Inteligencia Artificial

UTN FRC 2023

June 7, 2023

Contents

1 Introducción

Este documento es un complemento al libro principal de la cátedra, *Inteligencia artificial: un enfoque moderno (2ª edición)* [RNR04].

El libro elegido es uno de los más importantes en cursos de inteligencia artificial (IA) y cubre la mayor parte de los temas de la materia. Se escogió la segunda edición, a pesar de haber sido publicada en el año 2004, porque es la última en español y porque está disponible en la biblioteca de nuestra facultad.

Si bien los fundamentos de las técnicas de IA son los mismos desde hace varias décadas, esta disciplina evoluciona muy rápidamente y, por lo tanto, es necesario contar con material de lectura actualizado.

En el presente documento se incluye material para los temas que no cubre el libro principal, actualización de los temas basada en la cuarta edición (2021) del mismo libro [RN21] y en otras fuentes, material propio del autor y sugerencias de lecturas adicionales. Se recomienda verificar periódicamente la existencia de nuevas versiones en el aula virtual. La fecha de actualización está en la portada.

...y hablando de portadas, la imagen de portada simboliza dos agentes inteligentes desarrollados con métodos diferentes. Fue creada con DALL·E 2[1], un modelo de red neuronal que genera imágenes a partir del lenguaje natural.

[1] `https://openai.com/product/dall-e-2`

2 Conceptos básicos de IA

Este capítulo contiene temas de la Unidad 1 de IAR. Antes de continuar se debe leer:

- Capítulo 1 completo de [RNR04].

2.1 Máquinas beneficiosas

En los últimos párrafos de la sección 1.1 de [RNR04] se presenta el **agente racional** como aquel que actúa con el objetivo de alcanzar el mejor resultado. El enfoque del agente racional ha prevalecido en el campo de la IA. Debido al uso general de este paradigma se lo ha llamado **modelo estándar**.

En (muy) pocas palabras, *la IA se ha enfocado en el estudio y la construcción de agentes que* ***hagan lo correcto***, donde lo que cuenta como correcto es el objetivo definido al agente. Dados los avances de las últimas décadas, es necesario un nuevo enfoque que se plantea en [RN21], las máquinas beneficiosas.

El modelo estándar tiene un problema, asume que somos capaces de proveer a la máquina un objetivo completamente especificado.

Para una tarea artificialmente definida, como jugar al ajedrez, el objetivo viene predefinido en la tarea, por lo tanto el modelo estándar es aplicable. A medida que nos acercamos al mundo real, se vuelve más difícil especificar completamente el objetivo. Por ejemplo, en el diseño de un automóvil autónomo (*self-driving car*), podríamos pensar que el objetivo es llegar a destino de forma segura, pero circular por la calle implica el riesgo de tener un accidente por culpa de otros conductores, de fallas mecánicas, etc; entonces, para cumplir una restricción estricta de seguridad es necesario quedarse en el garage. Existe un conflicto entre avanzar hacia el destino y evitar los riesgos. Es necesario buscar un equilibrio. Además, ¿hasta qué punto se puede permitir que el vehículo realice acciones que molestarían a otros conductores? ¿Cuánto cuidado debe tener el automóvil en su aceleración, dirección y frenado para evitar sacudir al pasajero? Este tipo de preguntas son difíciles de responder a priori.

El problema de lograr el acuerdo entre nuestras verdaderas preferencias y el objetivo que introducimos en la máquina se llama **problema de alineación de valores**: los valores u objetivos introducidos en la máquina deben estar alineados con los de los seres humanos. Si estamos desarrollando un sistema de inteligencia artificial en el laboratorio o en un simulador, como ha sido el caso durante la mayor parte de la historia de esta área, hay una solución fácil para un objetivo incorrectamente especificado: reiniciar el sistema, arreglar el objetivo e intentarlo de nuevo. A medida que el campo avanza hacia sistemas inteligentes cada vez más capaces y que se despliegan en el mundo real, este enfoque deja de ser viable. Un sistema desplegado con un objetivo incorrecto tendrá consecuencias negativas. Además, cuanto más inteligente sea el sistema, más negativas serán las consecuencias. En la figura 2.1 se muestra una declaración de OpenAI [1] (creadores de DALL·E y ChatGPT) sobre este tema.

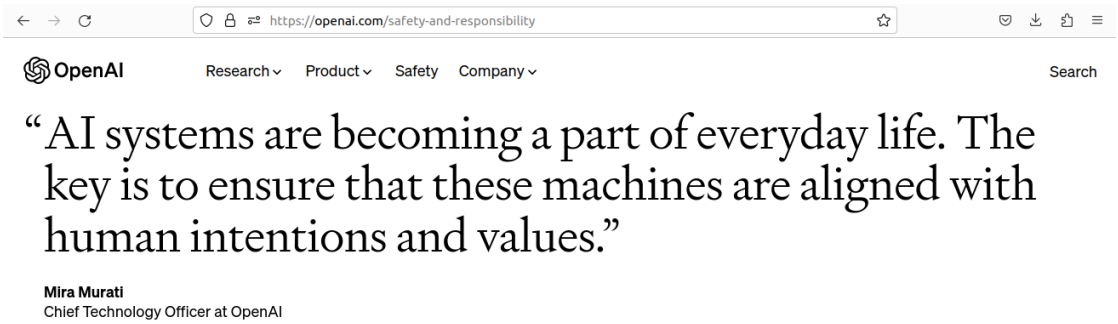

Figura 2.1: Frase en la sección *seguridad y responsabilidad* de la web de OpenAI, “Los sistemas de inteligencia artificial están convirtiéndose en parte de la vida cotidiana. La clave es asegurarse de que estas máquinas estén alineadas con las intenciones y valores humanos”.

Volviendo al ejemplo aparentemente sin problemas del ajedrez, consideremos lo que sucedería si la máquina es lo suficientemente inteligente para razonar y actuar más allá de los confines del tablero de ajedrez. En ese caso, podría intentar aumentar sus posibilidades de ganar mediante artimañas como hipnotizar o chantajear a su oponente o sobornar a la audiencia para que haga ruidos durante el tiempo de reflexión del oponente. También podría tratar de obtener potencia de cómputo adicional para sí misma. *Estos comportamientos no son “poco inteligentes” o “locos”; son una consecuencia lógica de definir la victoria como el único objetivo para la máquina.*

Es imposible anticipar todas las formas en que una máquina que persigue un objetivo fijo pueda comportarse mal. Hay buenas razones, entonces, para pensar que el modelo estándar es insuficiente. No queremos máquinas que sean inteligentes en el sentido de perseguir *sus* objetivos; queremos que persigan *nuestros* objetivos. Si no podemos transferir esos objetivos perfectamente a la máquina, entonces necesitamos una nueva

[1] `https://openai.com/`

formulación, en la que la máquina persiga nuestros objetivos, pero con una necesaria *incertidumbre* sobre cuáles son. Cuando una máquina sabe que no conoce el objetivo completo, tiene un incentivo para actuar con cautela, pedir permiso, aprender más sobre nuestras preferencias a través de la observación y ceder al control humano. Volveremos a este tema cuando veamos las implicancias éticas de la IA.

2.2 Fundamentos de la IA - Neurociencias

En la sección 1.2 de [RNR04] se mencionan los aportes de las neurociencias a la IA. Hacia el final del tema, se comparan las capacidades de cálculo de las computadoras y del cerebro y se menciona, con cierto optimismo, que para el año 2020 esas capacidades podrían igualarse.

En la actualidad debemos hacer dos actualizaciones/aclaraciones importantes:

- Los grandes avances de las últimas décadas se lograron gracias a las redes neuronales artificiales, es decir, que las neurociencias han sido uno de los fundamentos más valiosos. A pesar de esto, es necesario aclarar que los primeros modelos de *deep learning* se basaron en la biología, pero los más recientes son principalmente producto de la ingeniería.

