Capítulo 1

Introducción

> *"Learning" is making useful changes in the working of our minds.*
> Marvin Minsky, The society of mind.

Según el *Diccionario de la lengua española* (Real Academia Española 1992) aprender es

> "adquirir el conocimiento de alguna cosa por medio del estudio o de la experiencia."

En esta disertación nos vamos a concentrar en este segundo aspecto de la definición: el aprendizaje a partir de la experiencia. Más concretamente, nuestro interés se centrará en el *aprendizaje computacional* a partir de ejemplos.

Durante las últimas décadas, la denominada *inteligencia artificial* ha buscado sustituir a los seres humanos por ordenadores en la realización de las tareas más pesadas o que más tiempo nos ocupan. También se ha planteado, como objetivo más ambicioso, la posibilidad de reproducir los comportamientos característicos de la inteligencia humana. Sin embargo, la idea de que el cerebro humano no es más que un complejo ordenador sigue siendo una cuestión controvertida (Penrose

1992). Dada la extrema complejidad del cerebro, que contiene del orden de 10^{10} neuronas densamente interconectadas, no disponemos hoy en día de un modelo adecuado para describir su funcionamiento. Esto hace imposible el diseño de mecanismos que simulen globalmente la actividad inteligente, y nos obliga a plantearnos objetivos más modestos como que dichos mecanismos simulen comportamientos específicos o que realicen adecuadamente tareas restringidas. Aún así, esta tarea es difícil. Una posibilidad que reduciría la cantidad de información que debe incluirse a priori en el sistema es que el ordenador aprenda a reaccionar adecuadamente tras un período de entrenamiento. Idealmente, la máquina respondería tras el entrenamiento con un comportamiento semejante al de un ser humano con las mismas experiencias. Este proceso de adquisición de conocimientos a partir de la experiencia es lo que tradicionalmente se ha denominado *aprendizaje inductivo*, cuya historia revisamos rápidamente en la sección siguiente.

1.1 Aprendizaje inductivo

El procedimiento inductivo fue sometido a una demoledora crítica por parte de David Hume (Hume 1748:67)

> "Todas las inferencias realizadas a partir de la experiencia, por tanto, son efecto de la costumbre y no del razonamiento.",

cuyas consecuencias se extienden hasta el siglo XX. De hecho, Bertrand Russell llegó a afirmar a este respecto que Hume "representa la bancarrota de la racionalidad" (Russell 1946). El filósofo Karl Popper intentó justificar los métodos inductivos desde el punto de vista de la lógica. Según Popper (Popper 1972), es posible preferir una hipótesis a otras por lo que respecta a su verdad o falsedad basándose en justificaciones empíricas: dado que los hechos experimentales pueden refutar algunas de ellas "preferimos aquella cuya falsedad no haya sido demostrada". Sin embargo, desde este punto de vista, no existe ningún

motivo racional para elegir una entre las hipótesis no rechazadas, pues el hecho de que una hipótesis concuerde con los experimentos realizados, por muy grande que sea el número de éstos, no garantiza que el acuerdo se mantenga en los experimentos futuros. Dicho de otra forma, los experimentos no sirven para verificar una hipótesis, sólo para refutar algunas de ellas.

Para solventar este problema, el mismo Popper introdujo el concepto de simplicidad en el proceso de inducción:

> "Hemos de valorar más los enunciados sencillos que los menos sencillos, porque nos dicen más, porque su contenido empírico es mayor y porque se pueden contrastar mejor."

Aunque a primera vista resulta interesante, su definición de simplicidad como *grado de contrastabilidad* no resulta práctica, ni es lo suficientemente precisa como para ser aplicada con generalidad. A este respecto, puede leerse la crítica de Carl G. Hempel en Hempel (1966:77).

