Capítulo 1

Introducción

"Ayudadme a comprender lo que os digo y os lo explicaré"
Antonio Machado.

Para comenzar

Este primer Capítulo ofrece las definiciones y razonamientos que han motivado la perspectiva de trabajo del presente estudio. Significa por tanto un punto de referencia conveniente para entender los resultados desarrollados con posterioridad. Inicialmente se define y sitúa el marco de investigación general sobre el que se desarrolla el problema del aprendizaje supervisado. Se describe el modelo de un proceso de aprendizaje que permite enunciar el problema general de aprendizaje a partir de ejemplos como la minimización de un funcional de riesgo sobre la base de datos empíricos. Según sea la naturaleza de los valores en los datos de salida, el problema general da lugar al enunciado de diferentes tareas de aprendizaje, como pueden ser la estimación de regresiones, el reconocimiento de patrones o la regresión ordinal.

Una vez definido el problema general de aprendizaje supervisado y algunos casos particulares que se derivan, se hace necesario establecer un algoritmo que permita calcular una adecuada solución. Este algoritmo dependerá de una serie de factores que condicionaran su habilidad para generalizar. La elección del principio inductivo sobre el que se sienta la base del método de aprendizaje resultará crítica para conseguir el resultado deseado.

1.1 Problema de Aprendizaje a partir de Ejemplos

1.1.1 Planteamiento del Problema General

Tomando como punto de referencia los trabajos de [Vapnik, 1995], [Vapnik, 1998] y [Cherkassky and Mulier, 1998], se entenderá por aprendizaje a partir de ejemplos al proceso de estimar una dependencia desconocida entrada-salida de un sistema utilizando un número limitado de observaciones. El modelo general de un proceso de aprendizaje a partir de ejemplos se desarrolla sobre tres componentes:

- Un *generador* de vectores de entrada aleatorios, $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ — en el caso real multivariable —, producidos de forma no controlable por el usuario de forma independiente e idénticamente distribuida — en corto, i.i.d. — a partir de una densidad de probabilidad $p(\mathbf{x})$ fijada pero desconocida.
- Un *sistema* que retorna un valor de salida, $y \in \mathcal{Y}$, para todo vector de entrada $\mathbf{x}$, siguiendo una densidad condicional $p(y|\mathbf{x})$ también fijada pero desconocida. Este modelo general incluiría por ejemplo el caso de un sistema determinista que utilice alguna función $y = f(\mathbf{x}) + \epsilon$ donde ϵ es un ruido aleatorio de valor medio nulo, también denominado ruido blanco.
- Una *máquina de aprendizaje* — LM, del inglés *Learning Machine* — capaz de desarrollar un espacio de funciones

$$\mathcal{LM} = \left\{ f(\mathbf{x}, \omega) : \mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d, \omega \in \Omega \right\}, \tag{1.1}$$

  donde Ω es un conjunto de parámetros abstracto.

La máquina de aprendizaje observa los ℓ pares que constituyen el *conjunto de entrenamiento* conteniendo los vectores de entrada y la respuesta del sistema

$$\mathcal{T} = \{(\mathbf{x}_p, y_p)\}_{p=1}^{\ell} \subset \mathcal{X} \times \mathcal{Y} \sim P_{\mathcal{X}\mathcal{Y}}^{\ell}, \tag{1.2}$$

para construir a partir de ellos durante este período, denominado período de entrenamiento, algún operador que sirva de predictor de respuestas del sistema a entradas específicas producidas por el generador.

El problema general de aprendizaje deberá definirse como aquel de elegir entre el conjunto establecido de funciones $f(\mathbf{x}, \omega) \in \mathcal{LM}$ aquella que posea menor discrepancia con — "mejor aproxime" — la respuesta del sistema.

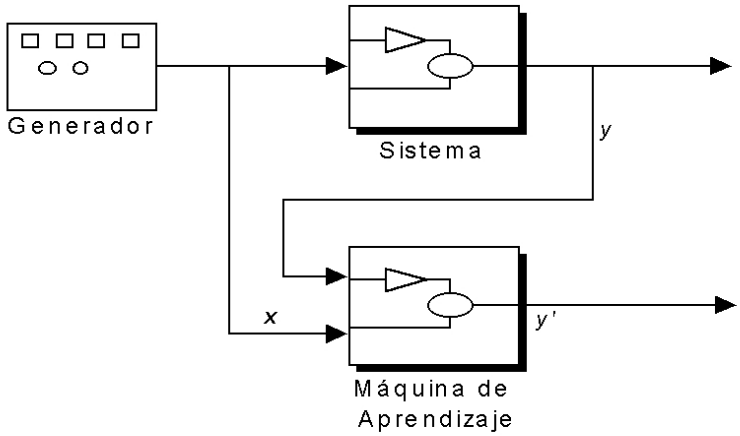

Figura 1.1: Modelo general de un proceso de aprendizaje a partir de ejemplos.

