un núcleo de Hilbert[2] , permite definir como *función decisión*

$$h(\mathbf{x}) = sign(f(\mathbf{x}, \omega))\ . \tag{3.3}$$

[1] Estos condicionantes hacen referencia a los espacios de aproximación que pueden ser considerados en función de los operadores de regularización P escogidos y las funciones de Green asociadas. Aunque en principio el abanico de posibilidades es amplio, en la práctica el número de funcionales considerados como penalizadores es reducido y cumplen un gran número de propiedades de regularidad.

[2] Cualquier función simétrica continua $k(\mathbf{a}, \mathbf{b})$ puede ser usada como un núcleo de Hilbert si satisface la condición de Mercer $\int\int k(\mathbf{a}, \mathbf{b})\, g(\mathbf{a})\, g(\mathbf{b})\, d\mathbf{a} d\mathbf{b} \geq 0 \quad \forall g$

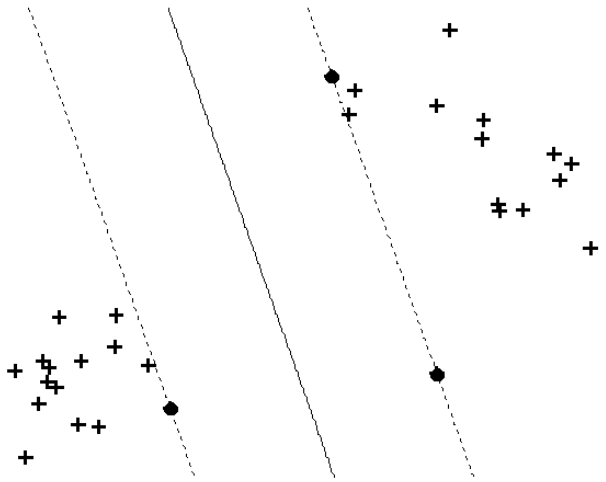

Figura 3.1: El margen es la distancia perpendicular entre el hiperplano separador y el hiperplano que pasa sobre los puntos más cercanos, los vectores soporte.

Con objeto de definir de forma única el hiperplano óptimo — *forma canónica* — deben ser añadidas las restricciones

$$y_i \cdot (\langle \omega, \mathbf{x}_i \rangle_{\mathcal{F}} + b) \geq 1 - \xi_i \quad i = 1, \ldots, \ell \,, \tag{3.4}$$

sobre el conjunto de entrenamiento $\mathcal{T}$, donde las variables artificiales

$$\xi_i \geq 0 \quad i = 1 \ldots, \ell \,, \tag{3.5}$$

son introducidas para permitir que existan ejemplos violando la restricción impuesta por el margen — *margen débil* — pues debe considerarse la posibilidad de que las clases a ser separadas se solapen o que los ejemplos contengan ruido.

En resumen, el hiperplano óptimo en forma canónica de margen débil es hallado solucionando el problema de optimización restringida:

$$\arg\min R_{SVMC}(\omega, \xi) = \frac{1}{2} \|\omega\|^2_{\mathcal{F}} + C \sum_{i=1}^{\ell} \xi_i \,, \tag{3.6}$$

sujeto a (3.4)–(3.5).

Introduciendo multiplicadores de Lagrange y realizando algunas substituciones, se obtiene el enunciado dual de Wolfe del problema de optimización original: Hallar multiplicadores $\alpha_i \geq 0$ que hagan mínimo el funcional

$$W(\alpha) = \frac{1}{2} \sum_{i,j=1}^{\ell} y_i \alpha_i \cdot k(\mathbf{x}_i, \mathbf{x}_j) \cdot y_j \alpha_j - \sum_{i=1}^{\ell} \alpha_i \,, \tag{3.7}$$

sujeto a

$$0 \leq \alpha_i \leq C \quad i = 1, \ldots, \ell \,, \tag{3.8}$$

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0 \,. \tag{3.9}$$

El hiperplano separador solución puede ser escrito como

$$f(\mathbf{x}) = \sum_{i=1}^{SV} \alpha_i y_i \cdot k(\mathbf{x}_i, \mathbf{x}) + b \,, \tag{3.10}$$

donde b es computado utilizando las condiciones complementarias de Karush-Kuhn-Tucker

$$\alpha_i \cdot [y_i \cdot (\langle \omega, \mathbf{x}_i \rangle_{\mathcal{F}} + b) - 1] = 0 \quad i = 1, \ldots, \ell \,. \tag{3.11}$$

De entre todos los elementos del conjunto de aprendizaje, sólo algunos de ellos poseen un peso asociado α_i no nulo en la expansión (3.10). Estos elementos caen sobre el margen — se cumple alguna restricción estricta en (3.4) — y reciben el nombre de *vectores soporte*.