En este punto, el concepto de Gold (1967) de *identificación en el límite* proporcionó un criterio riguroso para la elección de unas hipótesis sobre otras: en determinadas circunstancias, algunos procedimientos de formulación de hipótesis (pero no todos) garantizan que la hipótesis correcta será la única hipótesis propuesta después de un conjunto suficientemente grande de observaciones (si bien, no es posible especificar qué debe entenderse por "suficientemente grande"). Veremos en la sección 1.3 que dado cualquier método que permita ordenar las hipótesis, el criterio "proponer la primera hipótesis compatible con los experimentos" permite la identificación de la hipótesis correcta. Este resultado sugiere una definición rigurosa (aunque bastante flexible) del concepto de simplicidad y proporciona un indudable apoyo lógico al procedimiento inductivo.

1.2 Inferencia gramatical

El proceso de aprender la gramática correcta para un lenguaje a partir de ejemplos es conocido con el nombre de *inferencia gramatical*. La teoría de las gramáticas generativas fue desarrollada en los años cincuenta y sesenta a partir de las ideas del filólogo americano Noam Chomsky (Chomsky 1956). Su pretensión de encontrar un formalismo matemático para describir los lenguajes naturales resultó menos exitosa de lo esperado en cuanto a su objetivo de permitir el diseño de programas que pudieran interpretar o traducir textos. En cambio, la aplicación de estas ideas ha resultado especialmente provechosa en el ámbito de la informática, sobre todo en el desarrollo de lenguajes de programación y compiladores (Aho & Ullman 1972) y en la teoría de la computación, especialmente en el aprendizaje computacional (Laird 1988) y en los métodos sintácticos de reconocimiento de patrones (Fu 1982). Dado que, utilizando una codificación adecuada, cualquier conjunto de ejemplos puede describirse como cadenas de símbolos, el problema del aprendizaje del conjunto se reduce al de aprender las reglas de generación de estas cadenas. De una forma más general, los ejemplos se descomponen en estructuras elementales llamadas *primitivas* que aparecen formando *patrones* según ciertas reglas. Estas reglas de generación de patrones constituyen la gramática del lenguaje. La gran ventaja de la formulación gramatical o sintáctica es que un número reducido de reglas es capaz de describir un conjunto virtualmente infinito de patrones. Para ello, es suficiente con que la gramática incluya recursividad en sus reglas.

Una vez conocida la gramática mediante el proceso de inferencia, cualquier tarea de clasificación queda reducido a un problema de análisis sintáctico, esto es: se trata únicamente de decidir si el patrón por clasificar pertenece al lenguaje definido mediante la gramática. Para esta tarea existen diversos algoritmos eficientes de análisis, como el de Cocke, Younger (1967) y Kasami o el de Earley (1970), por lo que en esta memoria nos concentraremos en el problema de inferir la

gramática correcta.

1.3 Identificación en el límite

En este contexto, entenderemos el aprendizaje como la adquisición de la capacidad para realizar con éxito una tarea. Una definición rigurosa de esta idea intuitiva fue formulada por Gold en 1967 (Gold 1967), mediante el criterio de *identificación en el límite*. Según este criterio, se puede aprender un concepto si existe un procedimiento que garantiza que durante el proceso de aprendizaje sólo se producirá un número finito de errores.

De forma más precisa, dado un dominio Ω, llamaremos *lenguaje* o *concepto* a cualquier subconjunto $L \subset \Omega$, *ejemplo* de L a cualquier elemento $x \in L$ y *muestra positiva* a una secuencia infinita $S = \{x_1, x_2, ...\}$ de ejemplos de L, no necesariamente distintos. La *muestra finita* S_n está formada por los n primeros elementos de S. Diremos que el lenguaje L es identificable en el límite si existe un procedimiento $A(S, n)$ que, dada una muestra S del lenguaje L:

- para cada número natural $n \in \mathbb{N}$, propone como hipótesis para L un lenguaje $h_n = A(S, n)$;
- además, h_n converge a L en el sentido de que $h_n = L$ excepto para un número finito de valores de n, es decir, existe un valor n_0 tal que $h_n = L$ siempre que $n \geq n_0$.

Una clase de lenguajes F es identificable en el límite si todos los lenguajes de la clase son identificables en el límite mediante un mismo procedimiento A.

