

Disclaimer: Notebooks translated to Spanish are distributed as an optional aid to assist in your learning and comprehension. We make no guarantees that the translations are completely accurate nor that the translated code blocks will run properly.

Introducción a Ciencia de Datos, Machine Learning e Inteligencia Artificial

Objetivo (1 min)

Al final de este caso, esperamos haber definido adecuadamente los términos “Data Science” (o Ciencia de Datos), “Machine Learning” (o Aprendizaje Automático) e Inteligencia Artificial. Usted debería ser capaz de identificar ejemplos de cada uno de éstos en el mundo real.

Introducción (5 min)

¿Qué es Ciencia de Datos? ¿Machine Learning? ¿Inteligencia artificial? Todos estos términos se usan indistintamente en nuestra sociedad, como si significaran la mismo, pero en realidad son bastante diferentes. Sin embargo, antes de que podamos responder esas preguntas, es necesario abordar una pregunta mucho más fundamental: ¿Qué son los datos? Sorprendentemente, es bastante difícil encontrar una definición simple para ello.

Pregunta: (5 min)

¿Qué son los datos?

La respuesta a esta pregunta podría buscarse en Wikipedia o libros de texto, pero la mayoría de esas fuentes le darán una definición muy pedante y nada intuitiva. En cambio, vamos a seguir con la definición coloquial de los datos como “algo cuyo valor a usted le interesa”. No lo encontrará en ningún tratado formal del tema, pero por ahora, esta definición es suficiente para ilustrar qué son los datos. Su nombre, la edad y el número de teléfono son datos sobre usted. Los valores en su cuenta de ahorro, su dirección y los nombres de sus padres son datos que se relacionan con usted. Tenemos datos sobre todo, en todas partes.

Qué es Ciencia de Datos? (10 min)

Ahora que sabemos qué son los datos, es posible preguntar: “¿Qué es ciencia de datos?” Ciencia, en el lenguaje del [método científico](#), es:

1. Formulación de hipótesis o conjeturas sobre cómo funciona el mundo, basadas en observaciones del entorno que nos rodea.
2. Validar o invalidar esas hipótesis realizando experimentos.

Sin embargo, a diferencia de las ciencias puras, trabajar con datos no requiere necesariamente realizar experimentos (¡aunque podría ser!). Por el contrario, muchas veces los datos ya han sido recopilados y organizados por otra persona. Entonces, el método científico, tal como se aplica a los datos, puede resumirse como: **“Formular hipótesis basadas en el mundo que nos rodea, luego analizar datos relevantes para validar o invalidar nuestras hipótesis.”**

Ejercicio 1: (5 min)

Con base en lo expuesto anteriormente, ¿cuál de los siguientes enunciados puede describirse como un reflejo de todo el proceso de Ciencia de Datos? Seleccione todas las opciones que correspondan.

- (a) Se evidencia anecdóticamente que los millennials parecen responder más positivamente a las discusiones sobre la nueva versión del producto de su empresa que está en versión beta versus la existente, se implementa un proceso A / B testing que direcciona a los millenials por igual a ambas versiones, y

luego se realiza pruebas de significancia estadística en estos datos para verificar que prefieren la nueva versión.

- (b) Observar que los precios de Uber parecen estar correlacionados con un pequeño conjunto de factores, obtener datos abiertos sobre las tarifas de precios de Uber, luego construir un modelo de precios basado en esos factores y verificar que explican la mayor parte de la variación en las tarifas.
- (c) Convertir imágenes de círculos en píxeles estructurados y almacenarlos en una base de datos para su uso posterior.
- (d) Construir un algoritmo que permita que una computadora reconozca imágenes de perros y gatos.

Respuesta.

Tenga en cuenta que Ciencia de Datos **NO** es lo que a menudo aparece en los medios:

1. **NO** es el hecho de que un computador reconozca imágenes de perros y gatos
2. **NO** es IBM Watson detectando enfermedades en tejidos humanos
3. **NO** es AlphaGo ganándole al mejor jugador de Go del mundo

De hecho, la gran mayoría de los casos de uso de datos **NO son** como los tres ejemplos anteriores, sino mucho más similares a las ciencias tradicionales. La mayoría de los ejemplos de Ciencia de Datos son los que describimos en las opciones (a) y (b) en el Ejercicio 1. Un ejemplo común de Ciencia de Datos corporativa es una empresa que recolecta datos de usuarios, los analiza para clasificar a sus clientes, crea campañas de marketing para apuntar mejor a clientes potenciales, probando los efectos de las soluciones y actualizando sus materiales de marketing en consecuencia.