Definición 1.1. *Sea* $L(y, f(\mathbf{x}, \omega))$ *una medida de discrepancia o* función de coste *entre la respuesta del sistema,* y*, a una entrada,* $\mathbf{x}$*, y la respuesta producida por la máquina de aprendizaje,* $f(\mathbf{x}, \omega)$*. Sea el* funcional de riesgo $R(\omega)$ *la esperanza estadística de esta discrepancia*

$$R(\omega) = \int L(y, f(\mathbf{x}, \omega)) \, p(\mathbf{x}, y) \, , \tag{1.3}$$

entonces, se define el problema general de aprendizaje a partir de ejemplos — *en corto, PGAE — como aquel de elegir entre el conjunto establecido de funciones* $f(\mathbf{x}, \omega) \in \mathcal{LM}$ *aquella que minimice el funcional de riesgo,*

$$f\left(\mathbf{x}, \omega^{\mathcal{LM}}\right) = \arg\min_{\omega \in \Omega} R(\omega) \, , \tag{1.4}$$

cuando la densidad de probabilidad conjunta $p(\mathbf{x}, y)$ *es desconocida y la única información accesible está contenida en el conjunto de entrenamiento* $\mathcal{T}$*.*

Observando la definición, puede establecerse que la función elegida durante el proceso de aprendizaje debe ser seleccionada sobre tres principales restricciones:

1. Un amplio conjunto de funciones de aproximación, $\mathcal{LM}$, que definirá el *espacio de aproximación.*

2. Un número limitado de ejemplos, $\mathcal{T}$, denominado conjunto de entrenamiento.

3. Una medida de la discrepancia, $L\left(y, f\left(\mathbf{x}, \omega\right)\right)$, entre la respuesta del sistema y la respuesta de la máquina de aprendizaje.

Mientras que la segunda restricción viene determinada por el problema concreto a tratar, sobre la primera es necesaria la intervención del usuario puesto que depende en esencia del tipo de máquina de aprendizaje seleccionado. En cuanto a la tercera restricción, su uso en la definición teórica del funcional de riesgo no resulta en modo alguno dificultosa. Sin embargo, la imposibilidad de ser tratado de forma práctica debido al desconocimiento de la función densidad de probabilidad conjunta, obliga a su substitución por alguna otra medida que dependerá del principio inductivo seleccionado por el usuario y del conocimiento *a priori* que se desee insertar para conseguir la unicidad en la solución.

1.1.2 Casos Particulares del Problema General

El PGAE puede ser dividido en diferentes tipos según sea la naturaleza de las entradas y salidas tratadas en el modelo general. Puesto que las entradas han sido restringidas en la definición inicial al caso real multivariable, $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$, las posibles variaciones del problema general vendrán determinadas por la tipología de las variables de salida, $y \in \mathcal{Y}$, siendo los dos tipos más comunes de variable la numérica y la categórica.

Estimación de Regresiones

Se entenderá por *variable numérica* aquella cuyo valor pertenece a un conjunto sobre el que ha sido definida una relación de orden total, es decir, la relación de orden permite definir una distancia. El ejemplo más usual de variable numérica es el de las variables definidas sobre la recta real, $y \in \mathcal{Y} \subseteq \mathbb{R}$. En este caso, la ordenación sobre la recta permite definir una distancia, como por ejemplo la distancia euclídea.

Definición 1.2. *Se define el* problema general de estimación de una regresión *como el PGAE en el caso que el espacio de salida del sistema y de la máquina de aprendizaje sea un subconjunto de la recta real,* $y \in \mathcal{Y} \subseteq \mathbb{R}$.

La forma que adopta la función de coste para el caso de estimación de regresiones viene establecida por la siguiente afirmación,

Proposición 1.3. *Sean $y \in \mathcal{Y} \subseteq \mathbb{R}$ el tipo de respuesta de un sistema y $\mathcal{LM}$ un conjunto aproximación basado en funciones reales definido en (1.1). Sea la función de regresión definida como*

$$f^{opt}(\mathbf{x}) = \int y p(y|\mathbf{x}) \, , \tag{1.5}$$

entonces, si en la expresión del funcional de riesgo (1.3) es utilizada la función de coste

$$L(y, f(\mathbf{x}, \omega)) = (y - f(\mathbf{x}, \omega))^2 \, , \tag{1.6}$$

se puede afirmar que

- *Si* $f^{opt}(\mathbf{x}) \in \mathcal{LM} \Rightarrow \exists \omega^{\mathcal{LM}} \in \Omega,\ f\left(\mathbf{x}, \omega^{\mathcal{LM}}\right) = f^{opt}(\mathbf{x})$
- *Si* $f^{opt}(\mathbf{x}) \notin \mathcal{LM} \Rightarrow f\left(\mathbf{x}, \omega^{\mathcal{LM}}\right) = \underset{f \in \mathcal{LM}}{\arg\min} \sqrt{\int \left(f(\mathbf{x}, \omega) - f^{opt}(\mathbf{x})\right)^2 p(\mathbf{x})}$.