Gold también demostró que muchas clases importantes de lenguajes (entre ellos los lenguajes racionales que estudiaremos más adelante) no pueden ser identificados en el límite a partir de ejemplos. En muchos casos, una muestra positiva no es suficiente para identificar el lenguaje, esencialmente debido al problema de la generalización excesiva. La *generalización excesiva* se produce cuando se formula una

hipótesis que es más general que la correcta. Por ejemplo, h_2 generaliza a h_1 si $h_1 \subset h_2$. Si sólo disponemos de ejemplos, no siempre es posible elegir un procedimiento que identifique en cualquier caso la hipótesis correcta. Por ejemplo, si la clase F contiene el lenguaje Ω y además todos los lenguajes finitos $L \subset \Omega$, no existe un criterio para elegir entre Ω y L_n (siendo L_n el lenguaje finito formado exactamente por los ejemplos de S_n) que garantice la identificación en el límite.

Sin embargo, existen vías alternativas que evitan estas dificultades. Una de ellas es la utilización de muestras completas. Una *muestra completa* incluye no sólo los ejemplos del lenguaje sino también los contraejemplos, de forma que cada elemento aparece clasificado como perteneciente o no al concepto. De esta forma, $S = \{(x_1, \mu_1), (x_2, \mu_2)...\}$ donde $\mu_i = 1$ indica que x_i es un *ejemplo* ($x_i \in L$) y $\mu_i = 0$ indica que es un *contraejemplo* ($x_i \notin L$). Análogamente, $S_n = (x_1, \mu_1), ..., (x_n, \mu_n)$. La información contenida en la muestra completa S es suficiente para descartar todas las hipótesis que son demasiado generales.

Vamos a ver como un método enumerativo es suficiente para identificar un lenguaje que pertenece a un clase F recursivamente numerable de lenguajes[1] utilizando, para ello, muestras completas. Supongamos que el conjunto de hipótesis es $\{h_0, h_1, h_2, ...\}$ y que $h_r = L$ es la hipótesis correcta. Diremos que h_k es compatible con S_n si h_k contiene a todos los ejemplos de S_n y ninguno de sus contraejemplos o ejemplos negativos. El procedimiento enumerativo para identificar la hipótesis correcta consiste en elegir siempre como hipótesis aquella h_k con subíndice menor que es compatible con S_n.

Es evidente que de esta forma nunca se propondrá como hipótesis h_j si $j > r$. De hecho, no es posible que h_r sea rechazada, dado que nunca puede aparecer un ejemplo o contraejemplo incompatible con h_r. Por otro lado una hipótesis rechazada por S_n también lo será por todas las muestras finitas posteriores S_m tales que $m > n$. Dado que

[1] A nuestros efectos, C es un conjunto recursivamente numerable significa que existe un procedimiento algorítmico para asignar a cada $n \in \mathbb{N}$ uno de los elementos de C, de forma que dicho procedimiento es exhaustivo (suprayectivo) en C.

sólo existe un número finito de hipótesis h_i tales que $i < r$, resulta entonces evidente que no puede cambiarse de hipótesis más de r veces. Sólo resta, por tanto, justificar que todas las hipótesis h_i tales que $i < r$ acaban siendo descartadas para que la identificación en el límite siempre se alcance.

En efecto, si h_r no es compatible con h_i es porque existe al menos algún elemento en la diferencia simétrica de los dos conjuntos $h_r \oplus h_i$. Es decir, existe $x \in h_r$ tal que $x \notin h_i$ o bien, $x \in h_i$ tal que $x \notin h_r$. Si asumimos que x aparece en S, es decir que existe al menos un $m \in \mathbb{N}$ tal que $x = x_m$, entonces la hipótesis h_i es descartada por todas las S_n tales que $n \geq m$. En consecuencia, dada la muestra S, excepto para un número finito de valores de n, h_r es la primera hipótesis compatible con S_n. Nótese que, aunque queda garantizada la identificación en el límite, no disponemos de un criterio que nos permita afirmar cuándo se ha producido ésta. Además, ha sido necesario asumir que cualquier muestra contiene todos los ejemplos y contraejemplos necesarios para la identificación.

