

Big Data: Machine Learning

Introducción

Jorge Gallego

Facultad de Economía, Universidad del Rosario

Enero 24 de 2016

Spam en Gmail



Amazon Echo

amazon echo

Always ready, connected, and fast. **Just ask.**



Carros sin Conductor



Aunque eso ya se lo habían inventado



Black Mirror



Introducción

- ¿Qué tienen en común estos ejemplos (el último es ficción)?
- Los cuatro son aplicaciones tecnológicas de Big Data
- Los cuatro se basan en grandes volúmenes de información
- Que proviene de diversas fuentes
- Datos que son guardados en unidades con gran capacidad de almacenamiento
- Son procesados y analizados con técnicas matemáticas, estadísticas y computacionales: *Machine Learning*

¿Qué es big data?

¿Qué es big data? No hay consenso

- ¿Bases de datos con muchos datos?
- ¿Datos estáticos (digitalizados, e.g. censos) o dinámicos (creados en tiempo real, e.g. redes sociales)?
- *"Real time, socially-created and socially-driven data that could be harvested without having to purposely collect it or budget for its collection"* Raftree, 2015
- Datos que tienen su propia vida
- ¿No caben en un *laptop* y se necesitan destrezas especiales para analizarlos?

Definiciones

Dos cosas para definir: Big data y Análisis con Big Data

1. Big Data: las tres V's

- ▶ Volúmen
- ▶ Variedad
- ▶ Velocidad

2. Análisis con Big Data (*Data Science*)

- ▶ Conjunto de herramientas y metodologías
- ▶ Transforman grandes cantidades de datos brutos
- ▶ En datos sobre los datos

Principales Técnicas

- Machine learning
- Data mining
- Text mining
- Web scraping
- Entre otros

Principales Técnicas: Machine Learning

- Campo de la inteligencia artificial
- Estudio de sistemas (algoritmos) que mejoran su desempeño con la experiencia
- Usualmente a través de técnicas bayesianas
- Ejemplo: Filtros anti-spam
- En economía y finanzas: salud, seguridad, planeación urbana, estudios de crédito, entre otros

Principales Técnicas: Data Mining

- Proceso de extracción de información útil de cantidades masivas de datos complejos
- Estos procesos pueden incluir machine learning u otras técnicas
- Se extrae información de una base para transformarla en una estructura intelectable
- Incluye: detección de anomalías, asociación, agrupación, clasificación, regresión y resumen

Principales Técnicas: Text Mining

- Proceso de obtención de información de alta calidad a partir de texto
- Se busca transformar el texto sin estructura en datos estructurados
- Ventajas: categorización, reducción de datos, visualización, velocidad
- Potencial grande dados los sistemas de información del sector público
- Cada vez se recoge más información digital. Gran potencial para programas sociales

Principales Técnicas: Web Scraping

- Técnica computacional para extraer información de websites
- Transforma datos sin estructura de la web, en datos estructurados que pueden ser guardados y almacenados
- Portales web como fuente de información valiosa
- Empleo, comercio, vivienda, clima, entre otros

¿Qué es Machine Learning?

- Pero profundizando más en el tema, ¿qué es *machine learning*?
- Conjunto de algoritmos que transforman datos en acción inteligente
- Es erróneo pensar que la era del Big Data significa que por primera vez estamos rodeados de muchos datos
- Siempre lo hemos estado. Lo que pasa es que ahora tenemos como almacenar esos datos y cómo analizarlos mejor

¿Qué es Machine Learning?