3.1.2 SVM para Regresión

Para generalizar el método SV a la estimación de regresiones — método SVMR, del inglés *Support Vector Machine for Regression* — [Cortes and Vapnik, 1995], [Smola and Schölkopf, 1998b], es necesario construir un elemento análogo al margen en el espacio de valores de salida, $y \in \mathbb{R}$, mediante el uso de la función de coste ε-insensitiva[3] de Vapnik

$$|y - f(\mathbf{x}, \omega)|_{\varepsilon} = \max\{0, |y - f(\mathbf{x}, \omega)| - \varepsilon\} \,. \tag{3.12}$$

Para un $\varepsilon \geq 0$ dado, el problema de optimización restringida asociado para el caso de estimación de regresiones es

$$\arg\min R_{SVMR}\left(\omega, \varphi^{(*)}\right) = \frac{1}{2} \|\omega\|_{\mathcal{F}}^2 + D \sum_{i=1}^{\ell} (\varphi_i + \varphi_i^*) \,, \tag{3.13}$$

sujeto a

$$\begin{cases} (\langle \omega, \mathbf{x}_i \rangle_{\mathcal{F}} + b) - y_i & \leq \varepsilon + \varphi_i \\ y_i - (\langle \omega, \mathbf{x}_i \rangle_{\mathcal{F}} + b) & \leq \varepsilon + \varphi_i^* \end{cases} \quad i = 1, \ldots, \ell \,, \tag{3.14}$$

[3] Nótese que esta función es tomada como función de coste para el caso de un problema de regresión, a diferencia de la habitual norma L_2, tal como se definió en el Capítulo 1.

$$\varphi_i, \varphi_i^* \geq 0 \quad i = 1, \ldots, \ell \,. \tag{3.15}$$

Introduciendo multiplicadores de Lagrange, se obtiene el problema de optimización restringida: Hallar multiplicadores $\alpha_i, \alpha_i^* \geq 0$ que hagan mínimo el funcional

$$\begin{aligned} W(\alpha, \alpha^*) = & \varepsilon \sum_{i=1}^{\ell} (\alpha_i^* + \alpha_i) - \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) \, y_i + \\ & \frac{1}{2} \sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i) \cdot k(\mathbf{x}_i, \mathbf{x}_j) \cdot \left(\alpha_j^* - \alpha_j\right) \,, \end{aligned} \tag{3.16}$$

sujeto a

$$0 \leq \alpha_i, \alpha_i^* \leq D \quad i = 1, \ldots, \ell \,, \tag{3.17}$$

$$\sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) = 0 \,. \tag{3.18}$$

La regresión estimada toma la forma

$$f(\mathbf{x}) = \sum_{i=1}^{SV} (\alpha_i^* - \alpha_i) \cdot k(\mathbf{x}_i, \mathbf{x}) + b \,. \tag{3.19}$$

La solución se expande de nuevo en términos de un subconjunto del conjunto de entrenamiento, y b es calculado de (3.14) en su forma de igualdad sobre los vectores soporte.

3.2 Hacia un Marco Unificado para SVMs y RNs

Tal como se estableció en el Capítulo 1, los principios inductivos de regularización y SRM establecen la introducción de información *a priori* sobre la forma de la función solución sin necesidad de predeterminar la función densidad de probabilidad desconocida de los espacios de trabajo. El principio de regularización utiliza para ello un operador de regularización que permite asegurar la obtención de la solución de forma asintótica sobre espacios de funciones anidados si el número de elementos del conjunto de aprendizaje tiende a infinito, mientras que el principio SRM también basa su definición sobre espacios anidados pero asegurando una cota superior del funcional de riesgo con sólo un conjunto de datos empíricos finito.

Aunque es evidente la no equivalencia entre ambos procesos de inferencia, sus similitudes son exportadas a los métodos de aprendizaje que tienen su razón de ser en estos principios, lo que ha motivado que un gran número de investigadores que tratan el problema de aprendizaje desde perspectivas dispares hayan intentado establecer un marco de trabajo que permita tratar las SVMs y las RNs como casos particulares de un

mismo tipo de método de aprendizaje general, llámese Métodos Núcleo — en inglés, *Kernel Methods* — [Campbell, 2000], resaltando la importancia de la función núcleo generadora del espacio de características, o Clasificadores de Margen Amplio — en inglés, *Large Margin Classifiers* — , incidiendo en el motivo práctico que maximiza la generalización.