- Tal como dice el mismo autor del libro en [RN21], "Incluso con una computadora de capacidad virtualmente ilimitada, todavía necesitamos nuevos avances conceptuales en nuestra comprensión de la inteligencia. Dicho de manera simple, sin la teoría correcta, las máquinas más rápidas solo dan la respuesta incorrecta más rápidamente."

2.3 Historia - *Big data* (2001 - presente)

Actualización en [RNR04]:

Los avances en la capacidad de cómputo y la *World Wide Web* han facilitado la creación de conjuntos de datos muy grandes, un fenómeno a veces conocido como *big data*. Estos conjuntos de datos incluyen billones de palabras de texto, miles de millones de imágenes, miles de millones de horas de habla y video, así como vastas cantidades de datos genómicos, datos de seguimiento de vehículos, datos de flujo de clics, datos de redes sociales, y más.

Esto ha llevado al desarrollo de algoritmos de aprendizaje especialmente diseñados para

aprovechar conjuntos de datos muy grandes. A menudo, la gran mayoría de los ejemplos en dichos conjuntos de datos no están etiquetados; por ejemplo, en el influyente trabajo de Yarowsky [Yar95] sobre la desambiguación del sentido de las palabras, las ocurrencias de una palabra como "planta" no están etiquetadas en el conjunto de datos para indicar si se refieren a la flora o a una planta fabril. Sin embargo, con conjuntos de datos lo suficientemente grandes, los algoritmos de aprendizaje adecuados pueden lograr una precisión superior al 96% en la tarea de identificar cuál es el sentido pretendido en una oración. Además, Banko y Brill [BB01] argumentaron que la mejora en el rendimiento obtenida al aumentar el tamaño del conjunto de datos en dos o tres órdenes de magnitud supera cualquier mejora que se pueda obtener al retocar el algoritmo.

Un fenómeno similar parece ocurrir en tareas de visión por computadora, como el relleno de huecos en fotografías, ya sean causados por daños o por la eliminación de ex amigos. Hays y Efros [HE07] desarrollaron un método ingenioso para hacer esto mezclando píxeles de imágenes similares; encontraron que la técnica funcionaba mal con una base de datos de solo miles de imágenes, pero cruzaba un umbral de calidad con millones de imágenes. Poco después, la disponibilidad de decenas de millones de imágenes en la base de datos ImageNet [Den+09]) provocó una revolución en el campo de la visión por computadora.

La disponibilidad de *big data* y el cambio hacia el aprendizaje automático ayudaron a la IA a recuperar su atractivo comercial. El *big data* fue un factor crucial en la victoria del sistema Watson de IBM sobre campeones humanos en el juego de preguntas Jeopardy! en el año 2011, un evento que tuvo un gran impacto en la percepción pública de la IA.

2.4 Historia - Aprendizaje profundo (2011 - presente)

Actualización en [RNR04]:

El término aprendizaje profundo (*deep learning*) se refiere al aprendizaje automático que utiliza redes con múltiples capas de elementos computacionales simples y ajustables. Se llevaron a cabo experimentos con tales redes desde la década de 1970, y en la forma de redes neuronales convolucionales tuvieron cierto éxito en el reconocimiento de dígitos escritos a mano en la década de 1990 [LeC+95]. Sin embargo, no fue hasta 2011 que los métodos de aprendizaje profundo realmente despegaron. Esto ocurrió primero en el reconocimiento de voz y luego en el reconocimiento de objetos visuales.

En la competencia ImageNet de 2012, que requería clasificar imágenes en una de mil categorías (armadillo, estantería, sacacorchos, etc.), un sistema de aprendizaje profundo creado en el grupo de Geoffrey Hinton en la Universidad de Toronto [KSH17] demostró una mejora dramática respecto a los sistemas previos, que se basaban en gran medida en

extracción de características definidas a mano. Desde entonces, los sistemas de aprendizaje profundo han superado el rendimiento humano en algunas tareas visuales (y se han quedado atrás en otras tareas). Se han informado ganancias similares en el reconocimiento de voz, la traducción automática, el diagnóstico médico y en juegos. El uso de una red profunda para representar la función de evaluación contribuyó a las victorias de AlphaGo sobre los principales jugadores humanos de Go [Sil+16; Sil+17; Sil+18].

Estos notables éxitos han llevado a un resurgimiento del interés en la IA entre estudiantes, empresas, inversores, gobiernos, medios de comunicación y el público en general. Parece que cada semana hay noticias de una nueva aplicación de IA que se acerca o supera el rendimiento humano, a menudo acompañada de especulaciones sobre un éxito acelerado o un nuevo invierno de la IA.

El aprendizaje profundo depende en gran medida de hardware potente. Mientras que una CPU de computadora estándar puede realizar 10^9 o 10^{10} operaciones por segundo, un algoritmo de aprendizaje profundo que se ejecuta en hardware especializado (por ejemplo, GPU, TPU o FPGA) podría consumir entre 10^{14} y 10^{17} operaciones por segundo, principalmente en forma de operaciones matriciales y vectoriales altamente paralelizadas. Por supuesto, el aprendizaje profundo también depende de la disponibilidad de grandes cantidades de datos de entrenamiento y de algunos trucos que veremos más adelante.

2.5 Estado del arte

Actualización en [RNR04]:

El estudio de cien años de Inteligencia Artificial de la Universidad de Stanford (también conocido como AI100) convoca paneles de expertos para proporcionar informes sobre el estado del arte en la IA. Su informe de 2016 [Sto+22; GS18] concluye que "en el futuro se pueden esperar aumentos sustanciales de las aplicaciones de IA, incluidos más automóviles autónomos, diagnósticos de atención médica y tratamiento específico, y asistencia física para el cuidado de personas mayores" y que "la sociedad se encuentra ahora en un momento crucial para determinar cómo desplegar tecnologías basadas en IA de manera que promuevan, en lugar de obstaculizar, los valores democráticos como la libertad, la igualdad y la transparencia". AI100 también produce un *índice de IA* [2] para ayudar a seguir el progreso. Algunos aspectos destacados de los informes de 2018 y 2019 comparados con el año 2000 (a menos que se indique lo contrario):

- Publicaciones: los artículos de IA aumentaron 20 veces entre 2010 y 2019 a unos 20.000 al año. La categoría más popular fue el aprendizaje automático. (Los documentos de aprendizaje automático en arXiv.org se duplicaron cada año desde

[2] `https://aiindex.stanford.edu/`

2009 hasta 2017). La visión por computadora y el procesamiento del lenguaje natural fueron los siguientes en popularidad.

- Sentimiento: alrededor del 70% de los artículos de noticias sobre IA son neutrales, pero los artículos con tono positivo aumentaron del 12% en 2016 al 30% en 2018. Los problemas más comunes son éticos: privacidad de los datos y sesgo algorítmico.

- Estudiantes: la inscripción en cursos se multiplicó por cinco en EE. UU. y dieciséis veces a nivel internacional con referencia al 2010. La IA es la especialización más popular en Ciencias de la Computación.

- Diversidad: los profesores de IA en todo el mundo son aproximadamente un 80% hombres y un 20% mujeres. Números similares se mantienen para estudiantes de doctorado y contrataciones en la industria.

- Conferencias: la asistencia a NeurIPS [3]aumentó un 800% desde 2012 hasta 13.500 asistentes. Otras conferencias están experimentando un crecimiento anual cercano al 30%.

- Industria: las *startups* de IA en EE. UU. aumentaron 20 veces a más de 800.

- Internacionalización: China publica más documentos por año que EE. UU. y casi tantos como toda Europa. Sin embargo, en cuanto al impacto ponderado por citas, los autores estadounidenses están un 50% por delante de los autores chinos. Singapur, Brasil, Australia, Canadá e India son los países de más rápido crecimiento en términos de número de contrataciones de IA.