Por tanto, de forma teórica, siempre es posible hallar la función regresión solución, si ésta pertenece al conjunto de funciones que es capaz de implementar la máquina de aprendizaje o, en el caso que no pertenezca, la función del espacio $\mathcal{LM}$ que más se aproxima en norma 2. A este error en la búsqueda de la solución se le denomina *error de aproximación*, por ser consecuencia de la elección del espacio de aproximación $\mathcal{LM}$.

Se ha de hacer notar que la función de coste (1.6) continua estando integrada en el funcional de riesgo (1.3), que depende de una distribución de probabilidad desconocida que debe ser solucionada.

Reconocimiento de Patrones

Una *variable categórica* es aquella cuyo valor pertenece a un conjunto finito sobre el que no ha sido definido una relación de orden. Ejemplos de variables categóricas son aquellas que toman valor sobre conjuntos de elementos no numéricos — colores, marcas, tipos, ... — .

Definición 1.4. *Se define el* problema general de clasificación *o* de reconocimiento de patrones *como el PGAE en el caso que el espacio de salida del sistema y de la máquina de aprendizaje sea un conjunto cuyos elementos no están ordenados, $y \in \mathcal{Y} = \{\theta_1, \ldots, \theta_K\}$.*

Los elementos del conjunto de salida reciben el nombre de *etiquetas* definiendo la *clase* a la que puede ser asignado un elemento de entrada al sistema. Es muy habitual en la literatura hallar problemas de aprendizaje, denominados *dicotomías*, cuyo conjunto de salida categórico ha sido transformado en variables binarias numéricas, en general $\{0, 1\}$. La razón principal estriba en la facilidad de definir una función de coste que contabilice el número de errores de clasificación.

Sean $y \in \mathcal{Y} = \{0, 1\}$ la respuesta del sistema y $\mathcal{LM}$ un conjunto de aproximación basado en funciones indicador — funciones que sólo toman dos valores: cero y uno — . Es posible contar el número de errores de clasificación si en la expresión del funcional de riesgo (1.3) es utilizada la función de coste

$$L\left(y, f\left(\mathbf{x}, \omega\right)\right) = \begin{cases} 0 & \text{si } y = f\left(\mathbf{x}, \omega\right) \\ 1 & \text{si } y \neq f\left(\mathbf{x}, \omega\right) \end{cases} , \tag{1.7}$$

por lo que se puede establecer la siguiente definición.

Definición 1.5. *Se define el* problema de clasificación binaria *como el problema general de clasificación en el caso que el espacio de salida del sistema y de la máquina de aprendizaje sea el conjunto de etiquetas* $\mathcal{Y} = \{0, 1\}$ *y el funcional de riesgo (1.3) a minimizar tenga por función de coste la expresión (1.7).*

Regresión Ordinal

Además de los dos tipos generales de variables ya expresados, es posible definir un tipo intermedio de variables denominado ordinales que reúne características de las dos clases anteriores.

Una *variable ordinal* es aquella cuyo valor pertenece a un conjunto finito de etiquetas sobre el que ha sido definida una relación de orden u ordenación. Como ejemplo de variable ordinal sirva aquel de un conjunto de colores que han sido ordenados por su nivel de azul, o bien el de un conjunto de etiquetas definiendo las notas de los estudiantes de una asignatura. Nótese como las variables ordinales están fuertemente relacionadas con las variables borrosas o difusas y con las variables de tipo intervalar.

Definición 1.6. *Se define el* problema general de regresión ordinal *o* de ordenación *como el PGAE en el caso que el espacio de salida del sistema y de la máquina de aprendizaje sea un conjunto finito cuyos elementos poseen una ordenación,* $y \in \mathcal{Y} = \{\theta_1, \ldots, \theta_K\}$ *con* $\theta_K \succ_{\mathcal{Y}} \theta_{K\text{-}1} \succ_{\mathcal{Y}} \cdots \succ_{\mathcal{Y}} \theta_1$.

Se entenderá que la ordenación sobre las salidas permite establecer una ordenación sobre las entradas que sea de utilidad[1] . Por ejemplo, un producto comercial será preferido a otro en el caso que su salida *nivel de calidad* sea superior respecto al orden $\succ_{\mathcal{Y}}$.

Como en el caso del reconocimiento de patrones, lo más habitual cuando se trata con problemas de regresión ordinal es transformarlos en problemas de clasificación binaria, teniendo en cuenta durante la transformación la ordenación que poseen las etiquetas.

1.2 Algoritmo de Aprendizaje

1.2.1 Definición de Algoritmo de Aprendizaje

Un algoritmo de aprendizaje será aquel proceso capaz de dar respuesta al problema de aprendizaje a partir de ejemplos planteado. Continuando la definición de este problema, se sucede la siguiente definición, en congruencia con aquella planteada por los autores en [Vapnik, 1998] y [Cherkassky and Mulier, 1998].