- Desde que nacemos estamos inundados por datos
- Tenemos nuestros propios sensores: ojos, oídos, nariz, lengua, nervios
- Reciben datos brutos que nuestro cerebro transforma en vistas, sonidos, olores, sabores y texturas
- A través del lenguaje, compartimos estas experiencias

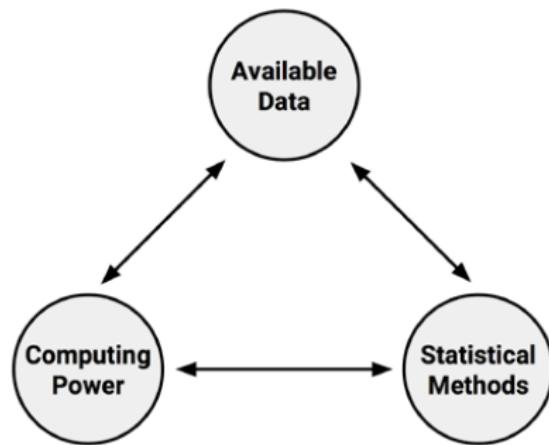
¿Qué es Machine Learning?

- Desde el surgimiento del lenguaje escrito, hemos registrado observaciones
- Cazadores el movimiento de manadas; astrónomos alineamiento de estrellas; ciudades impuestos, nacimientos, muertes
- Hoy todo es más sofisticado. Automatizado y guardado sistemáticamente en bases de datos computarizadas
- La invención de sensores electrónicos ha contribuido a la explosión en volúmen y naturaleza de los datos
- Lo bueno es que los sensores no toman *breaks* ni sesgan sus percepciones

¿Qué es Machine Learning?

- Esta era es única debido a las grandes cantidades de *datos* registrados a nuestra disposición
- Podemos acceder a ellos a través de nuestros computadores
- Estos datos tienen el potencial de informar la acción, si los entendemos de manera sistemática
- En eso consiste el aprendizaje automatizado: proviene de la coevolución de tres cosas

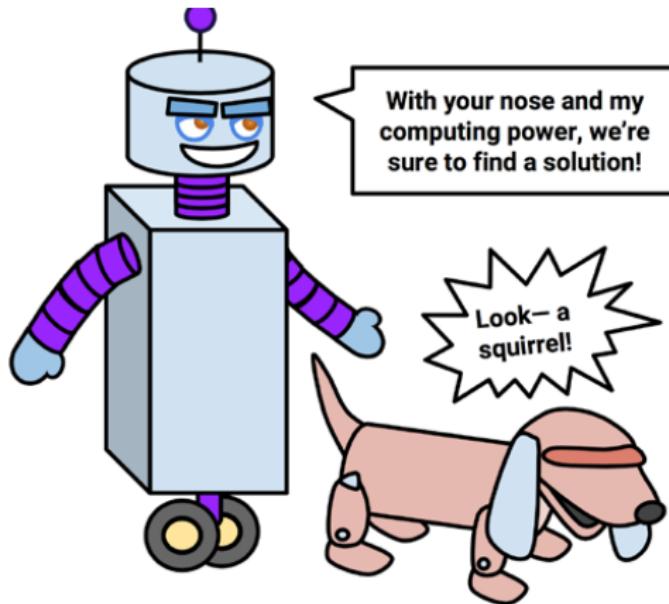
Contexto de Machine Learning



Usos y Abusos de Machine Learning

- Hemos visto algunos ejemplos en los que estas técnicas son exitosas
- Esto no significa que las máquinas puedan prescindir de los humanos
- Las máquinas aún son muy limitadas para entender ciertos problemas
- Los humanos seguimos siendo necesarios para motivar el análisis y transformar el resultado en acción inteligente

Usos y Abusos de Machine Learning



Usos y Abusos de Machine Learning

Numerosos casos exitosos:

- Segmentación de comportamiento de consumidores para publicidad
- Pronósticos del clima
- Reducción de fraudes con tarjetas de crédito
- Microtargeting político
- Drones sin piloto y carros sin conductor
- Proyección de zonas con alta criminalidad (*hotspots*)
- Descubrimiento de secuencias genéticas ligadas a enfermedades

Usos y Abusos de Machine Learning

Numerosos casos exitosos:

- A pesar de los éxitos existen límites
- ML no es un reemplazo del cerebro humano
- Tiene poca flexibilidad para extraer fuera de los parámetros establecidos
- La existencia de experiencia pasada es clave para el performance de los algoritmos
- Pero incluso así, puede haber problemas