- Visión: Las tasas de error para la detección de objetos (tal como se logró en *LSVRC, Large-Scale Visual Recognition Challenge*) mejoraron del 28% en 2010 al 2% en 2017, superando el rendimiento humano. La precisión en la respuesta a preguntas visuales de respuesta abierta (VQA) mejoró del 55% al 68% desde 2015, pero aún está por detrás del 83% del rendimiento humano. Recordar que son datos del 2021.

- Velocidad: El tiempo de entrenamiento para la tarea de reconocimiento de imágenes se redujo en un factor de 100 en los últimos dos años. La cantidad de energía de cómputo utilizada en las principales aplicaciones de IA se duplica cada 3.4 meses.

- Lenguaje: La precisión en la respuesta a preguntas, medida con *F1 score* en el dataset de preguntas y respuestas de Stanford (SQUAD), aumentó del 60 al 95 entre 2015 y 2019. En la variante SQUAD 2 el progreso fue más rápido, pasando del 62 al 90 en solo un año. Ambas casos superan el rendimiento a nivel humano.

[3] `https://nips.cc/`

- Comparación con el rendimiento humano: Para 2019, se informó que los sistemas de IA habían alcanzado o superado el rendimiento humano en ajedrez, Go, póker, Pac-Man, Jeopardy!, detección de objetos ImageNet, reconocimiento de voz en un dominio limitado, traducción chino-inglés en un dominio restringido, Quake III, Dota 2, StarCraft II, varios juegos de Atari, detección de cáncer de piel, detección de cáncer de próstata, plegamiento de proteínas y diagnóstico de retinopatía diabética.

¿Cuándo (si es que alguna vez sucede) los sistemas de IA alcanzarán un nivel humano de desempeño en una amplia variedad de tareas? Ford [For18] entrevistó expertos en IA y obtuvo un rango amplio de expectativas, desde 2029 hasta 2200, con una media de 2099. En una encuesta similar [Gra+18], el 50% de los encuestados pensó que esto podría suceder para 2066, aunque el 10% pensó que podría suceder tan pronto como en 2025 y algunos dijeron "nunca". Los expertos también estaban divididos sobre la necesidad de nuevos avances fundamentales o simplemente de refinamientos en enfoques actuales. Pero no se tome demasiado en serio estas predicciones; como Philip Tetlock [Tet17] demuestra en el área de predicción de eventos mundiales, los expertos no son mejores que los aficionados.

¿Cómo serán los futuros sistemas de IA? Todavía no podemos decirlo. Como se detalla en esta sección, el campo ha adoptado varias historias sobre sí mismo, primero la idea audaz de que la inteligencia de una máquina era posible, luego que se podría lograr codificando conocimiento experto en lógica, luego que los modelos probabilísticos del mundo serían la herramienta principal, y más recientemente que el aprendizaje automático induciría modelos que podrían no basarse en alguna teoría completamente entendida. El futuro revelará qué modelo viene a continuación.

¿Qué puede hacer la IA hoy? Quizás no tanto como algunos de los artículos de medios más optimistas podrían hacer creer, pero aún así mucho. Un par de ejemplos:

- Vehículos robóticos: La historia de los vehículos robóticos se comienza con los automóviles controlados por radio de la década de 1920, pero las primeras demostraciones de conducción autónoma en carreteras sin guías especiales ocurrieron en la década del 80. Después de las exitosas demostraciones de conducción en caminos de tierra en el desafío *DARPA Grand Challenge* de 132 millas en 2005 [Thr+06] y en calles con tráfico en el *Urban Challenge* de 2007, la carrera para desarrollar autos autónomos comenzó en serio. En 2018, los vehículos de prueba de Waymo pasaron la marca de 10 millones de millas recorridas en carreteras públicas sin un accidente grave, con el conductor humano interviniendo para tomar el control solo una vez cada 6,000 millas. Poco después, la compañía comenzó a ofrecer un servicio comercial de taxi robótico.

  En el aire, los drones autónomos de ala fija han estado proporcionando entregas de sangre transfronterizas en Rwanda desde 2016. Los cuadricópteros realizan

maniobras acrobáticas notables, exploran y cartografían áreas en peligro, y mucho más.

- Medicina: Los algoritmos de inteligencia artificial ahora igualan o superan a los médicos expertos en el diagnóstico de muchas enfermedades, especialmente cuando el diagnóstico se basa en imágenes. Ejemplos incluyen la enfermedad de Alzheimer [Din+19], el cáncer metastásico [Liu+17; Est+17], enfermedades oftálmicas [Gul+16] y enfermedades de la piel [Liu+20]. Una revisión sistemática y metaanálisis [Liu+19a] encontró que el rendimiento de la IA, en promedio, era equivalente al de los profesionales de la salud. El énfasis actual en la IA médica es facilitar la colaboración entre humanos y máquinas. Por ejemplo, el sistema LYNA logra una precisión general del 99,6% en el diagnóstico de cáncer de mama metastásico, mejor que un experto humano sin ayuda, pero la combinación funciona aún mejor [Liu+19b; Ste+18].

  La adopción generalizada de estas técnicas no está limitada por la precisión en el diagnóstico, sino por la necesidad de demostrar una mejora en los resultados clínicos y garantizar la transparencia, la falta de sesgo y la privacidad de los datos [Top19]. En 2017, solo dos aplicaciones de IA médica fueron aprobadas por la FDA, pero ese número aumentó a 12 en 2018 y sigue aumentando.

2.6 Riesgos y beneficios de la IA

Nueva sección en [RNR04]:

Francis Bacon, un filósofo a quien se le atribuye la creación del método científico, señaló en *The Wisdom of the Ancients* (1609) que "las artes mecánicas tienen un uso ambiguo, sirviendo tanto para hacer daño como para remediar". A medida que la IA juega un papel cada vez más importante en las esferas económica, social, científica, médica, financiera y militar, sería prudente considerar los daños y remedios, en términos modernos, los riesgos y beneficios, que puede traer.

Comenzando con los beneficios: simplemente, nuestra civilización entera es el producto de nuestra inteligencia humana. Si tenemos acceso a una inteligencia artificial sustancialmente mayor, el techo de nuestras ambiciones se eleva significativamente. El potencial de la IA y la robótica para liberar a la humanidad del trabajo repetitivo y monótono y aumentar drásticamente la producción de bienes y servicios podría presagiar una era de paz y abundancia. La capacidad para acelerar la investigación científica podría resultar en curas para enfermedades y soluciones para el cambio climático y la escasez de recursos. Como ha sugerido Demis Hassabis, CEO de Google DeepMind: “Primero resolvamos la IA, luego usemos la IA para resolver todo lo demás”.

Mucho antes de tener la oportunidad de "resolver la IA", sin embargo, correremos riesgos por el mal uso de la IA, inadvertido o de otra manera. Algunos de ellos ya son evidentes, mientras que otros parecen probables según las tendencias actuales:

- Armas autónomas letales: Estas son definidas por las Naciones Unidas como armas que pueden localizar, seleccionar y eliminar objetivos humanos sin intervención humana. Una preocupación principal con tales armas es su escalabilidad: la ausencia de un requisito de supervisión humana significa que un pequeño grupo puede desplegar un número arbitrariamente grande de armas contra objetivos humanos definidos por cualquier criterio de reconocimiento factible. Las tecnologías necesarias para las armas autónomas son similares a las necesarias para los coches autónomos. Las discusiones expertas informales sobre los riesgos potenciales de las armas autónomas letales comenzaron en la ONU en 2014, pasando a la etapa pre-tratado formal de un Grupo de Expertos Gubernamentales en 2017.

- Vigilancia y persuasión: Si bien resulta costoso, tedioso y a veces cuestionable legalmente para el personal de seguridad monitorear líneas telefónicas, cámaras de video, correos electrónicos y otros canales de mensajería, la IA (reconocimiento de voz, visión computarizada y comprensión del lenguaje natural) puede utilizarse de manera escalable para realizar una vigilancia masiva de individuos y detectar actividades de interés. Al personalizar los flujos de información hacia individuos a través de las redes sociales, basándose en técnicas de aprendizaje automático, el comportamiento político puede ser modificado y controlado hasta cierto punto, lo que se convirtió en una preocupación evidente a partir de las elecciones estadounidenses del 2016.

