Aprendizaje automático a partir de cuerpos de datos ralos: un enfoque basado en la inferencia inductiva

Sergio Gastón Romano (postulante) Santiago Figueira (director) Mariano Sigman (codirector)

1. Objetivos

En las últimas dos décadas, distintas técnicas de ingeniería reversa del aprendizaje en humanos han influenciado el éxito de los algoritmos de aprendizaje automático [19]. Sin embargo, pese a que aprender a partir de pocos datos es una capacidad cotidiana de la mente humana, las técnicas actuales de aprendizaje automático no dan buenas garantías al respecto [1, 25, 8].

Las explicaciones sobre esta capacidad humana de poder ir más allá de los datos observados para realizar buenas generalizaciones a partir de pocos datos proponen la existencia de otra fuente de conocimiento que permita generar y delimitar las hipótesis a analizar [19, 14], lo que en aprendizaje automático se conoce como *sesgo inductivo*.

En los últimos años, modelos computacionales de la cognición humana han desarrollado la idea de que la habilidad ubicua del ser humano para hacer inferencia (precisa) sobre cuerpos ralos de datos se basa en el uso de modelos *bayesianos*, donde el conocimiento se representa en espacios adecuadamente estructurados sobre los que se aplican reglas de inferencia bayesiana. Muchas veces es ésta una estructura jerárquica que permite representar a un mismo tiempo distintos niveles de abstracción de los datos en cuestión. En varios dominios del conocimiento, una forma eficiente de representar el conocimiento es en términos de un lenguaje simbólico específico, cuya gramática aparece como uno de los niveles en la estructura jerárquica que restringe el espacio de hipótesis a analizar en los niveles inferiores durante el proceso de inferencia [11, 24].

Por otro lado, en 1960 Ray Solomonoff desarrolla la teoría de inferencia inductiva para dar una respuesta teórica desde las ciencias de la computación al problema de inferencia sobre cuerpos de datos. Su trabajo puede entenderse como una formalización del principio de Occam, según el cual, la mejor teoría —entre todas las teorías que explican los datos— es la más *simple* [4]. Allí, Solomonoff utiliza como explicación *más simple* a la descripción efectiva *más corta* de un objeto y propone una distribución universal que permite predecir cualquier secuencia producida por un proceso computable con el mínimo absoluto de datos necesarios. Sin embargo, el hecho de utilizar la descripción efectiva más corta, o sea, su complejidad de Kolmogorov, convierte a dicha distribución en un problema no computable [21].

En este trabajo nos proponemos investigar la intersección de los modelos *bayesianos* con la teoría de inferencia inductiva de Solomonoff para desarrollar un modelo que permita realizar inferencia a partir de cuerpos ralos de datos. Estos modelos representarán al conocimiento sobre un lenguaje simbólico estructurado —como los utilizados en algunos modelos bayesianos— sobre el que se pueda inducir una función de complejidad computable y obtener así un algoritmo de predicción computable basado en la teoría de inferencia inductiva. En §2 veremos los primeros intentos en realizar este acercamiento entre ambas ideas y, en particular, proponemos empezar nuestra investigación trabajando con el lenguaje de geometría que explicamos en §2.1 ya que sus características lo hacen interesante tanto desde el punto de vista de la inferencia inductiva como desde el modelado bayesiano.

Como **objetivo general** pretendemos estudiar el proceso de inferencia a partir de cuerpos ralos de datos y desarrollar nuevos algoritmos de aprendizaje automático a partir de estas ideas. Los **objetivos específicos** son:

- Ajustar y validar el lenguaje de geometría, $\mathcal{LG}$, propuesto a través del estudio de tareas cognitivas en el dominio de intuiciones geométricas por parte de las personas.
- Obtener, ajustar y validar experimentalmente un modelo de predicción basado en la teoría de inferencia inductiva a partir de la complejidad de Kolmogorov inducida por $\mathcal{LG}$.
- Proponer y validar experimentalmente otros lenguajes formales que sirvan para adquirir conocimiento abstracto en diferentes dominios y tareas cognitivas realizadas por individuos.