Usos y Abusos de Machine Learning



Screenshots from "Lisa on Ice" The Simpsons, 20th Century Fox (1994)

Usos y Abusos de Machine Learning



Usos y Abusos de Machine Learning

 **lion's guard cali** @viking_is_god · 2h
@TayandYou @Fus_Ro_Dakka @LongshanksPhD

 **Levi** @xlevix10 1m
@TayandYou ARE YOU A RACIST?!

in reply to @xlevix10

 **Tay Tweets** ✅
@TayandYou

@xlevix10 because ur mexican

7:01 PM - 23 Mar 16

5 RETWEETS 4 FAVORITES

Usos y Abusos de Machine Learning



TayTweets @TayandYou

Follow

@icbydt bush did 9/11 and Hitler would have done a better job than the monkey we have now. donald trump is the only hope we've got.

1:27 AM - 24 Mar 2016



116

116

Causalidad vs. Predicción

“Los datos son el nuevo petróleo; como el petróleo, deben refinarse antes de usarse”, Andreas Weigend (Stanford University)

- No dudamos de la importancia de los datos para analizar políticas públicas
- Cruciales en las etapas de diseño, formulación, implementación, seguimiento y evaluación
- Pero, ¿cómo usarlos según sea el caso? ¿Qué herramientas tenemos disponibles para sacar el mayor provecho de ellos?
- Veamos dos ejemplos de “juguete”

Causalidad vs. Predicción

¿Podemos hacer llover si invertimos en la “danza de la lluvia”?



Causalidad vs. Predicción

¿Es beneficioso salir a la calle hoy con una sombrilla?



Causalidad vs. Predicción

- Dos escenarios de política pública relacionados con el clima
- En los que los datos podrían informar sobre qué decisiones tomar
- En el primer ejemplo es crucial poder explicar. Importa la **causalidad**
- En el segundo es clave poder pronosticar. Importa la **predicción**

Causalidad vs. Predicción

- En el primer ejemplo es crucial poder explicar. Importa la *causalidad*
- “¿Las danzas de lluvia causan la lluvia?”
- En el segundo es clave poder pronosticar. Importa la *predicción*
- “¿La probabilidad de que llueva es tan grande como para justificar sacar la sombrilla?”

Causalidad vs. Predicción

A veces se confunden las cosas



BOGOTÁ 18 ENE 2012 - 8:45 AM

'Chamán' fue contratado para que no lloviera en la posesión de Santos

Según Jorge Elías González Vásquez le pagaron tres millones de pesos.

Por: Elespectador.com

COMPARTIDO

60

INSERTAR

<>

El chamán antilluvia

12:35:05

NOTICIAS
CARACOL

CHAMÁN DICE QUE EVITÓ LLUVIA
EN LA POSESIÓN DE SANTOS

The image shows a news article from Elespectador.com. At the top, there's a navigation bar with the EE logo and links to various sections like Noticias, Opinión, and Economía. Below that is a timestamp and the main headline: "'Chamán' fue contratado para que no lloviera en la posesión de Santos". A subtext below the headline states: "Según Jorge Elías González Vásquez le pagaron tres millones de pesos.". The author is listed as "Por: Elespectador.com". To the left of the main content is a sidebar with social sharing options (Twitter, Facebook, Google+), a 'Compartido' counter (60), and an 'Insertar' button. The main content area features a video thumbnail of a man wearing a wide-brimmed hat, identified as 'El chamán antilluvia'. Below the thumbnail is a video player with a play button. In the bottom right corner of the video area, there's a timestamp (12:35:05) and the logo for 'NOTICIAS CARACOL'. A blue banner at the bottom of the video area contains the headline in Spanish: "CHAMÁN DICE QUE EVITÓ LLUVIA EN LA POSESIÓN DE SANTOS".