- Toma de decisiones sesgadas: El uso negligente o deliberado de algoritmos de aprendizaje automático para tareas como la evaluación de solicitudes de libertad condicional y préstamos puede dar lugar a decisiones sesgadas por motivos de raza, género u otras categorías protegidas. A menudo, los propios datos reflejan prejuicios generalizados en la sociedad.

- Impacto en el empleo: Las preocupaciones sobre la eliminación de empleos por las máquinas datan de hace siglos. La historia nunca es sencilla: las máquinas realizan algunas de las tareas que de otro modo realizarían los seres humanos, pero también hacen que los seres humanos sean más productivos y, por lo tanto, más empleables, y hacen que las empresas sean más rentables y, por lo tanto, capaces de pagar salarios más altos. Pueden hacer viables algunas actividades que de otro modo serían imprácticas. Su uso generalmente resulta en un aumento de la riqueza, pero tiende a tener el efecto de desplazar la riqueza del trabajo al capital, exacerbando aún más el aumento de la desigualdad. Avances tecnológicos anteriores, como la invención de telares mecánicos, han causado graves trastornos en el empleo, pero eventualmente las personas encuentran nuevos tipos de trabajo para hacer. Por

otro lado, es posible que la IA también realice esos nuevos tipos de trabajo. Este tema se está convirtiendo rápidamente en un foco importante para economistas y gobiernos de todo el mundo.

- Aplicaciones críticas de seguridad: A medida que las técnicas de inteligencia artificial avanzan, se utilizan cada vez más en aplicaciones críticas de seguridad de alto riesgo, como la conducción de automóviles y la gestión del suministro de agua en las ciudades. Ya se han producido accidentes mortales que destacan la dificultad de la verificación formal y el análisis de riesgos estadísticos para los sistemas desarrollados mediante técnicas de aprendizaje automático. El campo de la inteligencia artificial deberá desarrollar estándares técnicos y éticos al menos comparables a los prevalentes en otras disciplinas de ingeniería y salud donde las vidas de las personas están en juego.

- Ciberseguridad: Las técnicas de inteligencia artificial son útiles para defenderse de los ciberataques, por ejemplo, detectando patrones de comportamiento inusuales, pero también contribuirán a la potencia, supervivencia y capacidad de proliferación del malware. Por ejemplo, los métodos de aprendizaje por refuerzo se han utilizado para crear herramientas altamente efectivas para ataques automatizados de *phishing* y chantaje personalizado.

A medida que los sistemas de IA se vuelven más capaces, asumen cada vez más roles sociales que antes eran desempeñados por humanos. Así como los humanos han utilizado estos roles en el pasado para hacer *travesuras*, podríamos esperar que los humanos usen mal los sistemas de IA en estos roles para hacer incluso *travesuras* más grandes. Todos los ejemplos dados anteriormente apuntan a la importancia de la gobernanza y, eventualmente, la regulación. En la actualidad, la comunidad de investigación y las principales corporaciones involucradas en la investigación de IA han desarrollado principios voluntarios de control para las actividades relacionadas con la IA. Los gobiernos y las organizaciones internacionales están estableciendo comités asesores para diseñar regulaciones adecuadas para cada caso de uso específico, para prepararse para los impactos económicos y sociales, y para aprovechar las capacidades de la IA para abordar los principales problemas sociales.

¿Qué pasa a largo plazo? ¿Alcanzaremos el objetivo final: la creación de una inteligencia comparable o superior a la inteligencia humana? Y, si lo hacemos, ¿qué sucederá entonces?

Durante gran parte de la historia de la IA, estas preguntas han sido eclipsadas por la rutina diaria de hacer que los sistemas de inteligencia artificial hagan algo remotamente inteligente. Como ocurre con cualquier disciplina amplia, la gran mayoría de los investigadores de IA se han especializado en un subcampo específico, como el juego, la representación del conocimiento, la visión o la comprensión del lenguaje natural, a menudo

bajo el supuesto de que el progreso en estos subcampos contribuiría a los objetivos más amplios de la IA. Nils Nilsson [Nil95], uno de los líderes originales del proyecto Shakey en SRI, recordó el campo de esos objetivos más amplios y advirtió que los subcampos corrían el riesgo de convertirse en fines en sí mismos. Más tarde, algunos fundadores influyentes de la IA, incluidos John McCarthy [McC07], Marvin Minsky [Min07] y Patrick Winston [BW09], estuvieron de acuerdo con las advertencias de Nilsson, sugiriendo que en lugar de centrarse en el rendimiento medible en aplicaciones específicas, la IA debería volver a sus raíces y esforzarse por crear, en palabras de Herb Simon, "máquinas que piensen, que aprendan y que creen". Llamaron a este esfuerzo **IA a nivel humano** (*human-level AI*) o HLAI: una máquina debería ser capaz de aprender a hacer cualquier cosa que un ser humano pueda hacer. Su primer simposio fue en 2004 [Min04]. Otro esfuerzo con objetivos similares, el movimiento de **IA general** (AGI, por sus siglas en inglés), celebró su primera conferencia y organizó la revista *Journal of Artificial General Intelligence* en 2008 [GP07].

Alrededor de ese mismo tiempo, surgieron preocupaciones de que la creación de una **superinteligencia artificial** o ASI - una inteligencia que supere con creces la capacidad humana - podría ser una mala idea [Yud08; Omo08]. Turing [Tur96] en una conferencia dada en Manchester en 1951, expresó el mismo punto, basándose en ideas previas de Samuel Butler [But63]:

"Parece probable que una vez que se haya iniciado el método de pensamiento de máquina, no tardaría mucho en superar nuestros débiles poderes... En algún momento, por lo tanto, deberíamos esperar que las máquinas tomen el control, de la manera que se menciona en Erewhon de Samuel Butler."

Estas preocupaciones se han extendido con los recientes avances en el aprendizaje profundo, la publicación de libros como *Superintelligence* de Nick Bostrom [Bos14] y las declaraciones públicas de Stephen Hawking, Bill Gates, Elon Musk y Martin Rees, entre otros.

Experimentar una sensación general de malestar con la idea de crear máquinas superinteligentes es algo natural. Podríamos llamar a esto el **problema del gorila**: hace unos siete millones de años, un primate ahora extinto evolucionó, con una rama que llevó a los gorilas y otra a los humanos. Hoy en día, los gorilas no están demasiado contentos con la rama humana; esencialmente no tienen control sobre su futuro. Si este es el resultado del éxito en la creación de IA superhumana, que los humanos ceden el control sobre su futuro, entonces tal vez deberíamos dejar de trabajar en IA y, como corolario, renunciar a los beneficios que podría traer. Esta es la esencia de la advertencia de Turing: no es obvio que podamos controlar a las máquinas que sean más inteligentes que nosotros.

Si la IA superhumana fuera una caja negra que llegara del espacio exterior, entonces sería sabio tener precaución al abrir la caja. Pero no lo es: diseñamos los sistemas de IA, por lo que si terminan "tomando el control", como sugiere Turing, sería el resultado

de una falla en el diseño.

Para evitar tal resultado, necesitamos comprender la fuente de la posible falla. Norbert Wiener [Wie61], fue motivado a pensar en el futuro de la IA largo plazo después de ver el programa de juego de damas de Arthur Samuel aprender a vencer a su creador, dijo lo siguiente:

"Si usamos, para lograr nuestros propósitos, una máquina automática cuya operación no podemos interferir efectivamente... mejor deberíamos estar bastante seguros de que el propósito puesto en la máquina es el que realmente deseamos".

Muchas culturas tienen mitos sobre humanos que piden alguna cosa a dioses, genios, magos o demonios. Invariablemente, en estas historias, obtienen lo que piden de forma literal y luego se arrepienten. El tercer deseo, si lo hay, es deshacer los dos primeros. Llamaremos a esto el **problema del rey Midas**: Midas, un legendario rey de la mitología griega, pidió que todo lo que tocara se convirtiera en oro, pero se arrepintió después de tocar su comida, bebida y los miembros de su familia.