En la práctica los contraejemplos no son fáciles de conseguir, al menos en la cantidad deseada. Por ejemplo, en una tarea de reconocimiento de caracteres manuscritos podemos recoger grandes muestras de dígitos 0 y suponer que todas las formas de ceros acabarán siendo recogidas en la muestra. Sin embargo, aunque los ejemplos de otros dígitos pueden ser tomados como contraejemplos de ceros, es evidente que estos no son representativos de la clase complementaria (no todos los grafos que no representan un cero son otro dígito), por lo que el aprendizaje mostrará sistemáticamente una tendencia a la generalización excesiva.

La utilización de muestras completas puede ser evitada si se dispone de información adicional acerca del orden de presentación de los elementos de la muestra. Por ejemplo, si Ω está formado por cadenas de símbolos, es suficiente con que los elementos de la muestra positiva S aparezcan ordenados según su longitud, de forma que las cadenas que siguen a una dada sean todas de longitud mayor o igual que la

de ésta. Evidentemente, esto equivale a disponer de todos los contraejemplos. En efecto, dada $w \in S_n$, toda cadena $x \in \Omega$ de longitud menor que la de w tal que $x \notin S_n$ es necesariamente un contraejemplo del lenguaje. Otra forma de afrontar el problema de la generalización excesiva es asumir que los ejemplos han sido generados de acuerdo con una distribución probabilística preestablecida aunque desconocida. Esta opción parece corresponderse mejor con las situaciones reales. De hecho, en muchas aplicaciones como reconocimiento del habla, modelización del lenguaje natural etc, el proceso de aprendizaje involucra ejemplos ruidosos o aleatorios. La identificación de este tipo de lenguajes es discutida en la sección siguiente.

1.4 Identificación de lenguajes estocásticos

Angluin (1988) ha demostrado que es posible identificar en el límite un lenguaje a partir de muestras generadas aleatoriamente mediante distribuciones de probabilidad bastante generales, aunque sin llegar a proponer un método eficiente para ello. En cierto sentido, la regularidad estadística que presentan las muestras aleatorias es capaz de compensar la falta de datos negativos. Sin embargo, la suposición de que existe una fuente aleatoria de ejemplos no incrementa el conjunto de clases de lenguajes identificables.

Una *muestra aleatoria* del lenguaje L es una secuencia de ejemplos, $S = \{x_1, x_2, ...\}$, en la que los elementos de L aparecen repetidamente siguiendo una distribución de probabilidad $p(x)$ cuyo *soporte* coincide con L, es decir: $L = \{x \in \Omega : p(x) > 0\}$. Nótese que la muestra aleatoria no contiene información sobre los contraejemplos del lenguaje. La adición de una estructura probabilística al lenguaje no añade potencia a los métodos de identificación en el límite. Es decir, si una clase de lenguajes F es identificable en el límite utilizando muestras aleatorias, F es también identificable en el límite utilizando muestras completas (Angluin 1988).

Por otra parte, una clase recursivamente numerable de distribu-

ciones uniformemente aproximadamente computables es identificable en el límite (Angluin 1988). La distribución $p(x)$ es *uniformemente aproximadamente computable* si existe una función racional computable[2] $f : \Omega \times \mathbb{R} \to \mathbb{Q}$ que proporciona un número racional $f(x, \epsilon) \in \mathbb{Q}$ cuyo valor está como mucho a una distancia ϵ de $p(x)$. En particular, las distribuciones de probabilidad computables y racionales (es decir, aquellas que sólo toman valores en $\mathbb{Q}$), son identificables en el límite. Aunque en general los valores de las probabilidades en una distribución arbitraria pueden ser números reales, la distinción entre números reales y racionales es sólo relevante desde el punto de vista teórico, dado que en un ordenador todos ellos son representados mediante un número finito de bits. La cuestión de la identificación en el límite de probabilidades racionales será discutida con más detalle en la sección 3.6

1.5 Aprendizaje PAC

El criterio de aprendizaje formulado por Valiant (1984) es distinto del de Gold (1967), y se conoce con el nombre de *aprendizaje probablemente aproximadamente correcto* o, de forma abreviada, *aprendizaje PAC*. En este criterio, existe una distribución de probabilidad desconocida para los ejemplos y el algoritmo de aprendizaje extrae ejemplos al azar y trata de construir una hipótesis que no sea demasiado diferente del lenguaje correcto con probabilidad elevada.