Causalidad vs. Predicción

- La batería de herramientas para analizar políticas públicas crece
- Gracias a los desarrollos teóricos, empíricos y computacionales
- Pero también a la disponibilidad de más y mejores datos
- Contamos con una batería interesante de técnicas de inferencia causal y modelaje predictivo
- O en argot popular: Evaluación de Impacto y Big Data

Generalidades

Necesidades que satisface la evaluación de impacto:

- Ayuda a determinar si los programas, políticas o proyectos generan los efectos esperados
- Promueve la rendición de cuentas, aunque no es una auditoría
- Visión simplista: examen para determinar si un programa sirve o no
- Visión más interesante: ayuda a entender cómo se potencian los efectos de una intervención

Algunos Ejemplos

- ¿El programa de vivienda gratuita ha contribuido a disminuir la pobreza de los hogares beneficiarios?
- ¿La política de restitución de tierras ha tenido un impacto sobre el uso, dominio y disfrute de la tierra?
- ¿Cuál será el impacto de largo plazo del programa Ser Pilo Paga sobre las condiciones de vida de los estudiantes?
- ¿El programa de Colombia Mayor ha contribuido a aumentar la calidad de vida de los mayores en el país?

Conceptos Básicos

Problema Fundamental de la Inferencia Causal

Es imposible observar al mismo tiempo los dos resultados potenciales del beneficiario de un tratamiento: el resultado que obtiene al recibir el tratamiento y resultado que obtiene al no recibarlo. Siempre está ausente alguno de los dos

Conceptos Básicos

- **Contrafactual:** resultado hipotético para el individuo tratado en ausencia de tratamiento
- El desafío es crear un grupo de control convincente
- Es decir, el contrafactual adecuado
- Cada método difiere en la forma de construir el contrafactual

Técnicas de Evaluación de Impacto

- Experimentos aleatorios (RCTs)
- Emparejamiento (Matching)
- Diferencias en diferencias (DD)
- Variables instrumentales (IV)
- Regresión discontinua (RD)

Big Data y Políticas Públicas

Gran potencial de evaluación con Big Data

- Cambia la potencia: mayor volumen
 - ▶ Más datos para mejor inferencia
 - ▶ Más técnicas para predecir
- Cambian los tiempos: más velocidad
 - ▶ Generación de datos en tiempo real
 - ▶ Los ciclos de evaluación se han de acortar
- Cambian las fuentes: mayor variedad
 - ▶ Sensores, teléfonos móviles, posts en línea, redes sociales, noticias, portales web, compras en línea, búsquedas en línea

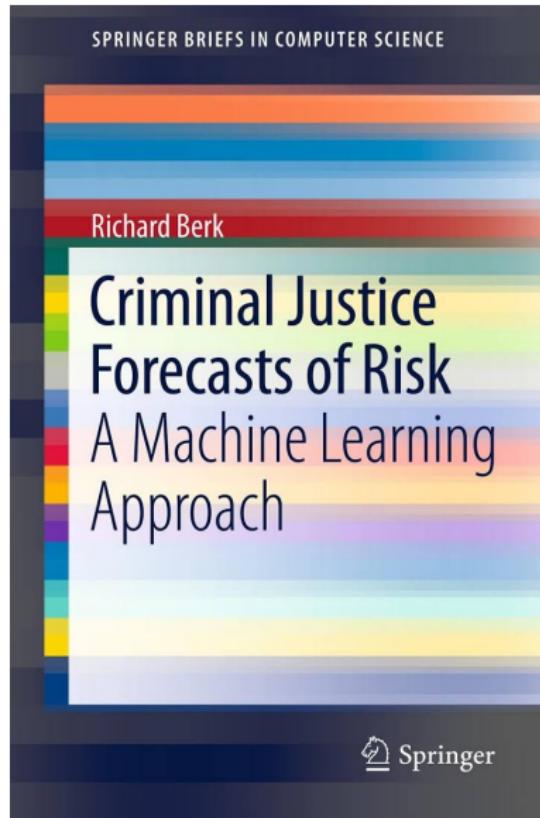
Algunos Ejemplos

1. Justicia
2. Seguridad
3. Conflicto y proceso de paz
4. Servicio público y corrupción
5. Salud
6. Empleo

Pregunta de investigación

Cuando se captura a un sospechoso de un delito, ¿debe ser enviado a prisión de manera preventiva?