- Por cada lenguaje, obtener un modelo de predicción del comportamiento en el aprendizaje con escasez de datos basado en la teoría de inferencia inductiva y validarlo experimentalmente.
- A partir de los estudios anteriores, diseñar algoritmos de aprendizaje automático que —inspirados en el comportamiento humano— permitan aprender múltiples estructuras de representación de los datos o que permitan modificar la estructura actual a partir de la adquisición de nuevos datos.
- Estudiar y comparar la eficacia de estos nuevos algoritmos de aprendizaje automático con otros algoritmos conocidos en escenarios con escasez de datos de entrenamiento.

2. Antecedentes

Un problema fundamental en aprendizaje automático es el compromiso entre el sesgo y la varianza. Por un lado, uno quiere minimizar el sesgo (cuán preciso es el modelo para nuevos datos de entrenamiento) y por otro uno quiere minimizar la varianza (cuán ajustado está nuestro modelo con respecto a los datos actuales de entrenamiento). Cuando los datos son escasos, este problema es todavía más relevante, ya que se vuelve mucho más probable el riesgo de sobreajustar el modelo y perder generalización.

De aquí, la importancia de contar con grandes cantidades de datos de entrenamiento se ha vuelto fundamental para los algoritmos actuales de aprendizaje automático, a tal punto que existe un aforismo en el área que dice: *no gana aquel con el mejor algoritmo, sino aquel que más datos tenga* [1]. Por el contrario, hemos mencionado que la mente humana sí posee la capacidad de aprender y encontrar eficientemente modelos lo suficientemente genéricos a partir de pocos datos [3, 9]. Además, los algoritmos de aprendizaje automático no supervisados actuales, como las técnicas de clustering o el análisis de componentes principales (*PCA* en inglés) [2, 19], encuentran estructuras en los datos observados pero asumen en ellos una única estructura fija [20]. Sin embargo, esto también difiere de la manera de aprender que tenemos los seres humanos, pues somos capaces de adaptar la estructura de representación del conocimiento con el paso del tiempo y luego de la adquisición de nuevos datos [13].

Hemos mencionado en §1 que en el escenario de datos ralos es necesaria la introducción de alguna restricción adicional sobre el espacio de hipótesis, llamada *sesgo inductivo* [15]. Un ejemplo clásico de sesgo inductivo es el principio de Occam [4], según el cual entre todas las teorías que explican los datos, la mejor es la más *simple*. Inspirado en este principio, Solomonoff crea en 1960 la teoría de la inferencia inductiva [21], donde utiliza como explicación *más simple* a la descripción efectiva *más corta* de un objeto. Su enfoque está íntimamente relacionado con la teoría algorítmica de la información, que permite clasificar a las cadenas finitas σ sobre un alfabeto finito Σ. Esta clasificación se realiza a través de la complejidad de Kolmogorov, que mide el tamaño de los programas p para una máquina de Turing M. La propiedad que interesa es la longitud del programa más corto que, ejecutado en M, genera la cadena σ en cuestión. De entre todos los programas que generan σ cuando son ejecutados en la máquina M, nos quedamos con aquellos que tengan la menor longitud; esa longitud es la complejidad de Kolmogorov de σ relativa a M. Más formalmente,

$$K_M(\sigma) := \begin{cases} \text{mín}\{|p| : M(p) = \sigma\} & \text{si } \sigma \text{ está en la imagen de } M \\ \infty & \text{si no,} \end{cases}$$

donde $|p|$ representa la longitud del programa p.