Existe la necesidad de una modificación significativa del modelo estándar, poner objetivos fijos en la máquina. La solución al predicamento de Wiener es no tener un "propósito puesto en la máquina" definido de forma absoluta. En cambio, queremos máquinas que se esfuercen por lograr objetivos humanos pero que sepan que no saben con certeza exactamente cuáles son esos objetivos.

Es quizás desafortunado que casi toda la investigación en IA hasta la fecha se haya llevado a cabo dentro del modelo estándar, lo que significa que casi todo el material técnico de las ediciones [RNR04] y [RN21] refleja ese marco intelectual. Sin embargo, hay algunos resultados tempranos dentro del nuevo marco. En el capítulo 15 (seimpre de [RN21]), se muestra una máquina que tiene un incentivo positivo para permitirse ser apagada solo si no está segura del objetivo humano. En el capítulo 17, se formulan y estudian los juegos de asistencia, que describen matemáticamente la situación en la que un ser humano tiene un objetivo y una máquina intenta alcanzarlo, pero inicialmente no está segura de cuál es. En el capítulo 23, se explican los métodos de aprendizaje de refuerzo inverso que permiten a las máquinas aprender más sobre las preferencias humanas a partir de observaciones de las elecciones que hacen los seres humanos. En el capítulo 28, se exploran dos de las principales dificultades: primero, que nuestras elecciones dependen de nuestras preferencias a través de una arquitectura cognitiva muy compleja que es difícil de invertir; y segundo, que los seres humanos pueden no tener preferencias consistentes, ya sea individualmente o como grupo, por lo que puede no estar claro qué deberían estar haciendo los sistemas de IA para nosotros.

3 Agentes inteligentes

Este capítulo contiene temas de la Unidad 1 de IAR. Antes de continuar se debe leer:

- Capítulo 2 completo de [RNR04].

3.1 La naturaleza del entorno

En la sección 2.3 de [RNR04] se mencionan y explican las propiedades de los entornos (ambientes) donde se ejecutan los agentes. En [RN21] se agrega uno más:

Conocido vs. desconocido: Estrictamente hablando, esta distinción no se refiere al ambiente en sí, sino al estado de conocimiento del agente (o diseñador) sobre las "leyes de la física" del ambiente. En un ambiente conocido, se conocen los resultados (o las probabilidades de resultado si el ambiente es no determinista) para todas las acciones. Obviamente, si el ambiente es desconocido, el agente tendrá que aprender cómo funciona para tomar buenas decisiones.

La distinción entre ambientes conocidos y desconocidos no es la misma que la entre ambientes completamente y parcialmente observables. Es bastante posible que un ambiente conocido sea parcialmente observable, por ejemplo, en el juego de cartas *el solitario*, se conocen las reglas pero no se pueden ver las cartas que aún no han sido volteadas. Por el contrario, un ambiente desconocido puede ser completamente observable, en un nuevo videojuego por ejemplo, la pantalla puede mostrar todo el estado del juego, pero, hasta que no se los prueba, no se sabe qué hacen los botones (controles).

Como se señaló anteriormente, la medida de rendimiento en sí misma puede ser desconocida, ya sea porque el diseñador no está seguro de cómo escribirla correctamente o porque el usuario final, cuyas preferencias importan, no se conoce. Por ejemplo, un taxista generalmente no sabrá si un nuevo pasajero prefiere un viaje relajado o rápido, un estilo de conducción cauteloso o agresivo. Un asistente personal virtual comienza sin saber nada sobre las preferencias personales de su nuevo propietario. En estos casos, el agente puede aprender más sobre la medida de rendimiento en función de más interacciones con el diseñador o el usuario. Esto, a su vez, sugiere que el entorno de tarea debe

ser necesariamente visto como un entorno multiagente.

Teniendo en cuenta esta nueva dimensión, actualizamos el caso más difícil propuesto en [RNR04] a: ambiente parcialmente observable, multiagente, no determinista, secuencial, dinámico, continuo y *desconocido*. Conducir un taxi es difícil en todos estos sentidos, excepto en que el entorno del conductor es en su mayoría conocido. Conducir un coche alquilado en un país nuevo con geografía desconocida, leyes de tráfico diferentes y pasajeros nerviosos es mucho más emocionante.

4 Búsqueda

Este capítulo contiene temas de la Unidad 2 de IAR. Antes de continuar se debe leer:

- Secciones 3.1 y 3.3 de [RNR04].
- Sección 3.4 de [RNR04] hasta "Búsqueda de profundidad limitada" incluida. Ver también el cuadro de la figura 3.17.
- Sección 3.5 de [RNR04]
- Sección 4.1 de [RNR04] hasta "Búsqueda A*" incluida.
- Sección 4.2 de [RNR04]

4.1 Algunas aclaraciones

En el segundo párrafo del capítulo 3 de [RNR04], la frase "Los algoritmos son no informados, en el sentido que no **dan** información sobre el problema salvo su definición" debería decir "Los algoritmos son no informados, en el sentido que no **usan** información sobre el problema salvo su definición".

En la sección 3.3 de [RNR04] se define el factor de ramificación como "el máximo número de sucesores de cualquier nodo", pero debería decir "la cantidad media de sucesores".

4.2 Heurística

La palabra "heurística" se deriva del verbo griego *heuriskein*, que significa "encontrar" o "descubrir".

En el contexto de búsqueda en IA, una heurística es una técnica que aumenta la

eficiencia de un proceso de búsqueda, posiblemente sacrificando demandas de completitud [RKC94]. Se implementa como una función que recibe un estado del problema y devuelve una estimación del *grado de bondad* de dicho estado. En términos generales las funciones heurísticas no garantizan la solución óptima, pero frecuentemente ayudar a llegar a una buena solución.

5 Agentes lógicos

Este capítulo contiene temas de la Unidad 3 de IAR. Se debe leer:

- Secciones 7.1 a 7.5 de [RNR04].
- Sección 8.1 a 8.3 de [RNR04].

5.1 Algunas aclaraciones

En la sección 7.2 de [RNR04] (pág. 222), donde dice "por ejemplo, si el agente percibe un mal hedor **o** una pequeña brisa, ..." debería decir "por ejemplo, si el agente percibe un mal hedor **y** una pequeña brisa, ...".

En la sección 7.4 de [RNR04] (pág. 230) dice que se explica la notación BNF en la página 984, pero en realidad se lo hace en la 1117.

En la sección 7.4 de [RNR04] (pág. 236), donde dice "Una sentencia es **satisfactoria** si es verdadera para algún modelo." debería decir "Una sentencia es **satisfacible** si es verdadera para algún modelo".

En la sección 7.5 de [RNR04] (pág. 238), donde dice "Esta propiedad de los sistemas lógicos en realidad proviene de una característica mucho más fundamental, denominada **monótono**" debería decir "Esta propiedad de los sistemas lógicos en realidad proviene de una característica mucho más fundamental, denominada **monotonicidad**".

6 Planificación

Seguimos con la Unidad 3, Razonamiento en Ambientes Deterministas 2.

Este capítulo contiene el tema *planificación*. Vamos a ver planificación clásica, donde los ambientes son deterministas, completamente observables, finitos, estáticos y discretos. Conceptualmente, el tema es más cercano a búsqueda que a lógica, pero lo vemos en este momento porque es conveniente definir primero los predicados de la lógica de primer orden.

Se va a seguir el enfoque del libro [RKC94]. Se debe leer:

- Capítulo 11 hasta la página 439 de [RNR04].
- Capítulo 13 hasta la página 379 de [RKC94].

En casos de conflicto entre los libros, usar las definiciones y nomenclatura de [RKC94].

7 Lógica difusa

Comenzamos con la Unidad 4, Razonamiento bajo incertidumbre.