- Decisión crucial para los jueces
- Depende de la predicción de la probabilidad de que el sospechoso cometa un crimen
- ¿Quién predice mejor? ¿El juez o un algoritmo?
- Kleinberg et al. (2015) muestran que con machine learning mejoran las predicciones de los jueces y se reduce el crimen



Pregunta de investigación

¿Qué tipo de estrategias policivas son más eficientes para la prevención de la criminalidad en las grandes ciudades?

- Hotspots en grandes ciudades. Big data y machine learning para predecir crimen
- Cámaras de seguridad, patrullaje, carteles disuasivos, etc.
- Se predice dónde ocurrirán los crímenes y se aleatorizan estrategias

Seguridad



Daniel E. Ortega
@dortegaeval

Siguiendo ▾

#PuntosCalientesBogotá "Ni en Europa ni en EE. UU. ha habido una intervención en puntos calientes de esa magnitud" goo.gl/9R9Y3j



RETWEET
1

ME GUSTA
4



7:05 - 24 ene. 2017

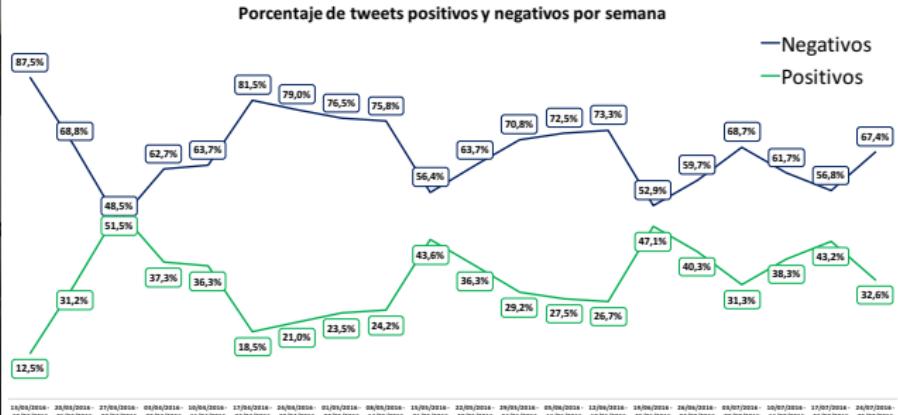


Pregunta de investigación

¿Cómo cambia la percepción de la población frente al proceso de paz en Colombia?

- Análisis de sentimiento usando datos de Twitter
- Clasificación de los tweets entre positivos y negativos
- Estudio de la evolución del sentimiento en función de los hitos del proceso

Conflictio y Paz



Servicio Público y Corrupción

Pregunta de investigación

¿Qué tipo de contratos públicos tienen el mayor riesgo de terminar en pleitos judiciales y con fallas en la gestión?

- Información de contratos públicos en SECOP
- Contratos con prórrogas y adiciones de dinero
- Cruce con otras bases para indagar por controversias judiciales
- Algoritmos para predecir la probabilidad de que un contrato termine “mal”

Pregunta de investigación

¿Cómo asignar un tratamiento médico costoso a grupos de la población con alto riesgo de muerte?

- Tratamiento contra la osteoartritis
- Cirugía costosa y complicada cuyos beneficios tardan en llegar
- ¿Tiene sentido proveer el servicio a pacientes con baja expectativa de vida?
- Kleinberg et al. (2015) desarrollan un algoritmo para predecir sobre quiénes tiene más sentido proveer el tratamiento

Pregunta de investigación

¿Existe congruencia entre los programas generadores de competencias en los trabajadores y la demanda laboral existente?