Definición 1.7. *Se define un* algoritmo de aprendizaje a partir de ejemplos *como aquel proceso capaz de elegir una única función a partir del conjunto de entrenamiento dando respuesta al problema planteado de aprendizaje a partir de ejemplos.*

Un algoritmo de aprendizaje precisa seleccionar:

1. Un espacio de aproximación $\mathcal{LM}$ amplio y flexible.
2. Un *conocimiento a priori* que establezca una ordenación de las funciones de aproximación de acuerdo a alguna medida de su flexibilidad para adecuarse a los datos del conjunto de entrenamiento.
3. Un *principio inductivo* que determine en qué medida se combina el conocimiento *a priori* con el conjunto de entrenamiento disponible.
4. Un *método de aprendizaje* que constituya una implementación computacional constructiva de un principio inductivo para un espacio de aproximación dado.

[1] La ordenación total existe sobre las etiquetas, pero es sólo un orden parcial sobre las entradas.

Los dos elementos iniciales a seleccionar dependen generalmente de la elección que realice el usuario para el problema concreto que está trabajando, aunque la necesidad de traducir el conocimiento *a priori* en términos de la metodología algorítmica empleada restringen esta segunda elección y la estandarizan en gran manera.

Espacio de Aproximación $\mathcal{LM}$

Definición 1.8. *La* amplitud *de un espacio de aproximación se define como la capacidad para aproximar cualquier función continua, $f \in \mathcal{C}(\mathcal{X}, \mathcal{Y})$, con una precisión especificada cualquiera. Se dirá que un espacio de aproximación $\mathcal{LM}$ es* denso*, en $\mathcal{C}(\mathcal{X}, \mathcal{Y})$, si cumple la propiedad de aproximación universal.*

Tradicionalmente, han sido considerados buenos espacios de aproximación aquellos que son densos. Sin embargo, esta gran capacidad de aproximación provoca que el número de funciones solución que se ajustan al conjunto de entrenamiento sea muy elevado y de características muy diferentes. Por ejemplo, el uso de perceptrones multicapa — MLP, del inglés *Multi Layer Perceptron* — como máquinas de aprendizaje se ha justificado en numerosas ocasiones por su propiedad de aproximador universal [Hornik et al., 1989], aunque no existe un algoritmo de aprendizaje que permita hallar la solución al problema de manera única.

Definición 1.9. *La* flexibilidad *de un espacio de aproximación se define como la capacidad del espacio para estimar dependencias arbitrarias a partir de un conjunto de datos finito.*

Así pues, la flexibilidad es una propiedad que depende en gran medida del tipo de máquina de aprendizaje definitoria del espacio de aproximación y de su habilidad para trabajar sobre un conjunto finito de datos.

Conocimiento *a priori*

La necesidad de asumir cierto conocimiento de antemano sobre la forma del modelo buscado es esencial para conseguir la unicidad de la solución. Tal conocimiento será insertado en el algoritmo de aprendizaje en función del principio inductivo seleccionado. Algunos de los requerimientos más comunes suelen ser la exigencia de suavidad en la solución, la reducción en la talla de los pesos, la existencia de una determinada función de distribución de probabilidad conocida, la maximización del margen entre clases, ...

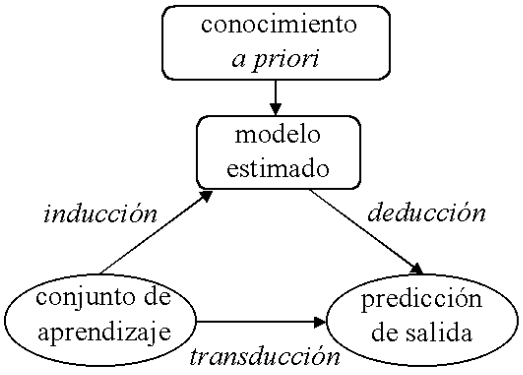

Figura 1.2: Procesos de inferencia: inducción-deducción y transducción.

1.2.2 Principios Inductivos

Observando el funcional de riesgo (1.3) a ser minimizado según la definición del PGAE, se deduce que la función desconocida debe ser estimada para todo posible valor del espacio de entrada $\mathcal{X}$, ya que la esperanza es tomada sobre alguna distribución de probabilidad desconocida del espacio completo. La función solución no sólo debe de tener un buen comportamiento sobre el conjunto de entrenamiento, sino que debe ser asegurado que posea la propiedad de *generalizar bien*, por lo que se hace necesario estimar un modelo general que permita predecir la salida a cualquier entrada del espacio $\mathcal{X}$, siguiendo el proceso de inferencia inductiva, o de inducción-deducción, que se puede observar en la Figura 1.2. En paralelo, la misma definición restringe la creación de nuestro algoritmo a la de un espacio aproximador predefinido — aunque tan amplio y flexible como se desee — , un conjunto de entrenamiento finito no controlable y una medida de discrepancia basada en una distribución de probabilidad desconocida. El principio inductivo deberá establecer *cómo* dar respuesta al problema de aprendizaje generalizando bien y cumpliendo las restricciones del planteamiento.