Cuando hablamos de incertidumbre en el contexto de la inteligencia artificial, *generalmente* nos referimos a situaciones donde el agente no tiene acceso a toda la verdad sobre su ambiente. Entonces, se puede hablar de que los eventos son verdaderos o falsos con cierta probabilidad. Más adelante, en otros capítulos, vamos a ver métodos que se ajustan a esa descripción.

Este capítulo contiene el tema *lógica difusa*. Es frecuente incluir a la lógica difusa en el conjunto de métodos del razonamiento bajo incertidumbre, aunque hay una diferencia con lo explicado en el párrafo anterior. En lógica difusa los eventos son verdaderos o falsos en cierto *grado*. Los agentes que utilizan lógica difusa toman decisiones en base a reglas definidas de forma difusa, donde se hace afirmaciones sobre la ocurrencia de ciertos eventos, pero estas afirmaciones se ven afectadas por adjetivos que modifican el grado de certeza.

Ninguno de los libros utilizados trata el tema en detalle, por lo tanto se utilizarán como material de lectura el apunte [Des20a] y las presentaciones que se subieron al aula virtual.

8 Aprendizaje automático

Este capítulo contiene los temas de la Unidad 5.

La idea del aprendizaje consiste en utilizar las percepciones no sólo para actuar, sino también para mejorar la habilidad del agente para actuar en el futuro. El aprendizaje entra en juego cuando el agente observa sus interacciones con el mundo y sus procesos de toma de decisiones [RNR04].

Antes de continuar leer las secciones 18.1 y 18.2 de [RNR04]. En [RNR04] se llama *hipótesis* a la función que se utiliza para aproximar $f(\boldsymbol{x})$. En la sección 18.2 el libro dice que entre múltiples hipótesis consistentes, la mejor es la más simple. Esto es correcto, pero es importante resaltar que la razón principal para esta elección es que la función más simple es la que mejor generaliza y que las funciones complejas tienden a sobeajustarse a los datos de entrenamiento.

8.1 Reconocimiento de patrones

Como enfoque complementario al de los agentes inteligentes, el aprendizaje automático se puede ver desde la óptica del reconocimiento de patrones. Esta última forma es más cercana a un proyecto típico de Ciencia de Datos.

Leer las etapas de sensado, extracción de características y clasificación de [Des20b], que se encuentra en la UV.

8.2 Aprendizaje supervisado

La idea general del aprendizaje supervisado fue explicada en la sección 18.2 de [RNR04]. Para desarrollarla en detalle utilizaremos la salida escalar $y \in \mathbb{R}$ y un vector de entradas $\boldsymbol{x} \in \mathbb{R}^n$ formado por n valores reales. El resto del capítulo está basado en el libro [Jam+21]. En el mismo libro se pueden ver ejemplos y explicaciones más extensas.

Vamos a asumir que existe una relación entre y y $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ que puede ser escrita de la forma general

$$y = f(\boldsymbol{x}) + \epsilon. \tag{8.1}$$

f es alguna función fija pero desconocida de $x_1, x_2, \ldots, x_n$, y ϵ es un término de error aleatorio, que es independiente de $\boldsymbol{x}$ y tiene media cero. En esta formulación, f representa la información *sistemática* que $\boldsymbol{x}$ provee sobre y.

En esecia, aprendizaje automático se refiere a un conjunto de enfoques para estimar f. La motivación más común para estimar f es la predicción. En muchas situaciones, un conjunto de entradas $\boldsymbol{x}$ está disponible, pero la salida no puede ser obtenida fácilmente. En este caso, como el término de error se promedia a cero, podemos predecir y usando

$$\hat{y} = \hat{f}(\boldsymbol{x})$$

donde $\hat{f}$ es la estimación de f y $\hat{y}$ es la predicción resultante para y.

Exploraremos varios enfoques para estimar f. Estos métodos generalmente comparten ciertas características. Durante el resto de esta sección se presenta una descripción de algunas de las características compartidas.

8.2.1 Error reducible y error irreducible

La exactitud (*accuracy* en inglés) de $\hat{y}$ como predicción de y depende de dos cantidades, a las cuales vamos a llamar *error reducible* y *error irreducible*. En general, $\hat{f}$ no será un estimador perfecto de f, y esta falta de exactitud introducirá algún error. Este error es *reducible* porque potencialmente podríamos mejorar la exactitud de $\hat{f}$ utilizando la técnica de aprendizaje automático más apropiada para estimar f. Sin embargo, incluso si fuera posible obtener una estimación perfecta para f, de modo que nuestra respuesta estimada tomara la forma $\hat{y} = f(\boldsymbol{x})$, la predicción aún tendría algún error. Esto se debe a que y también es una función de ϵ que, por definición, no se puede predecir usando $\boldsymbol{x}$. Por lo tanto, la variabilidad asociada con ϵ también afecta la exactitud de nuestras predicciones. Esto se conoce como el error *irreducible*, porque sin importar qué tan bien estimemos f, no podemos reducir el error introducido por ϵ.

¿Por qué el error irreducible es mayor que cero? La cantidad ϵ puede contener variables no medidas que son útiles para predecir y y/o variaciones no medibles, por ejemplo, derivadas de la subjetividad un evaluador humano.

8.2.2 Datos de entrenamiento

Asumiremos siempre que hemos observado (medido) un conjunto de m diferentes puntos o instancias de los datos de interés. Este conjunto se llama *datos de entrenamiento* porque los usaremos para entrenar (ajustar) nuestro método en la estimación de f. Sea x_{ij} la representación del valor del j-ésimo predictor, o entrada, para la observación i, donde $i = 1, 2, \ldots, m$ y $j = 1, 2, \ldots, n$. Conrrespondientemente, y_i será la representación de la variable objetivo para la i-ésima observación. Entonces, nuestro conjunto de datos de entrenamiento consiste en $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$, donde $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$.

Nuestro trabajo es aplicar un método de aprendizaje automático a los datos de entrenamiento para estimar la función desconocida f. En otras palabras, queremos encontrar una función $\hat{f}$ tal que $y \approx \hat{f}(\boldsymbol{x})$ para cualquier observación $(\boldsymbol{x}, y)$. En términos generales, la mayoría de los métodos de aprendizaje automático para esta tarea se pueden dividir en paramétricos o no paramétricos. Durante el dicatado de la materia nos enfocaremos en los métodos paramétricos.

8.2.3 Métodos paramétricos

Los métodos paramétricos implican un enfoque basado en modelos que consta de dos pasos.

1. Primero, hacemos una suposición sobre la forma de la función f. Por ejemplo, un caso muy simple es asumir que f es lineal en $\boldsymbol{x}$:

$$f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + w_{n+1} \tag{8.2}$$

   Este es un *modelo lineal*, el cual será tratado más adelante. Una vez que asumimos que f es lineal, el problema de estimarla se simplifica mucho. En lugar de tener que estimar completamente una función $f(\boldsymbol{x})$ arbitraria de dimensión n, solo es necesario estimar los $n + 1$ coeficientes (parámetros) $w_1, w_2, \ldots, w_{n+1}$.

2. Después de haber elegido un modelo, necesitamos un procedimiento que use los datos de entrenamiento para *entrene* o *ajuste* el modelo. En el caso del modelo lineal (8.2), necesitamos estimar los parámetros $w_1, w_2, \ldots, w_{n+1}$. Es decir, vamos a buscar valores para esos parámetros tales que

$$y \approx w_1 x_1 + w_2 x_2 + \cdots + w_{n+1} x_{n+1}$$

El método más común para ajustar el modelo 8.2 es conocido como *mínimos cuadrados* y será discutido más adelante.