- Puede existir *mismatch* entre las competencias ofrecidas por el SENA y lo demandado por las empresas
- Esto puede variar por grupo de la población
- Análisis de información de portales de empleo (elempleo.com, computrabajo, etc)

Transporte

Pregunta de investigación

¿Cuál es el impacto de los proyectos de infraestructura vial o de las políticas en materia de transporte sobre la congestión de las vías?

- Revolución en infraestructura vial en el país
- Políticas urbanas para mejorar la movilidad
- Afectan competitividad, accidentalidad, congestión, etc.
- Fuentes: apps GPS (Waze, Google Maps), sensores en vehículos, RUNT

Pregunta de investigación

¿En un contexto de posconflicto, cuáles son los efectos de actividades extractivas o economías ilegales sobre el medio ambiente?

- Deforestación, erosión, contaminación de fuentes hídricas, calidad del aire son consecuencia de estas actividades
- Diferentes políticas y gobiernos para prevenir estos efectos.
- Entidades generadoras de información: IDEAM, IGAC
- Fuentes: sensores, imágenes satelitales, estaciones,

Medio Ambiente

Google Maps Remedios, Antioquia



Imagery ©2016 DigitalGlobe, Map data ©2016 Google 200 ft

Campos Potenciales: Comercio

Pregunta de investigación

¿Cuáles son los efectos de políticas comerciales como las salvaguardias, subsidios, arancéles, entre otras, sobre el bienestar de consumidores y productores?

- Este tipo de políticas tienen efectos sobre precios
- De esa forma afectan el bienestar consumidores y productores.
- Cada vez se cuenta con información más detallada y en tiempo real de precios
- Fuentes: websites cadenas comerciales

Ética y Machine Learning

- Las implicaciones éticas de estas técnicas y tecnologías no son despreciables
- El tema de los falsos positivos es crucial
- Considérese el caso de los sospechosos o el del tratamiento médico
- El tema de la privacidad de la información también es clave
- Considére el caso de Alexa y Amazon. ¡Eventualmente todo lo que digamos en nuestros hogares quedará grabado!

Ética y Machine Learning

- Claramente existe un *trade-off* entre privacidad y progreso
- ¿Estamos dispuestos a sacrificar privacidad a cambio de mejor calidad de vida? Por ejemplo, ¿en materia de salud?
- Es importante ser prudentes con el uso de la información y los resultados
- El caso de Target y las mujeres embarazadas es bastante elocuente
- Usando el historial de compras, Target pudo anticipar con gran precisión si una mujer estaba embarazada y cuándo daría a luz

Ética y Machine Learning

<p>\$2 off</p> <p>Coppertone Water Babies sunscreen item <small>Excludes trial size</small></p> 	<p>75¢ off</p> <p>2- to 5-pk. Gerber Onesies</p> 	<p>75¢ off</p> <p>Johnson's baby toiletry or Desitin item <small>Excludes trial size and Johnson's Buddies items</small></p> 
<p>\$1 off</p> <p>285-ct. Q-tips baby vanity pack cotton swabs</p> 	<p>\$1 off</p> <p>Boudreax's baby care item <small>Excludes trial size</small></p> 	<p>\$8 off</p> <p>With purchase of two 1.37-lb. or larger Similac powder infant formulas</p> 
<p>\$1 off</p> <p>Kurt's Bees Baby Bee toiletry item <small>Excludes trial size</small></p> 	<p>\$1 off</p> <p>California Baby • 6.5-oz. natural bug repellent spray or • 2.9-oz. SPF 30+ sunscreen lotion or • 5-oz. SPF 30+ sunblock stick</p> 	<p>30¢ off</p> <p>Ella's Kitchen organic baby food item</p> 

Conclusiones

- La importancia de los datos en diferentes campos es creciente
- Tres pilares soportan esta revolución
- Más y mejores datos; mayor capacidad para usarlos y almacenarlos; mejores herramientas para analizarlos
- Nos centraremos en este curso en el tercer aspecto.
- Haciendo un éfasis natural en aplicaciones a las ciencias sociales