Minimización del Riesgo Empírico

El principio inductivo de minimización del riesgo empírico — ERM, del inglés *Empirical Risk Minimization* — es el más comúnmente utilizado en los procesos de aprendizaje clásico. Tomando como espacio de aproximación alguno que sea denso, queda eliminada la primera de las tres restricciones impuestas por la definición del PGAE. De las otras dos, puesto que el conjunto de entrenamiento no es controlable,

el principio ERM fija su atención en redefinir el funcional de riesgo basándose en el siguiente razonamiento:

> "Para obtener una buena generalización es suficiente con elegir los parámetros de la función aproximadora que aseguren el número mínimo de errores sobre el conjunto de entrenamiento".

Siguiendo esta aseveración, el principio inductivo ERM substituye la minimización del funcional de riesgo (1.3) con función densidad de probabilidad desconocida por el siguiente esquema:

1. El funcional de riesgo $R(\omega)$ es reemplazado por el denominado *funcional de riesgo empírico*
$$R_{emp}(\omega) = \frac{1}{\ell}\sum_{p=1}^{\ell} L\left(y_p, f\left(\mathbf{x}_p, \omega\right)\right) , \tag{1.8}$$
construido sobre el conjunto de entrenamiento $\mathcal{T}$. Generalmente la función de coste sigue la expresión (1.6) de la norma 2 al cuadrado.

2. Se aproxima la función (1.4), $f\left(\mathbf{x}, \omega^{\mathcal{LM}}\right)$, que minimiza el riesgo (1.3) por la función $f\left(\mathbf{x}, \omega^{\mathcal{T}}\right)$ que minimiza el riesgo (1.8)
$$f\left(\mathbf{x}, \omega^{\mathcal{T}}\right) = \arg\min_{\omega\in\Omega} R_{emp}(\omega) . \tag{1.9}$$

Definición 1.10 ([Vapnik, 1995]). *Se define el* error de generalización *cometido por un algoritmo de aprendizaje al solucionar un problema de aprendizaje como la cota de la suma del error de aproximación que se comete al elegir el espacio de aproximación más el error de estimación que provoca la finitud del conjunto de entrenamiento en el que se basa el proceso de aprendizaje.*

Siguiendo la definición, se dirá que la función solución estimada generaliza bien si comete un error de generalización pequeño. Nuevamente esta medida es puramente teórica e imposible de calcular, pero permite definir una base teórica sobre la que definir principios de inferencia que aseguren una buena generalización (Figura 1.3).

Inconvenientes del Principio Inductivo ERM Este procedimiento inductivo puede entenderse como aquel que determina la densidad de probabilidad desconocida

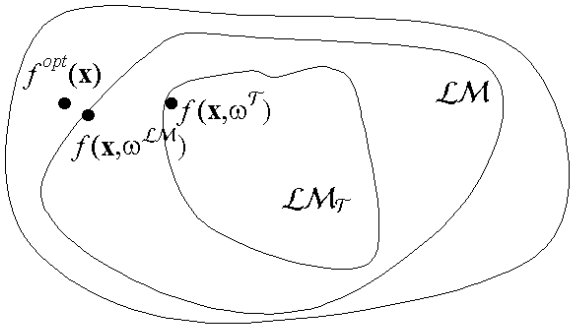

Figura 1.3: Visualización del error de aproximación, de estimación y de generalización.

y la establece *a priori* como

$$p\left(\mathbf{x},y\right)=\begin{cases}1 & \text{si } (\mathbf{x},y)=(\mathbf{x}_p,y_p)\\ 0 & \text{si } (\mathbf{x},y)\neq(\mathbf{x}_p,y_p)\end{cases}\quad, \tag{1.10}$$

por lo que es posible afirmar que para dar respuesta al problema de aprendizaje el principio inductivo ERM necesita establecer *a priori* la distribución de probabilidad conjunta.

El uso del principio ERM en procesos de aprendizaje clásicos, como por ejemplo máquinas de aprendizaje MLP, conduce a un proceso de interpolación sobre el conjunto de entrenamiento que conlleva el fenómeno conocido como sobrentrenamiento u 'overfitting', que habitualmente es solucionado mediante el uso de criterios de parada temprana del entrenamiento — en inglés, *early stopping rules* — o mediante decaimiento de pesos — en inglés, *weigth decay* — [Bishop, 1995]. Estas técnicas son una introducción de nueva información *a priori* durante el proceso de aprendizaje, que resulta del todo necesaria para asegurar la unicidad de la solución.

Por otra parte, en numerosos problemas extraídos de la vida cotidiana los datos de aprendizaje contienen ruido, por lo que no resultaría la mejor opción realizar un proceso de aprendizaje que conduzca en su estadio final a una interpolación de los datos empíricos facilitados.

Regularización o Penalización

El proceso de inducción de un modelo general a partir de un conjunto de entrenamiento es un problema mal situado en el sentido que no existe una única solución. La técnica de regularización [Tikhonov and Arsenin, 1977] asegura, bajo ciertas pequeñas restricciones sobre los espacios de trabajo, que si en vez del funcional de riesgo $R(\omega)$ se minimiza el denominado *funcional de riesgo regularizado*

$$R^{reg}(\omega) = R(\omega) + \lambda \cdot \phi(f) \,, \tag{1.11}$$

donde $\phi(f)$ es algún tipo de funcional y λ es una constante positiva escogida apropiadamente, entonces se obtiene una única solución sobre el espacio definido por el funcional $\phi(f)$.