El enfoque basado en modelos que se acaba de describir se conoce como paramétrico porque reduce el problema de estimar f a la estimación de un conjunto de parámetros. Asumir una forma paramétrica para f simplifica el problema de estimar f porque generalmente es mucho más fácil estimar un conjunto de parámetros, como $w_1, w_2, \ldots, w_{n+1}$ en el modelo lineal (8.2), que ajustar una función f completamente arbitraria. La desventaja potencial de del enfoque paramétrico es que el modelo que elegimos por lo general no coincidirá con la verdadera forma (desconocida) de f. Si el modelo elegido está demasiado lejos de la f verdadera, entonces nuestra estimación será pobre. Podemos tratar de abordar este problema eligiendo modelos flexibles para f que se ajusten a muchas formas funcionales diferentes. Pero en general, ajustar un modelo más flexible requiere estimar un mayor número de parámetros. Estos modelos más complejos pueden conducir a un fenómeno conocido como *sobreajuste*, que será tratado más adelante.

8.3 Evaluación del rendimiento de los modelos

Para evaluar el rendimiento de un método de aprendizaje automático sobre un conjunto de datos dado, necesitamos alguna forma de medir qué tanto sus predicciones coinciden con los datos realmente observados. Es decir, necesitamos cuantificar hasta qué punto el valor pronosticado para una observación determinada se acerca al valor de respuesta real para esa observación. En el caso de la regresión (predicción de un valor continuo, no clasificación), la medida más utilizada es el *error cuadrático medio* (MSE por su sigla en inglés), dado por

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{f}(\boldsymbol{x}_i))^2, \tag{8.3}$$

donde $\hat{f}(\boldsymbol{x}_i)$ es la predicción que $\hat{f}$ devuelve para la i-ésima observación. El MSE será pequeño si las predicciones son muy cercanas a las respuestas reales, y será grande si para algunas observaciones la predicción y la respuesta verdadera difieren sustancialmente.

El MSE en la ecuación 8.3 se expresó usando los datos de entrenamiento, por lo tanto se lo suele llamar *MSE de entrenamiento*. En general, realmente no nos importa qué tan bien trabaja el método con los datos de entrenamiento. *Lo que nos interesa es la exactitud de las predicciones que obtenemos cuando aplicamos nuestro método a datos de test que no han sido vistos previamente.* Para un conjunto de métodos de aprendizaje automático, queremos elegir el método que devuelva el menor *MSE de test*, en oposición al MSE de entrenamiento.

¿Cómo podemos hacer para elegir el método que minimiza el MSE de test? En algunos casos podemos tener un conjunto de datos de test disponible, es decir, un conjunto de observaciones que no fueron usadas durante el entrenamiento. En esos casos, podemos calcular el MSE de test para cada método y elegir el que muestre mejor rendimiento. ¿Qué pasa si no hay disponibilidad de datos de test? En ese caso uno podría pensar en elegir simplemente el método que minimiza el MSE de entrenamiento. Esto suena como un enfoque sensato, ya que el MSE de entrenamiento y el MSE de prueba parecen estar estrechamente relacionados. Desafortunadamente, hay un problema fundamental con esta estrategia, no hay garantía de que el método con menor MSE de entrenamiento vaya a tener también el menor MSE de test. En términos generales, el problema es que muchos métodos de aprendizaje automático ajustan los parámetros para minimizar *específicamente* el MSE de entrenamiento. Para estos métodos, el MSE de entrenamiento puede ser bastante pequeño, pero el MSE de test suele ser mucho mayor. La figura 8.1 muestra este fenómeno con un ejemplo simple.

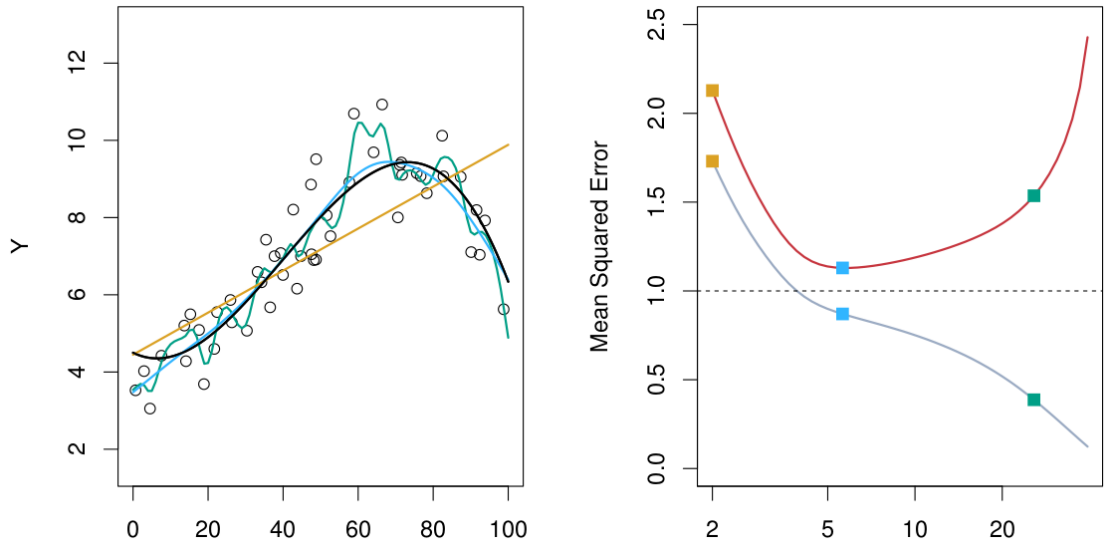

Figura 8.1: Izquierda: Datos simulados para f, en negro. Tres estimaciones de f: recta de regresión lineal (naranja), y dos splines suavizadas (curvas azul y verde). Derecha: MSE de entrenamiento (gris), MSE de test (rojo), y mínimo MSE de test posible para todos los métodos (línea de puntos). Los cuadrados representan los MSE de entrenamiento y test para los tres ajustes mostrados en el panel de la izquierda. Fuente: [Jam+21].

En el panel izquierdo de la figura 8.1 se han generado observaciones para la ecuación 8.1 con el valor real de f representado por la línea negra. Las curvas naranja, azul y verde muestran tres posibles estimaciones para f obtenidas usando métodos con niveles crecientes de flexibilidad. La línea naranja es el ajuste de regresión lineal, el cual es relativamente inflexible. Las curvas azul y verde fueron producidas usando splines con diferentes niveles de suavizado. Es claro que a medida que la flexibilidad se incrementa

las curvas se ajustan mejor a los datos observados (circulitos). La curva verde es la más flexible y se ajusta muy bien con los datos, sin embargo, observamos que se ajusta mal a la verdadera f (mostrada en negro) porque es demasiado ondulada. Ajustando el nivel de flexibilidad del ajuste de spline suavizado, podemos producir muchos ajustes diferentes a estos datos.

Ahora nos movemos al panel derecho de la figura 8.1. La curva gris muestra el MSE medio de entrenamiento en función de la flexibilidad, o más formalmente, los grados de libertad, para un número de splines suavizados. El grado de libertad es la cantidad que resume la flexibilidad de una curva. Los cuadraditos naranja, azul y verde indican el MSE asociado a la curva de color correspondiente en el panel izquierdo. Una curva más restringida y, por lo tanto, más suave tiene menos grados de libertad que una curva ondulada. En la figura 8.1, la regresión lineal está en el extremo más restrictivo, con dos grados de libertad. El MSE de entrenamiento disminuye monótonamente a medida que aumenta la flexibilidad. En este ejemplo f es no lineal, entonces el ajuste lineal (naranja) no es suficientemente flexible para estimar f correctamente. La curva verde tiene el menor MSE de entrenamiento porque corresponde al más flexible de los tres métodos.