Qué es machine learning? (10 min)

De cualquier manera, la opción (d) del Ejercicio 1, al igual que los ejemplos 1 y 2 de la lista anterior, cae dentro de lo que puede considerarse como **machine learning**. ¿Qué significa esto?

“Aprender” (“Learn” en inglés) significa “obtener o adquirir conocimiento o habilidad en algo a través de la experiencia”. Entonces uno podría definir “machine learning” como “la manera en que una máquina obtiene o adquiere conocimiento a través de la experiencia”. ¿Cómo gana experiencia una máquina? Todas las entradas de la máquina son esencialmente cadenas binarias de 0s y 1s, lo que realmente es: ¡lo adivinaste, datos! Entonces, el aprendizaje automático es realmente **la manera en que una computadora adquiere conocimiento a través de los datos**.

Por supuesto, esto no da una idea del “cómo” en absoluto; solo dice que hay algo que se hace con los datos de entrada para generar este conocimiento como salida. Para hacer una analogía matemática, el aprendizaje automático es una función f tal que

$$\text{\$ conocimiento} = f(\text{datos})\text{\$}$$

y aparte de eso no hay otras estipulaciones reales sobre f ! así que f podría ser tan mecánica como una simple función matemática (por ejemplo, la suma de todos los puntos de datos) y calificarse como aprendizaje automático. Y en la práctica, esto es lo que son la mayoría de los algoritmos comunes de aprendizaje automático, incluidos:

1. Regresión logística
2. Random forests
3. Support vector machines
4. k - means clustering
5. Redes neuronales

(Usted aprenderá sobre esto más tarde en el curso.) Esto puede parecer decepcionante, dada la forma en que los medios promocionan la “inteligencia artificial” y hace que parezca que hay algo “inteligente” en las técnicas de machine learning, pero de hecho, muchos métodos mecánicos satisfacen las condiciones requeridas para ser clasificado como machine learning. Esto no significa que estos métodos mecánicos tengan una utilidad limitada; de hecho, son bastante potentes si se usan correctamente, pero sí significa que no se parecen a nada que naturalmente asociaríamos con inteligencia similar a la humana.

Más específicamente, el aprendizaje automático es “mecánico” en el sentido de que la forma en que estos algoritmos “aprenden” se basa estrictamente en principios matemáticos. Por ejemplo, la regresión lineal es un algoritmo que aprende ajustando los coeficientes de los datos de entrada para predecir mejor un valor de salida. La forma en que cambian los coeficientes se basa completamente en protocolos matemáticos (en este caso, los gradientes de los datos de entrada). Una aplicación común de regresión lineal sería predecir los precios de la vivienda en función de diversos datos de entrada, como el tamaño, el número de habitaciones y la antigüedad de la casa. El modelo tomaría los datos y aprendería de ellos eligiendo el conjunto de coeficientes que minimiza el error de sus predicciones frente a los precios reales.

Ejercicio 2: (5 min)

Según la definición anterior, ¿cuál de las siguientes tareas probablemente implicaría el uso de “machine learning”? Seleccione todas las que apliquen.

- (a) Construir el modelo que soporta el sistema de reconocimiento facial de los iPhone.
- (b) Construir el modelo que soporta el sistema de recomendación de películas de Netflix basado en lo que sus usuarios han visto anteriormente.
- (c) Investigar factores que afectan los precios de Airbnb y desarrollar una herramienta de precios basada en este análisis.
- (d) Desarrollar un sistema automatizado para aprobar o rechazar solicitudes de préstamos hipotecarios.

Respuesta.

Machine learning puede ser, y por lo general es, una parte propia de Data Science o la Ciencia de Datos. **Ciencia de Datos es fundamentalmente un proceso, mientras que Machine Learning es una herramienta que puede ser inmensamente útil para llevar a cabo el proceso de Ciencia de Datos.**

Qué es Inteligencia Artificial? (25 min)

Pero aún no hemos llegado al fondo del asunto: aunque algunos métodos mecánicos “sencillos” ó “tontos” se pueden calificar como machine learning, esto no excluye que los métodos “inteligentes” similares a los humanos sean clasificados como tales. Y aunque esto es completamente cierto, muchas personas han optado por nombrarlo de manera completamente distinta: **inteligencia artificial**.

¿Pero por qué? ¿Por qué dar a los métodos “inteligentes” un nombre completamente diferente si también pueden caer en la categoría de machine learning? Esa es la pregunta que exploraremos durante el resto de este módulo.