La complejidad de Kolmogorov, en sus distintas variantes, tiene importantes aplicaciones teóricas. Por ejemplo, cuando tomamos una máquina universal U, la función K_U, se convierte en una herramienta fundamental para definir matemáticamente una noción de aleatoriedad para secuencias infinitas [16, 7]. También es K_U la función que clasifica a las cadenas de acuerdo a su "simpleza" en el marco de la teoría de la inferencia inductiva de Solomonoff. A efectos prácticos, la principal desventaja de K_U es que resulta no computable, esto es, no existe ningún algoritmo que dada una cadena σ calcule $K_U(\sigma)$.

En la tesis de grado del postulante [17] y en una publicación posterior [18] se contempla un lenguaje simbólico, L (en el fondo, una máquina de Turing), lo suficientemente poderoso como para describir adecuadamente a todas las secuencias sobre un cierto alfabeto Σ (o sea, la imagen de L es Σ^*), pero lo suficientemente débil como para inducir una complejidad de Kolmogorov K_L computable (de hecho, en

tiempo polinomial). La idea es reemplazar K_U por K_L en la teoría de la inferencia inductiva y obtener así un modelo de predicción basado en esta teoría. Este modelo teórico es luego puesto a prueba exitosamente en algunos aspectos de la cognición humana: se prueba la eficacia de este lenguaje para encontrar regularidades en cadenas producidas por individuos, tratando de generar cadenas aleatorias para, por un lado, dar argumentos a favor de la naturaleza algorítmica del pensamiento humano y, por otro, realizar inferencias sobre el algoritmo y el lenguaje que describe los patrones del pensamiento humano.

La propuesta de este plan es primero profundizar los intentos de modelado de conocimiento de alto nivel en tareas cognitivas con modelos bayesianos como los realizados en distintos estudios sobre aleatoriedad [12], adquisición del lenguaje [5], predicción de eventos diarios [10], detección de similitudes [23], pero utilizando las técnicas descriptas anteriormente para realizar inferencia sobre los datos, donde a partir del lenguaje definido para estructurar el conocimiento, se induce una complejidad de Kolmogorov que es utilizada en un modelo de predicción basado en la teoría de inferencia inductiva. Particularmente, nos interesa trabajar con lenguajes más complejos y expresivos que el analizado en [18], como es el caso del lenguaje de geometría que presentamos a continuación.

2.1. Un lenguaje de geometría

Filósofos desde Platón a Descartes o Kant han argumentado que los conceptos desarrollados en la geometría resultan naturales para la mente humana, incluso en personas que carecen de toda instrucción matemática. Estudios recientes de cognición en niños o tribus sin ninguna formación escolar clásica muestran que las personas manejan intuiciones geométricas como la de paralelismos, ángulos agudos y simetría, entre otras [6, 22].

Nuestra propuesta es entonces definir un lenguaje simple, que denominamos $\mathcal{LG}$, para representar este conocimiento geométrico. $\mathcal{LG}$ permite describir secuencias de puntos sobre el alfabeto $\{0,\ldots,7\}$ dispuestos en circulo, como muestra la Figura 1. El lenguaje permite describir los puntos secuencialmente con instrucciones que indican cómo pasar al siguiente punto a partir del punto actual (siguiente inmediato en sentido horario o anti-horario, siguiente del siguiente inmediato, etc.), reflejar puntos de acuerdo a los ejes A, B, V y H que se muestran en la Figura 1 y ciertas estructuras simples para formar ciclos. Definimos a continuación la sintaxis y semántica formal de nuestro lenguaje de geometría.

Sintaxis. Definimos las instrucciones atómicas $At = \{+0, +1, +2, +3, -1, -2, -3, P, A, B, H, V\}$. La gramática del lenguaje se define como sigue. Un *programa atómico* $[i]$ es un programa que consiste de exactamente una instrucción atómica $i \in At$. Las reglas de formación de programas está dada por: 1) todo programa atómico es un programa; 2) si P y Q son programas entonces PQ (concatenación de P y Q) es un programa; 3) si P es un programa y $s \in \mathbb{N}$, entonces $[P]^s$ es un programa; 4) Si P es un programa, $s \in \mathbb{N}$ y $i \in At$, entonces $[P^s\{i\}]$ es un programa; y 5) Nada más es un programa.