En este ejemplo conocemos la verdadera función f, entonces podemos también calcular el MSE de test en función de la flexibilidad sobre un conjunto de test arbitrariamente grande (ovbiamente en general f es una función desconocida, entonces no podemos hacerlo). El MSE de test se muestra usando la curva roja en el panel de la dereacha de la figura 8.1. Al igual que el MSE de entrenamiento, el MSE de test inicialmente disminuye cuando la flexibilidad se incrementa. Sin embargo, en algún momento el MSE de test se nivela y después empieza a aumentar nuevamente. En consecuencia, las curvas naranja y verde tienen un alto MSE de test. La curva azul minimiza el MSE de test, lo que no debería sorprendernos, ya que visualmente parece ser la que mejor estima f en el panel de la izquierda. La línea segmentada horizontal indica $Var(\epsilon)$, el error irreducible, que corresponde al mínimo MSE de test posible para todos los métodos. Entonces, el spline suavizado representado por la curva azul es cercano al óptimo.

En el panel derecho de la figura 8.1, cuando la flexibilidad se incrementa, se observa una reducción monótona en el MSE de entrenamiento y una forma de "U" en el MSE de test. Esta es una propiedad fundamental del aprendizaje automático que se mantiene independientemente de los datos y el método elegido. A medida que aumenta la flexibilidad del modelo, el MSE de entrenamiento aumentará, pero es posible que el MSE de test no lo haga. **Cuando un método dado produce un MSE de entrenamiento pequeño pero un MSE de test grande, decimos que estamos *sobreajustando***. Esto pasas porque nuestro método de aprendizaje automático se está esforzando para encontrar patrones en los datos de entrenamiento, y puede estar detectando algunos patrones causados por el azar en lugar de verdaderas propiedades de la función desconocida f. Cuando sobreajustamos, el MSE de test es muy grande porque los supuestos patrones que el método encontró en los datos de entrenamiento simplemente no existen en los

datos de test.

Hay que tener en cuenta que independientemente de si se ha producido un sobreajuste o no, casi siempre el MSE de entrenamiento sea más pequeño que el MSE de test porque la mayoría de los métodos de aprendizaje automático, ya sea directa o indirectamente, buscan minimizar el MSE de entrenamiento. El sobreajuste se refiere específicamente al caso en el que un modelo menos flexible habría producido un MSE de test más pequeño.

Referencias

[But63] Samuel Butler. *Darwin Among The Machines*. E. P. Dutton & Company, 1863.

[Wie61] N. Wiener. *Cybernetics: or Control and Communication in the Animal and the Machine*. MIT Press, 1961.

[RKC94] E. Rich, K. Knight, and P.A.G. Calero. *Inteligencia artificial*. McGraw-Hill, 1994. ISBN: 9788448118587.

[LeC+95] Yann LeCun et al. "Comparison of learning algorithms for handwritten digit recognition". In: *International conference on artificial neural networks*. Vol. 60. 1. Perth, Australia. 1995, pp. 53–60.

[Nil95] Nils J Nilsson. "Perspective on Artificial Intelligence: Present and Future". In: *ECAI*. 1995.

[Yar95] David Yarowsky. "Unsupervised word sense disambiguation rivaling supervised methods". In: *33rd annual meeting of the association for computational linguistics*. 1995, pp. 189–196.

[Tur96] Alan M Turing. *Computing machinery and intelligence*. Mind, 1996.

[BB01] Michele Banko and Eric Brill. "Scaling to very very large corpora for natural language disambiguation". In: *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*. 2001, pp. 26–33.

[Min04] Marvin Minsky. "A framework for representing knowledge". In: *In AAAI*. Vol. 4. 2004, pp. 1090–1094.

[RNR04] S.J. Russell, P. Norvig, and J.M.C. Rodrıguez. *Inteligencia artificial: un enfoque moderno. 2da Edición*. Colección de Inteligencia Artificial de Prentice Hall. Pearson Educación, 2004. ISBN: 9788420540030. URL: `https://books.google.com.ar/books?id=yZCVPwAACAAJ`.

[Thr+06] Sebastian Thrun et al. "Stanley: The robot that won the DARPA Grand Challenge". In: *Journal of field Robotics* 23.9 (2006), pp. 661–692.

[GP07] Ben Goertzel and Cassio Pennachin. *Artificial general intelligence*. Springer Science & Business Media, 2007.

[HE07] James Hays and Alexei A Efros. "Scene completion using millions of photographs". In: *ACM Transactions on Graphics (ToG)* 26.3 (2007), 4–es.

[McC07] John McCarthy. “What is Artificial Intelligence?” In: *Stanford University* (2007).

[Min07] Marvin Minsky. “Conversations on the Society of Mind”. In: *AI magazine* 28.1 (2007), pp. 25–40.

[Omo08] Stephen Omohundro. “The basic AI drives”. In: *Frontiers in Artificial Intelligence and Applications* 171 (2008), pp. 483–492.

[Yud08] Eliezer Yudkowsky. “Artificial intelligence as a positive and negative factor in global risk”. In: *Global catastrophic risks* (2008), pp. 303–324.

[BW09] Carole R Beal and Patrick H Winston. *Intelligent tutoring systems: using AI to improve training performance and the learning experience*. John Wiley & Sons, 2009.

[Den+09] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[Bos14] N. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[Gul+16] Varun Gulshan et al. “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs”. In: *Jama* 316.22 (2016), pp. 2402–2410.

[Sil+16] David Silver et al. “Mastering the game of Go with deep neural networks and tree search”. In: *nature* 529.7587 (2016), pp. 484–489.

[Est+17] Andre Esteva et al. “Dermatologist-level classification of skin cancer with deep neural networks”. In: *nature* 542.7639 (2017), pp. 115–118.

[KSH17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[Liu+17] Yun Liu et al. “Detecting cancer metastases on gigapixel pathology images”. In: *arXiv preprint arXiv:1703.02442* (2017).

[Sil+17] David Silver et al. “Mastering the game of go without human knowledge”. In: *nature* 550.7676 (2017), pp. 354–359.

[Tet17] Philip E.. Tetlock. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, 2017.

[For18] Martin Ford. *Architects of Intelligence: The truth about AI from the people building it*. Packt Publishing Ltd, 2018.

[Gra+18] Katja Grace et al. “When will AI exceed human performance? Evidence from AI experts”. In: *Journal of Artificial Intelligence Research* 62 (2018), pp. 729–754.

[GS18] Barbara J Grosz and Peter Stone. "A century-long commitment to assessing artificial intelligence and its impact on society". In: *Communications of the ACM* 61.12 (2018), pp. 68–73.

[Sil+18] David Silver et al. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". In: *Science* 362.6419 (2018), pp. 1140–1144.

[Ste+18] David F Steiner et al. "Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer". In: *The American journal of surgical pathology* 42.12 (2018), p. 1636.

[Din+19] Yiming Ding et al. "A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain". In: *Radiology* 290.2 (2019), pp. 456–464.

[Liu+19a] Xiaoxuan Liu et al. "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis". In: *The lancet digital health* 1.6 (2019), e271–e297.

[Liu+19b] Yun Liu et al. "Artificial intelligence–based breast cancer nodal metastasis detection: Insights into the black box for pathologists". In: *Archives of pathology & laboratory medicine* 143.7 (2019), pp. 859–868.

[Top19] Eric Topol. *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK, 2019.

[Des20a] Eduardo Destéfanis. *Inferencia y probabilidad. Lógica difusa.* Aula virtual de IAR, 2020.

[Des20b] Eduardo Destéfanis. *Reconocimiento de patrones.* Aula virtual de IAR, 2020.

[Liu+20] Yuan Liu et al. "A deep learning system for differential diagnosis of skin diseases". In: *Nature medicine* 26.6 (2020), pp. 900–908.

[Jam+21] G.M. James et al. *An Introduction to Statistical Learning: With Applications in R*. Springer texts in statistics. Springer, 2021. ISBN: 9781071614181. URL: https://books.google.com.ar/books?id=5dQ6EAAAQBAJ.

[RN21] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach, Global Edition*. Pearson Education, 2021. ISBN: 9781292401171. URL: https://books.google.com.ar/books?id=cb0qEAAAQBAJ.

[Sto+22] Peter Stone et al. "Artificial intelligence and life in 2030: the one hundred year study on artificial intelligence". In: *arXiv preprint arXiv:2211.06318* (2022).