El contenido de esta Sección ha sido elaborado con la finalidad de mostrar los nexos de unión entre ambas visiones de un mismo problema, al tiempo que establece un esquema de estudio que permite justificar el posterior uso de máquinas de aprendizaje SVM en el problema de clasificación multiclase.

3.2.1 Conexión vía Operadores de Regularización

Puede observarse la similitud en la expresión del funcional de riesgo para el principio SRM y para el de regularización en el caso de regresión si el problema de optimización restringida (3.13)–(3.14)–(3.15) que debe ser resuelto para hallar la máquina SVMR solución es traducido en una expresión equivalente a

$$\min R_{SVMR}(\omega) = R_{emp}(\omega) + \lambda \cdot \|\omega\|_{\mathcal{F}}^2 , \tag{3.20}$$

mientras que el problema variacional que conduce a las RNs ya es conocido que adopta la forma (2.6)

$$\min R_{emp}^{reg}(\omega) = R_{emp}(\omega) + \lambda \cdot \|Pf\|^2 , \tag{3.21}$$

si el penalizador se considera en el formato (2.9).

En cuanto a la estructura de la función regresión estimada, resulta en una expansión del estilo (3.19) sobre todos los patrones de aprendizaje como solución del problema dual (3.16)–(3.17)–(3.18) para el caso de las máquinas SVMR. En el caso de las RNs, si se establece (3.19) como estructura solución deseable[4] y se utiliza la función ε-insensitiva de Vapnik (3.12) como función de coste, entonces el problema dual de Wolfe a solucionar tiene como expresión a minimizar

$$\begin{aligned} W(\alpha, \alpha^*) =& \varepsilon \sum_{i=1}^{\ell} (\alpha_i^* + \alpha_i) - \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) y_i \\ &+ \frac{1}{2} \sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i) \cdot \left(\mathbf{K}\mathbf{D}^{-1}\mathbf{K}\right)_{ij} \cdot \left(\alpha_j^* - \alpha_j\right) \end{aligned} \tag{3.22}$$

donde $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ y $D_{ij} = \left(\left(\widehat{P}k\right)(\mathbf{x}_i, \cdot) \cdot \left(\widehat{P}k\right)(\mathbf{x}_j, \cdot)\right)$.

Por tanto puede asegurarse la equivalencia entre ambos métodos, con las restricciones que ya han sido asumidas, en el caso siguiente

[4] Nótese que k es alguna función simétrica que no necesariamente cumple la condición de Mercer.

Proposición 3.1 ([Smola and Schölkopf, 1998a]). *Sea P un operador de regularización y sea G la función de Green de $\widehat{P}P$, entonces G es un núcleo de Hilbert tal que $\mathbf{D} = \mathbf{K}$. Las SVMRs que utilizan G minimizan el funcional de riesgo $R_{emp}^{reg}(\omega)$ con P como operador regularización.*

Esta proposición establece una condición suficiente pero no necesaria para satisfacer $\mathbf{D} = \mathbf{K}$. La implicación contraria sólo puede enunciarse en el caso discreto

Proposición 3.2 ([Smola et al., 1998]). *Dado un operador de regularización P que permite una expansión de $\widehat{P}P$ en un sistema discreto de vectores y valores propios (η_i, ψ_i) y un núcleo k con*

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^{\infty} \frac{d_i}{\eta_i} \psi_i(\mathbf{x}_i) \psi_i(\mathbf{x}_j) \tag{3.23}$$

donde $d_i \in \{0, 1\}$ y $\sum_{i=1}^{\infty} \frac{d_i}{\eta_i}$ es convergente, entonces $\mathbf{K} = \mathbf{D}$.

Se establece así una cierta equivalencia basada en la adopción de un operador de regularización y una función de Green apropiados. Un estudio sobre las implicaciones de esta equivalencia puede ser hallado en [Angulo and Català, 1999].

3.2.2 Espacios de Hilbert definidos por Núcleos

Al igual que en la conexión del apartado anterior, la búsqueda de un marco de trabajo único para las SVMs y las RNs tiene por finalidad la importación de las buenas características que presentan las SVMs, en especial su desarrollo sobre un número pequeño de vectores soporte, hacia la teoría de regularización. Es por ello que el siguiente paso hacia la unificación tiene por objetivo la derivación del algoritmo SVM desde el marco de la teoría de regularización mediante el uso de espacios generados por núcleos.