Semántica. Un programa P se ejecuta con entrada $n \in \{0,\ldots,7\}$, y su salida se es una cadena formada por los símbolos $\{0,\ldots,7\}$ representa por $P(n)$. La semántica de las instrucciones atómicas se define en la Figura 2.

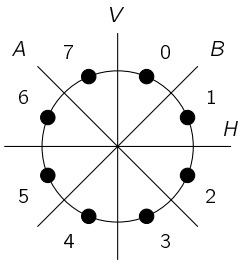

Figura 1. Círculo de puntos y reflexiones

- $[+0]$: identidad, i.e, $[+0](n) \stackrel{\text{def}}{=} n$
- $[+1]$: suma 1 mód. 8, i.e. $[+1](n) \stackrel{\text{def}}{=} n{+}1 \mod 8$
- $[-1]$: resta 1 mód. 8, i.e. $[-1](n) \stackrel{\text{def}}{=} n{-}1 \mod 8$
- $[+2]$: suma 2 mód. 8, i.e. $[+2](n) \stackrel{\text{def}}{=} n{+}2 \mod 8$
- $[-2]$: resta 2 mód. 8, i.e. $[-2](n) \stackrel{\text{def}}{=} n{-}2 \mod 8$
- $[+3]$: suma 3 mód. 8, i.e. $[+3](n) \stackrel{\text{def}}{=} n{+}3 \mod 8$
- $[-3]$: resta 3 mód. 8, i.e. $[-3](n) \stackrel{\text{def}}{=} n{-}3 \mod 8$
- $[H]$: H-reflexión, i.e. $[H](n) \stackrel{\text{def}}{=} 3-n \mod 8$
- $[V]$: V-reflexión, i.e. $[V](n) \stackrel{\text{def}}{=} 7-n \mod 8$
- $[A]$: A-reflexión, i.e. $[A](n) \stackrel{\text{def}}{=} 5-n \mod 8$
- $[B]$: B-reflexión, i.e. $[B](n) \stackrel{\text{def}}{=} 1-n \mod 8$
- $[P]$: opuesto, i.e. $[P](n) \stackrel{\text{def}}{=} 4+n \mod 8$

Figura 2. Semántica de instrucciones At

- $PQ(n) \stackrel{\text{def}}{=} P(n)\, Q(n')$, donde n' es el último símbolo de $P(n)$
- $[P^s](n) \stackrel{\text{def}}{=} \underbrace{P \ldots P}_{s}(n)$.
- $[P^s\{i\}](n) \stackrel{\text{def}}{=}$
  $P(n)P([i](n))P([i,i](n)) \ldots P[\underbrace{i \ldots, i}_{s-1}](n)$.

Figura 3. Semántica de programas

- Si P es atómico, $|P| = 2$
- $|PQ| = |P| + |Q|$
- $|[P^s]| = |P| + \lceil \log_2 s \rceil$
- $|[P^s\{i\}]| = |P| + \lceil \log_2 s \rceil + |[i]|$

Figura 4. *Tamaño* de un programa

La semántica formal de los programas está definida en la Figura 3. La concatenación tiene una semántica natural. Por ejemplo, $[+1,+1][+2,+2](0) = [+1,+1](0)\ [+2,+2](2) = 1246$. El programa $[P^s]$ representa una repetición simple. Por ejemplo, $[+1]^4(0) = 1234$, $[0]^4(3) = 3333$, $[+1,-1]^4(1) = 21212121$, y $[[+1]^4, [-2,[+1]^3]^2](0) = 123423453456$. Finalmente, $[P^s\{i\}]$ representa una repetición con cambios. Por ejemplo: $[[[V]^2]^4\{-1\}](3) = 43526170$, y $[[[+1]^4]^3\{+1\}](0) = 123423453456$.