En principio, una medida de la semejanza entre el lenguaje L y la hipótesis h podría definirse introduciendo una tasa de error, como el cociente del tamaño de la diferencia simétrica $L \oplus h$ y y el tamaño de L. Sin embargo, esta medida no es útil si L es un lenguaje infinito. Si disponemos de una distribución de probabilidad $p(x)$ para los elementos del dominio Ω, la probabilidad de las cadenas mal clasificadas

[2] La función f es computable si existe un algoritmo que proporciona el resultado de la función para cualquier entrada de los argumentos. Para una definición más rigurosa puede consultarse, por ejemplo, el libro de Hopcroft y Ullman (1979).

$p(h \oplus L)$ es una medida de la semejanza entre la hipótesis h y el lenguaje L, que denotaremos como $\mathrm{err}_L(h) = p(h \oplus L)$. Además, $p(x)$ nos permite generar muestras aleatorias finitas S_n de tamaño n arbitrario.

Todo lo anterior nos permite definir de forma más rigurosa un nuevo criterio de aprendizaje. Dado un dominio Ω y una clase F de lenguajes sobre Ω, F es aprendible (PAC) si existe un procedimiento A tal que para cualquier distribución de probabilidad $p(x)$ para los elementos de Ω y cualquier $L \in F$:

- toma $\epsilon > 0$ y $1 \geq \alpha > 0$ como entrada;
- puede llamar a un procedimiento $B(n)$ para generar S_n, una muestra aleatoria de L de tamaño n;
- proporciona un concepto $h = A(\epsilon, \alpha, S)$ tal que $\mathrm{err}_L(h) \leq \epsilon$ con probabilidad mayor que $1 - \alpha$.

Si el procedimiento A requiere un tiempo polinómico en términos de α y ϵ, F es aprendible *eficientemente*. Obviamente, para que esto se produzca A debe elegir también un valor para n polinómico en función de α y ϵ.

Este criterio es muy estricto, en el sentido de que no se realiza ninguna suposición sobre la distribución de probabilidad $p(x)$, ni tan siquiera que ésta sea computable. En particular, la posibilidad de aprender (PAC) una familia de lenguajes F viene dada por la dimensión de Vapnik-Chervonenkis de F, que es el tamaño del mayor conjunto $C \subset \Omega$ desmenuzado ("shattered") por F. Un conjunto C es *desmenuzado* por F si para cualquier subconjunto $X \subset C$ existe un lenguaje $L \in F$ tal que $F \cap C = X$. Es decir, existen suficientes lenguajes en F como para elegir uno que coincida (en el dominio propio de C) con cualquier subconjunto de C. De alguna forma, la dimensión de Vapnik-Chervonenkis mide la complejidad, en el sentido de capacidad expresiva, de la familia F.

Por ejemplo, si para algún elemento $x \in \Omega$ existe un lenguaje de F que contiene a x y otro que no lo contiene, la dimensión de Vapnik-

Chervonenkis de F es mayor o igual que uno. Para que la dimensión de F sea mayor o igual que 2, debe existir un par de elementos $x_1, x_2 \in \Omega$ y cuatro lenguajes $L_{00}, L_{01}, L_{10}, L_{11} \in F$ tales que $x_i \in L_{j_1 j_2}$ si y sólo si $j_i = 1$. De forma análoga se puede razonar para las condiciones con dimensión n.

Como veremos más adelante, los lenguajes racionales o regulares incluyen a todos los lenguajes finitos, y por tanto, su dimensión de Vapnik-Chervonenkis es mayor que cualquier valor n que se elija, o lo que es lo mismo, es infinita. Resulta que sólo las familias con dimensión de Vapnik-Chervonenkis finita son aprendibles (PAC) polinómicamente (Anthony & Biggs 1992), por lo que en adelante, utilizaremos el criterio de Gold para caracterizar el éxito en el aprendizaje.