Básicamente se trata de traducir las condiciones *a priori* que se quieren y/o se deben imponer para conseguir solución única en la forma de un funcional penalizador. Este método de inferencia inductiva, al ser combinado con el principio ERM, permite definir el *funcional de riesgo empírico regularizado* a ser minimizado para solucionar el problema de aprendizaje como

$$R^{reg}_{emp}(\omega) = R_{emp}(\omega) + \lambda \cdot \phi(f) \,, \tag{1.12}$$

demostrando que el razonamiento sobre el que se basa el método ERM de minimizar $R_{emp}(\omega)$ para asegurar una buena generalización no es 'auto-evidente'.

En el Capítulo 2 se verá más ampliamente el método de trabajo de la técnica de regularización. Por ahora tan solo apuntar que la teoría que da pie a esta técnica no trabaja con determinaciones *a priori* de la función densidad de probabilidad desconocida, sino que se desarrolla sobre la hipótesis de espacios de aproximación anidados del estilo

$$\mathcal{M}_c = \{f : \phi(f) \leq c\} \,, \quad c \geq 0 \,, \tag{1.13}$$

cumpliendo la condición de ser todos compactos.

Inconvenientes de la Técnica de Regularización La técnica de regularización fue desarrollada para solucionar ecuaciones integrales generadoras de problemas mal situados. Esto significa que la finalidad última de la máquina de aprendizaje es la de *identificar* el sistema del modelo de aprendizaje, es decir construir un operador que sea próximo al sistema. Por contra, el problema de aprendizaje, siguiendo la definición

que se ha realizado inicialmente, puede resolverse construyendo un operador capaz de *imitar* al sistema, es decir capaz de proveer para un generador fijado la mejor predicción de las salidas del sistema.

El proceso de la técnica de regularización de resolver un problema más complicado — identificar — que el que finalmente se quiere solucionar — imitar — conlleva que hasta el momento los resultados derivados de esta teoría sean válidos sólo de forma asintótica cuando el número de observaciones tiende a infinito. No existe forma alguna de evaluar la bondad de la solución si es utilizado un número finito de observaciones, requerimiento que ha sido expresado explícitamente en el enunciado de la definición del PGAE.

Se ha de reconocer sin embargo que el intento de identificar el sistema puede ser apropiado si el problema de aprendizaje se plantea por ejemplo desde la perspectiva de identificación de sistemas no lineales en Control Automático, en cuyo caso la técnica de regularización resulta adecuada [Johansen, 1996], [Johansen, 1997].

Inferencia Bayesiana

La inferencia bayesiana codifica información *a priori* adicional sobre las funciones de aproximación en forma de una *distribución de probabilidad a priori*, la probabilidad de que una función del espacio $\mathcal{LM}$ sea la auténtica función desconocida. De esta forma responde a la restricción sobre la definición del funcional de riesgo (1.3) asumiendo una cierta probabilidad sobre el modelo y añadiendo información con intención de obtener un único modelo predictivo.

Este tipo de inferencia está basado en la fórmula clásica de Bayes de actualización de probabilidades *a priori* utilizando la evidencia proporcionada por los datos

$$\mathrm{P}\left[\text{modelo}|\text{datos}\right] = \frac{\mathrm{P}\left[\text{datos}|\text{modelo}\right] \cdot \mathrm{P}\left[\text{modelo}\right]}{\mathrm{P}\left[\text{datos}\right]}, \tag{1.14}$$

donde P[modelo|datos] es la probabilidad *a posteriori* que se desea conocer y P[modelo] es la probabilidad *a priori* usada antes de que la máquina de aprendizaje observe los datos.

Inconvenientes de la Visión Bayesiana El gran inconveniente de la técnica bayesiana es su restricción a la necesidad de que el conjunto de aproximación $\mathcal{LM}$

coincida con el conjunto de problemas que la máquina tiene que resolver. Tal como se ejemplariza en [Vapnik, 1995], no tiene sentido aplicar inferencia bayesiana a un problema de aproximación por polinomios si la función regresión no es polinómica, puesto que la probabilidad *a priori* de que cualquier función de $\mathcal{LM}$ sea la función de regresión es igual a 0. El hecho de que la visión bayesiana funcione en situaciones generales tiene más que ver con el procedimiento humano-máquina frente al problema concreto que con la auténtica capacidad de la máquina construida sobre inferencia bayesiana.

Minimización del Riesgo Estructural

El principio inductivo de Minimización del Riesgo Estructural — SRM, del inglés *Structural Risk Minimization* — es un proceso de inferencia desarrollado sobre la Teoría del Aprendizaje Estadístico — SLT, del inglés *Statistical Learning Theory* [Vapnik, 1998]— específicamente para trabajar con problemas de aprendizaje a partir de un conjunto de entrenamiento pequeño.