Complejidad de Kolmogorov. Para una cadena $\sigma \in \{0,\ldots,7\}^+$, definimos la complejidad de Kolmogorov de σ relativa a j como $K^j_{\mathcal{LG}}(\sigma) \stackrel{\text{def}}{=} \text{mín}\{|P|\colon P(j) = \sigma\}$, donde $|P|$ representa el *tamaño* de P definido en la Figura 4. La complejidad de Kolmogorov de σ se define por $K_{\mathcal{LG}}(\sigma) \stackrel{\text{def}}{=} \text{mín}\{K^j_{\mathcal{LG}}(\sigma) \mid j \in \{0,\ldots,7\}\}$. Se puede ver que $K_{\mathcal{LG}}$ resulta una función computable.

3. Actividades y metodología

El presente proyecto articula el estudio teórico-computacional y experimental de la representación del conocimiento en el proceso cognitivo. Así, se plantea un diálogo constante entre la computación y la neurociencia (facilitados por el director y el codirector), como también entre la teoría y los datos que emerjan de los experimentos.

La metodología general a aplicar en el desarrollo de esta investigación consiste principalmente en iteraciones continuas de las siguientes tareas: 1) Revisión bibliográfica: estudio detallado de los conceptos a partir de bibliografía y trabajos relacionados. 2) Definición de un dominio de tareas sobre las que estudiar la representación del conocimiento. Empezaremos por la representación de intuiciones geométricas pero dejamos abierto el estudio a otros lenguajes y tareas cognitivas más adelante. 3) Desarrollo, análisis e implementación computacional de un lenguaje de representación para dicho dominio. En el caso del dominio de geometría, proponemos trabajar con el lenguaje descripto, $\mathcal{LG}$. 4) Diseño de experimentos y prueba con conjuntos de participantes acotados para la validación inicial del modelo propuesto y la estimación de las variables iniciales para el experimento principal. 5) Experimentación con un mayor número de participantes, análisis estadístico de los resultados y comparación con otras técnicas del estado del arte aplicables a los mismos conjuntos de datos. 6) Publicación y discusión de los resultados con otros expertos del área y a través de informes, presentaciones en congresos y artículos en revistas.

Todos los experimentos involucrados en este plan de trabajo han sido aprobados por el Comité de Ética de la Dirección de Investigación del Centro de Educación Médica e Investigación Clínica "Norberto Quirno" (CEMIC), Unidad Asociada del CONICET (Protocolo No 435) o están en trámite.

4. Factibilidad

El Departamento de Computación de la Universidad de Buenos Aires cuenta con la infraestructura y el equipamiento necesario para llevar a cabo esta investigación a los cuales el postulante tendrá acceso.

El Dr. Figueira —director propuesto en este plan— dirige el grupo de Lógica y Computabilidad (GLyC) de este centro, y dispone de un escritorio y computadora con acceso a internet que serán destinados a las actividades diarias del postulante. Su experiencia y conocimientos en computabilidad y teoría algorítmica de la información, resultarán fundamentales para el desarrollo de los modelos computacionales basados en inferencia inductiva.

Por otro lado, el Dr. Sigman —codirector propuesto— realiza sus actividades en el Laboratorio de Neurociencia Integrativa (LNI) del Departamento de Física de la Universidad de Buenos Aires, y posee una extensa trayectoria en diversos temas de neurociencia y cognición. Su experiencia resultará clave para el diseño y la orientación del proyecto desde la perspectiva de la neurociencia, contribuyendo a delinear las preguntas relevantes y el modo óptimo de abordarlos tanto en el diseño de los experimentos como en el análisis de los datos. El LNI posee los recursos necesarios y una vasta trayectoria para la implementación de experimentos en humanos con interfaz gráfica de usuario como los utilizados en la tesis de grado o del tipo que proponemos desarrollar en este proyecto. Así, contaremos con el conocimiento y la infraestructura necesaria para la rápida implementación de estos.

A su vez, estos centros cuentan con la bibliografía específica sobre la temática y también poseen acceso a publicaciones científicas de primer nivel a través de la biblioteca electrónica del MinCyT.