Comencemos por echar un vistazo a una demostración icónica de esta llamada inteligencia: [AlphaGo beating the world's top human Go player](#).

```
[2]: from IPython.display import IFrame
```

```
[3]: IFrame('https://www.youtube.com/embed/8tq1C8spV_g?rel=0&controls=0&↵showinfo=0', width=560, height=315)
```

```
[3]: <IPython.lib.display.IFrame at 0x7faf9d0b4208>
```

¡Muy impresionante! ¿Pero esta hazaña por sí sola prueba que una máquina exhibe inteligencia humana?

Pregunta: (5 min)

¿Qué, para usted, cuenta como “inteligencia artificial”? ¿Qué demostraciones de aptitud le convencerían, sin lugar a dudas, de que algo es tan inteligente como nosotros los humanos?

Es posible que haya descubierto que es bastante difícil encontrar respuestas a la segunda parte de la pregunta anterior, y que la mayoría de las ideas que ha tenido: 1) parece que podrían tener una solución mecánica “tonta” como en la discusión anterior de machine learning, a pesar de parecer impresionante al principio; o 2) realmente le hizo preguntarse cuán única es realmente la inteligencia humana y si prácticamente todo lo que hacemos podría reducirse a métodos mecánicos tan “tontos”.

¿Qué pueden y no pueden hacer las máquinas? Que hay de nosotros los humanos? (3 min)

Entonces, ¿hay alguna prueba sensata que podamos usar para determinar si algo es tan inteligente como un ser humano? Ha habido muchas propuestas a lo largo del tiempo. La prueba de aptitud más famosa desarrollada fue el **test de Turing**, nombrado en honor al matemático inglés y famoso criptógrafo de la Segunda Guerra Mundial Alan Turing. En este examen, hay un evaluador humano y dos compañeros de conversación: una máquina y un humano. El evaluador mantendría una conversación con cada uno a través de un canal de solo texto. Si el evaluador no puede distinguir confiablemente la máquina del humano, se dice que la máquina ha pasado la prueba.

Turing no declaró explícitamente que su prueba podría usarse como una medida de inteligencia, pero muchos de los que vinieron después de él pusieron su prueba en el centro de atención. Por supuesto, la implicación es que si una computadora puede conversar como un ser humano, entonces es efectivamente tan inteligente como un ser humano.

Pregunta: (5 min)

¿Cuáles son algunas deficiencias de esta propuesta? ¿Qué idea aporta esto a los criterios que una prueba rigurosa de inteligencia humana debería satisfacer?

Además de las fallas con la prueba de Turing (y de hecho con casi cualquier otra prueba que se pueda encontrar), esto trae a la luz una de las obsesiones poco saludables de nuestra sociedad cuando se trata del campo de la inteligencia artificial: su enfoque singular en imitar la inteligencia humana a través de máquinas. Pero, ¿qué pasa si la inteligencia artificial es fundamentalmente diferente (no peor ni mejor) que la inteligencia humana? ¿Qué pasa si las máquinas son más “inteligentes” que nosotros para algunas cosas, y viceversa?

Pregunta: (7 min)

Realice una lluvia de ideas con un compañero, y luego discutiremos con el resto de la clase:

1. ¿Cuáles cosas las máquinas ya pueden hacer mejor que nosotros? ¿Qué específicamente sobre ellas les permite hacerlo mejor?
2. ¿Cuáles cosas podemos hacer mejor que las máquinas? ¿Crees que esta ventaja probablemente se mantendrá con el tiempo? ¿Por qué?

Pregunta: (3 min)

Ahora que hemos definido Ciencia de Datos de manera adecuada, ¿Quién ha usado Ciencia de Datos en su trabajo? Por favor proporcione un ejemplo específico.

Conclusiones e ideas que nos debemos llevar (5 min)

En este caso, usted aprendió lo que realmente son “Ciencia de Datos” y “Machine Learning”, en contraste con las connotaciones engañosas que a menudo se dan en la discusión pública. También aprendió que “inteligencia artificial” es un término muy ambiguo: nadie está realmente de acuerdo con su definición exacta y no está claro si su enfoque actual en imitar la inteligencia humana es incluso el enfoque correcto.

A lo largo de este programa, nos centraremos principalmente en Data Science o ciencia de datos y Machine Learning o aprendizaje automático, y no tanto en la inteligencia artificial. Sin embargo, las preguntas filosóficas que rodean la inteligencia artificial son fascinantes, y lo alentamos a que continúe reflexionando sobre ellas a medida que se involucre más en este nuevo y emocionante campo. En las próximas semanas, haremos una introducción a varias técnicas (como Regresión Lineal, k - means clustering) y herramientas (como [pandas](#), AWS) de tal manera que pueda llevar a cabo procesos de Ciencia de Datos en su vida profesional.