A partir de la obtención de una cota sobre el funcional de riesgo $R(\omega)$ válida para cualquier conjunto dado de funciones $\mathcal{LM}$, se concluye que para asegurar su minimización, fijado el conjunto de entrenamiento, es necesario minimizar simultáneamente el funcional de riesgo empírico $R_{emp}(\omega)$ y la VC dimensión del conjunto de funciones $\mathcal{LM}$, $\Phi(h_k, \ell)$[2] , que es una medida de la *amplitud* del espacio de aproximación

$$R(\omega) \leq R_{emp}(\omega) + \Phi(h_k, \ell) = R_{srm}(\omega) \ . \tag{1.15}$$

Para entender mejor esta medida de la amplitud del espacio de aproximación, sirva la siguiente definición restringida para el caso de un problema de clasificación binaria.

Definición 1.11 (Vapnik-Chervonenkis). *La VC dimensión de un conjunto de funciones indicador $\mathcal{LM}$ es igual al número* h *máximo de vectores $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)$ que pueden ser separados en dos clases diferentes en todas las 2^h posibles maneras utilizando este conjunto de funciones $\mathcal{LM}$.*

Corolario 1.12 (Vapnik-Chervonenkis). *Si el espacio de aproximación $\mathcal{LM}$ es denso, entonces la VC dimensión es infinita, $h = \infty$, por lo que la minimización del riesgo empírico no puede asegurar, siguiendo la expresión (1.15), la minimización del riesgo funcional, $R(\omega) \leq R_{emp}(\omega) + \infty = \infty$.*

[2] La función $\Phi(h_k, \ell)$ es directamente proporcional a la VC dimensión e inversamente proporcional al número de elementos que componen el conjunto de entrenamiento $\mathcal{T}$.

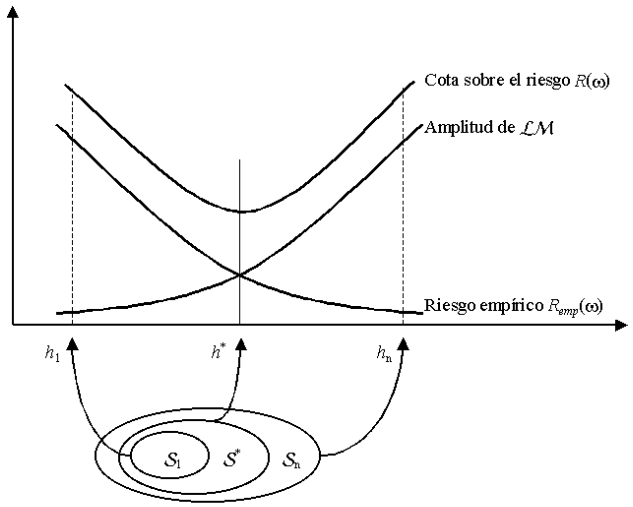

Figura 1.4: Visualización del error de aproximación, de estimación y de generalización.

Definición 1.13. *Se dirá que un espacio de aproximación $\mathcal{LM}$ es* capaz *si permite asegurar la minimización del funcional de riesgo estructural $R_{srm}(\omega)$.*

El funcional de riesgo estructural constituye una cota del funcional de riesgo, por lo que se trata de una condición suficiente para asegurar la minimización de $R(\omega)$, pero en ningún caso es una condición necesaria. Siempre es posible en casos de trabajo sobre problemas reales emplear un espacio de aproximación denso aunque se aplique el principio inductivo SRM y obtener buenos resultados, sin que la cota del riesgo funcional pueda asegurar de antemano estos buenos resultados.

Para conseguir que la VC dimensión sea una variable controlada, el principio SRM considera un esquema, semejante al principio de regularización, de espacios de aproximación anidados, del estilo

$$\mathcal{S}_k = \{f(\mathbf{x}, \omega) : \omega \in \Omega_k\} , \tag{1.16}$$

cumpliendo algunos requisitos teóricos, de forma que para un conjunto de entrenamiento $\mathcal{T}$ dado, el método SRM escoge el elemento $\mathcal{S}_k$ de la estructura anidada que permite obtener la cota más pequeña en (1.15), tal como se observa en la Figura 1.4. Si el diseño de la máquina de aprendizaje es demasiado complejo, la amplitud del espacio de aproximación será muy elevada, por lo que la reducción del riesgo empírico no asegura una buena generalización dando lugar al fenómeno conocido como *overfitting*. Para evitarlo, deben construirse máquinas sobre espacios con VC dimensión pequeña pero lo suficientemente capaces como para aproximar los datos de entrenamiento.

Inconvenientes del Principio SRM La principal desventaja del presente principio inductivo es su orientación hacia soluciones que imiten el sistema, desconsiderando la posibilidad de identificarlo. Aunque algún trabajo ha sido realizado en el sentido de solucionar problemas mal situados estocásticos con la ayuda de técnicas de regularización, los resultados tan solo son asintóticos por lo que no se adecúan completamente al caso de problemas de aprendizaje sobre un conjunto de datos empíricos finito.