Proposición 3.3. *Para todo espacio de Hilbert generado por núcleo — RKHS, del inglés Reproducing Kernel Hilbert Space — , $\mathcal{H}$, es posible hallar una función definida positiva $k(\mathbf{x}, \mathbf{y})$[5] , denominada* núcleo generador *de $\mathcal{H}$ con la siguiente* propiedad de generación

$$f(\mathbf{x}) = \langle f(\mathbf{y}), k(\mathbf{y}; \mathbf{x}) \rangle_{\mathcal{H}} . \tag{3.24}$$

Siguiendo la exposición de [Girosi, 1998], si se toma como problema variacional la minimización del funcional de riesgo tal como está definido en (3.20) con función de coste ε-insensitiva sobre un espacio $\mathcal{H}$ que es RKHS con núcleo

$$k(\mathbf{y}; \mathbf{x}) = \sum_{i=1}^{\infty} \eta_i \psi_i(\mathbf{x}) \psi_i(\mathbf{y}) , \tag{3.25}$$

[5] *La función k actúa de forma similar a la función delta en L_2, aunque L_2 no es un RKHS.*

y el funcional penalizador es $\|f\|^2_{\mathcal{H}}$, entonces la función solución en $\mathcal{H}$ tiene una única expansión de la forma

$$f(\mathbf{x}, \mathbf{c}) = \sum_{i=1}^{\infty} c_i \psi_i(\mathbf{x}) + b\,, \tag{3.26}$$

y su norma es

$$\|f\|^2_{\mathcal{H}} = \sum_{i=1}^{\infty} \frac{c_i^2}{\eta_i}\,, \tag{3.27}$$

por lo que la solución toma la forma de una expansión sobre las entradas del conjunto de entrenamiento $\mathcal{T}$

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} a_i k(\mathbf{x}; \mathbf{x}_i) + b\,. \tag{3.28}$$

Otros trabajos a ser considerados siguiendo este intento de obtener las SVMs desde el marco de la regularización son [Poggio and Girosi, 1998] y [Wahba, 1999].

3.2.3 Conexión vía el Principio Inductivo SRM

Los intentos anteriores de establecer un marco común de trabajo para las SVMs y las RNs han tenido más bien como objetivo la derivación desde el marco de la regularización de las máquinas SVM. Es por ello que el problema de aprendizaje principalmente tratado ha sido el de regresión, que es el problema original tratado por la teoría de regularización.

En [Evgeniou et al., 1999] se establece una extensión técnica del principio inductivo SRM de Vapnik que permite que las RNs y las SVMs puedan ser entendidas dentro del marco SRM, considerando como funcional de riesgo expresiones que facilitan esta unificación de la forma

$$R_{uni}(\omega) = \frac{1}{\ell} \sum_{p=1}^{\ell} L(y_p, f(\mathbf{x}_p, \omega)) + \lambda \cdot \|f\|^2_{\mathcal{H}}\,, \tag{3.29}$$

donde $L(\cdot, \cdot)$ es una función de coste y $\|f\|^2_{\mathcal{H}}$ es una norma en un RKHS $\mathcal{H}$ definido por una función definida positiva k. La adopción de este formato de funcional de riesgo permite hacer corresponder las RNs y las SVMs con la minimización de $R_{uni}(\omega)$ en la ecuación (3.29) para diferentes elecciones de la función de coste:

- Redes de Regularización clásicas (RN)

$$L(y_p, f(\mathbf{x}_p, \omega)) = (y_p - f(\mathbf{x}_p, \omega))^2\,, \tag{3.30}$$

- Máquinas de Soporte Vectorial para Regresión (SVMR)

$$L\left(y_p, f\left(\mathbf{x}_p, \omega\right)\right) = \left|y_p - f\left(\mathbf{x}_p, \omega\right)\right|_\varepsilon , \tag{3.31}$$

- Máquinas de Soporte Vectorial para Clasificación (SVMC)

$$L\left(y_p, f\left(\mathbf{x}_p, \omega\right)\right) = \left|1 - y_p f\left(\mathbf{x}_p, \omega\right)\right|_+ , \tag{3.32}$$

donde $|\cdot|_\varepsilon$ es la función de coste ε-insensitiva de Vapnik, $|\mathbf{x}|_+ = \mathbf{x}$ si $\mathbf{x}$ es positivo y 0 en cualquier otro caso, y $y_p \in \mathbb{R}$ en RN y SVMR, mientras que $y_p \in \{-1, 1\}$ en SVMC.