Las actividades del becario estarán enmarcadas en los proyectos *Aplicaciones de la Teoría de la Aleatoriedad Algorítmica* (PICT-2011-0365 y UBACyT 20020110100025), dirigidos por el Dr. Figueira.

Referencias

[1] M. Banko y E. Brill. Scaling to very very large corpora for natural language disambiguation. En *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, páginas 26–33, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.

[2] C. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.

[3] P. Bloom. *How children learn the meanings of words*. Learning, Development, and Conceptual Change. MIT Press, 2002.

[4] A. Blumer, A. Ehrenfeucht, D. Haussler, y M. K. Warmuth. Occam's razor. *Inf. Process. Lett.*, 24(6):377–380, Abril 1987.

[5] N. Chater y C. D. Manning. Probabilistic models of language processing and acquisition. *Trends Cogn. Sci. (Regul. Ed.)*, 10(7):335–344, Jul 2006.

[6] S. Dehaene, V. Izard, P. Pica, y E. Spelke. Core knowledge of geometry in an amazonian indigene group. *Science*, 311(5759):381–384, 2006.

[7] R. Downey y D. Hirschfeldt. *Algorithmic Randomness and Complexity*. Theory and Applications of Computability Series. Springer, 2010.

[8] G. Forman y I. Cohen. Learning from little: Comparison of classifiers given little training. En *in 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD*, páginas 161–172, 2004.

[9] A. Gopnik y A. Meltzoff. *Words, Thoughts, and Theories*. A Bradford book. MIT Press, 1997.

[10] T. L. Griffiths y J. B. Tenenbaum. Optimal predictions in everyday cognition. *Psychol Sci*, 17(9):767–773, Sep 2006.

[11] T. L. Griffiths, N. Chater, C. Kemp, A. Perfors, y J. B. Tenenbaum. Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8):357–364, June 2010.

[12] T. L. Griffiths y J. B. Tenenbaum. Randomness and coincidences: Reconciling intuition and probability theory. *23rd Annual Conference of the Cognitive Science Society*, páginas 370–375, 2001.

[13] E. Markman. *Categorization and Naming in Children: Problems of Induction*. Bradford Books. MIT Press, 1991.

[14] D. A. McAllester. Some pac-bayesian theorems. En *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, páginas 230–234, New York, NY, USA, 1998. ACM.

[15] T. M. Mitchell. The need for biases in learning generalizations. Technical report, 1980.

[16] A. Nies. *Algorithmic Randomness and Complexity*. Clarendon Press, Oxford, 2009.

[17] S. Romano. Generación de azar en humanos: modelo computacional. Tesis de grado, Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina, 2012.

[18] S. Romano, M. Sigman, y S. Figueira. LT^2C^2: A language of thought with Turing-computable Kolmogorov complexity. *Papers in Physics*, 5:050001, 2013.

[19] S. Russell y P. Norvig. *Artificial Intelligence: Pearson New International Edition: A Modern Approach*. Always learning. Pearson Education, Limited, 2013.

[20] R. N. Shepard. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210:390–398, 1980.

[21] R. Solomonoff. A preliminary report on a general theory of inductive inference. Technical Report V-131, Zator Co. and Air Force Office of Scientific Research, Cambridge, Mass., Feb 1960.

[22] E. Spelke, S. A. Lee, y V. Izard. Beyond core knowledge: Natural geometry. *Cognitive Science*, 34(5):863–884, 2010.

[23] J. B. Tenenbaum y T. L. Griffiths. Generalization, similarity, and Bayesian inference. *Behav Brain Sci*, 24(4):629–640, Aug 2001.

[24] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, y N. D. Goodman. How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, 331(6022):1279–1285, Marzo 2011.

[25] G. M. Weiss y F. Provost. Learning when training data are costly: The effect of class distribution on tree induction. *J. Artif. Int. Res.*, 19(1):315–354, Octubre 2003.