1.2.3 Métodos de Aprendizaje

Un método de aprendizaje resulta de la implementación constructiva del principio inductivo elegido en el algoritmo de aprendizaje. Generalmente corresponde a un proceso de optimización de un cierto funcional de riesgo que ha sido determinado siguiendo el principio inductivo.

Para cada principio inductivo existen muchos métodos de aprendizaje que lo implementan correspondientes a los diferentes espacios de aproximación $\mathcal{LM}$ y a las diferentes técnicas de optimización. En la presente Sección se hará mención a algunos métodos generales que se utilizan sobre los principios inductivos de regularización y SRM.

Tres Métodos sobre el Principio de Regularización

Aunque el método inicial de minimizar el funcional de riesgo empírico regularizado, $R_{emp}^{reg}(\omega) = R_{emp}(\omega) + \lambda \cdot \phi(f)$, que substituye al funcional de riesgo teórico $R(\omega)$, es habitualmente solucionado en su formulación original, también podría ser expresado como

$$\text{minimizar} \quad R_{emp}(\omega) \quad \text{con} \quad \phi(f) \leq C\,, \tag{1.17}$$

es decir, aplicar el principio ERM mientras se mantiene la complejidad del modelo acotada mediante el término de regularización $\phi(f)$. Un tercer método podría plantearse si el problema inicial es desarrollado como

$$\text{minimizar} \quad \phi(f) \quad \text{con} \quad R_{emp}(\omega) \leq \delta\,. \tag{1.18}$$

Algorítmicamente los tres enunciados corresponden a problemas de optimización convexa que pueden ser solucionados eficientemente y son equivalentes [Smola, 1998].

La Dualidad del Principio SRM

Para solucionar el problema de minimización de la cota de riesgo estructural (1.15) asociado al principio SRM, es necesario hallar un resultado que evite el *overfitting* y que minimice el número de errores sobre $\mathcal{T}$. Los métodos de aprendizaje pueden realizar esta tarea siguiendo una de estas dos visiones:

1. Elegir una arquitectura apropiada para la máquina de aprendizaje que permita mantener la amplitud del espacio de aproximación acotada y minimizar el riesgo empírico
$$\text{minimizar} \quad R_{emp}(\omega) \quad \text{con} \quad \Phi(h_k, \ell) \leq C\,. \tag{1.19}$$
2. Mantener el valor del funcional de riesgo empírico fijado — por ejemplo igual a 0 — y minimizar la amplitud de $\mathcal{LM}$
$$\text{minimizar} \quad \Phi(h_k, \ell) \quad \text{con} \quad R_{emp}(\omega) \leq \delta\,. \tag{1.20}$$

Según se tenga en consideración una u otra expresión del problema es posible implementar dos tipos de máquinas de aprendizaje diferentes:

1. Redes Neuronales Artificiales — ANN, del inglés *Artificial Neural Networks* — tipo MLP y similares.
2. Máquinas de Soporte Vectorial — SVM, del inglés *Support Vector Machines* —.

Debe precisarse que la definición de ANN es muy amplia y puede llevar a equívocos ya que muchos autores consideran las SVMs como ANNs. Con la distinción anterior sólo se pretende resaltar que las SVMs están basadas sobre una realización del principio SRM diferente a aquella sobre la que se basan los MLPs y máquinas de aprendizaje similares, que habitualmente sí se han definido como ANNs.

1.3 En Resumen

En este primer Capítulo ha sido definido el problema general de aprendizaje a partir de un conjunto finito de datos empíricos. Si bien el espacio de entradas al sistema que

	$\mathcal{LM}$ usual	densidad de probabilidad	información *a priori*
ERM	denso	unitaria sobre los datos	ninguna
Regularización	denso	desconocida	funcional penalizador
Bayesiana	denso	establecida por usuario	probabilidad *a priori*
SRM	capaz / denso	desconocida	VC dimensión

Tabla 1.1: Sumario de características de los principios inductivos descritos.

se desea modelizar ha sido restringido al caso real multidimensional, la indefinición del espacio de salida permite considerar varios casos del problema general como son la estimación de regresiones, el reconocimiento de patrones y la regresión ordinal. En el caso de la clasificación se ha hecho notar que en numerosas ocasiones cuando el espacio de salida es multiclase el problema es reconducido hacia esquemas de clasificación binaria.

El deseo de hallar una respuesta al problema hace necesaria la definición de un algoritmo de aprendizaje que tenga en cuenta las restricciones que conlleva la definición del problema. Un punto crítico en la definición del algoritmo radica en la elección del principio inductivo sobre el que se construye el método de aprendizaje. En la Tabla 1.1 se sumarizan las principales propiedades sobre las que están basadas los principios inductivos descritos, de los cuales también han sido resaltados sus principales inconvenientes de aplicación.

Por último, debe ser elegido el método de aprendizaje que implementará el principio inductivo. A este respecto, han sido enunciadas algunas metodologías generales referentes a la técnica de regularización y a la aplicación de la minimización del riesgo estructural que más adelante serán utilizadas.