Observando las funciones de coste, cabe destacar que la RN posee la norma L_2 como función de coste tal como se estableció en (1.6) para el caso de problemas de regresión, mientras que la SVMR todavía hace uso de una norma especialmente diseñada. Es este uno de los puntos flacos en el debe de las SVMRs y uno de los motivos que incitan a los investigadores a intentar hallar la conexión entre RN y SVMR que permitiría por una parte adoptar la norma original para regresión y por otra explotar el buen comportamiento de las SVMs favorecido por el principio SRM en el que están basadas. En cuanto al caso de clasificación, si bien la nueva función de coste coincide para los valores negativos con la definición original (1.7), su definición para valores positivos está estrechamente relacionada con las restricciones (3.4) que aparecen en el problema de optimización primal (3.6) y toma el valor de las variables artificiales (3.5) que se introducen en el caso de considerar margen débil.

3.2.4 Implicaciones Bayesianas

La visión del aprendizaje supervisado desde la perspectiva de espacios de Hilbert desarrollados sobre una función generadora, RKHS, no es única a los investigadores y desarrolladores de máquinas de aprendizaje a partir del principio SRM o el de regularización. Una importante comunidad científica implementa algoritmos basados en el principio de inferencia bayesiana que, como ya se comentó en el Capítulo 1, establece *a priori* la forma de la densidad de probabilidad. Aunque este tipo de inferencia no ha sido trabajada en el presente estudio, sí creo oportuno mencionar alguno de los intentos de conexión entre el principio bayesiano y el SRM.

En [Tipping, 2000] es presentada un tipo de máquina, la Máquina de Vectores Relevantes — RVM, del inglés *Relevance Vector Machine* — que representa un tratamiento bayesiano de un modelo lineal generalizado de forma funcional idéntica a la SVM. Asumiendo que, por ejemplo en el caso de regresión, la distribución de los datos es gaussiana, $p\left(y|\mathbf{x}\right) \sim \mathcal{N}\left(y|f\left(\mathbf{x}\right), \sigma^2\right)$, se obtiene una RVM con un rendimiento similar a la SVM sobre problemas tipo pero con un número menor de vectores sobre los que expandir la función solución. En el caso de clasificación, a diferencia de las

SVMs que concentran los vectores soporte sobre el margen delimitador de las clases, las RVMs toman como vectores relevantes aquellos que se hallan más alejados del margen separador.

En el trabajo [Evgeniou et al., 1999], los autores también intentan establecer una somera interpretación bayesiana de los principios de regularización y SRM, mientras que en [Herbrich et al., 1999a] es introducida la Máquina del Punto de Bayes — BPM, del inglés *Bayes Point Machine* —, en un enfoque puramente bayesiano, constituyendo un puente interesante entre la visión bayesiana de las máquinas de aprendizaje y la teoría de aprendizaje estadístico. Este estudio permite descubrir que la SVMC corresponde al centro de la hiperesfera de radio máximo inscribible en el espacio de características $\mathcal{F}$ siendo las fronteras del espacio $\mathcal{F}$ con las cuales la hiperesfera realiza un contacto tangencial los vectores soporte, mientras que el punto de Bayes es una aproximación al centro de masas del espacio de características.

3.3 En Resumen

Una breve introducción a las Máquinas de Soporte Vectorial se hace necesaria por razones de claridad en la exposición. El cambio en el uso de máquinas de aprendizaje SVMs en vez de las RNs expresadas en el Capítulo 2 viene motivado por la voluntad de atravesar el puente de unión desde el principio de regularización hacia el principio de inducción SRM. Se ha realizado un estudio sobre cuáles son estos puentes lanzados entre ambas visiones que han motivado este cambio de orientación en el estudio. Se trata sin duda de un campo de investigación abierto y altamente dinámico, pero los resultados obtenidos hasta el momento y el punto de vista que brinda el marco SRM parecen ofrecer mayores garantías teóricas en la búsqueda de una solución que generalice bien sobre problemas basados en un conjunto de datos empírico finito. En cualquier caso, aunque la integración de máquinas de aprendizaje sobre modelos mixtos se hace cada día más clara, ha sido descartado el uso de la inferencia bayesiana por su condicionamiento a la necesidad de definición de antemano por el usuario de una función de probabilidad.