Sesión 8: Aplicaciones

ML-IA: Módulo de Machine learning

Jacob Muñoz-Castro

Universidad de los Andes

26 de noviembre de 2020

¿Qué es ciencia de datos?

Definamoslo de la forma más simple:

- Conjunto de metodologías para usar miles de formas de datos que son disponibles para lograr conclusiones significativas
- i.e. responder preguntas!

¿Qué pueden hacer los datos?

- Describir el estado actual de una organización o proceso
- Detectar eventos anómalos
- Diagnosticar las causas de eventos y comportamientos (p.e. Spotify, Netflix)
- Predecir eventos futuros

¿Por qué tan popular ciencia de datos?

- Por hoy, recolectamos más datos que nunca

Flujo de trabajo de Ciencia de Datos

1. Recolección de datos y almacenamiento
2. Preparación de datos
3. Exploración y visualización
4. Experimentos y predicciones

Ejercicio 1: Segmentación de clientes

María está a cargo de un grupo de ciencia de datos en una compañía de envíos de productos caninos por suscripción. Su grupo se le ha pedido investigar la pérdida de clientes también conocido como 'Abono de clientes'.

Ejercicio 1: Segmentación de clientes

María está a cargo de un grupo de ciencia de datos en una compañía de envíos de productos caninos por suscripción. Su grupo se le ha pedido investigar la pérdida de clientes también conocido como 'Abono de clientes'.
Organice los siguientes pasos:

- Reformatear la fecha de envíos de todas las entradas para que tengan el mismo huso horario
- Agrupar los usuarios en diferentes caracterizaciones de persona y realizar una regresión para predecir abono para cada grupo
- Descargar los datos
- Crear un gráfico de líneas para mostrar la caída de suscripciones por cohorte

Ejercicio 2: Creando un chatbot de servicio al cliente

Ritesh co. y su equipo de datos están trabajando en un chatbot de servicio al cliente. El chatbot es un programa de computador que usa ciencia de datos para responder preguntas básicas de clientes por medio de un servicio de mensajería. El equipo va a usar transcripciones de 300,000 interacciones de servicio al cliente para entrenar al chatbot para responder preguntas de clientes

Ejercicio 2: Creando un chatbot de servicio al cliente

Ritesh co. y su equipo de datos están trabajando en un chatbot de servicio al cliente. El chatbot es un programa de computador que usa ciencia de datos para responder preguntas básicas de clientes por medio de un servicio de mensajería. El equipo va a usar transcripciones de 300,000 interacciones de servicio al cliente para entrenar al chatbot para responder preguntas de clientes Clasifique en **Recolección de datos y almacenamiento (1)**, **Exploración y visualización (2)** y **Experimentación y predicción (3)** las siguientes tareas:

- Crear un gráfico de barras del número de conversaciones por cada tipo
- Recolectar las marcas de tiempo para cada transcripción
- Usar un modelo de machine learning para predecir posibles respuestas para cada pregunta
- Cargar las transcripciones en la base de datos del equipo de datos
- Mirar el número de conversaciones vs el tiempo en el día
- Crear un algoritmo que clasifique la pregunta inicial del cliente
- Reunir Información del cliente para cada conversación

Ejercicio 2: Creando un chatbot de servicio al cliente

Ritesh co. y su equipo de datos están trabajando en un chatbot de servicio al cliente. El chatbot es un programa de computador que usa ciencia de datos para responder preguntas básicas de clientes por medio de un servicio de mensajería. El equipo va a usar transcripciones de 300,000 interacciones de servicio al cliente para entrenar al chatbot para responder preguntas de clientes Clasifique en **Recolección de datos y almacenamiento (1)**, **Exploración y visualización (2)** y **Experimentación y predicción (3)** las siguientes tareas:

- **2** - Crear un gráfico de barras del número de conversaciones por cada tipo
- **1** - Recolectar las marcas de tiempo para cada transcripción
- **3** - Usar un modelo de machine learning para predecir posibles respuestas para cada pregunta
- **1** - Cargar las transcripciones en la base de datos del equipo de datos
- **2** - Mirar el número de conversaciones vs el tiempo en el día
- **3** - Crear un algoritmo que clasifique la pregunta inicial del cliente
- **1** - Reunir Información del cliente para cada conversación

Aplicaciones de Ciencia de datos: más casos de estudio

- Aprendizaje de máquinas tradicional
- IoT (el internet de las cosas)
- Aprendizaje profundo

Caso de estudio: Detección de fraude

- ¿Qué datos necesitamos?

Caso de estudio: Detección de fraude

- ¿Qué datos necesitamos?
  - Monto
  - Fecha
  - Ubicación
  - Tipo de compra
  - Dirección del titular

Caso de estudio: Detección de fraude

- ¿Qué datos necesitamos?
  - Monto
  - Fecha
  - Ubicación
  - Tipo de compra
  - Dirección del titular
  - **Una etiqueta que diga si hubo fraude**
- La pregunta: ¿Qué probabilidad hay de que una nueva transacción sea fraudulenta?

Aquí entra ML

¿Qué necesitamos de machine learning?

- Una pregunta bien definida: *¿Qué probabilidad hay de que esta transacción sea fraudulenta?*
- Un conjunto de datos: *transacciones antiguas etiquetadas con 'fraudulento' o 'válido'*
- Un nuevo conjunto de datos que para usar nuestro algoritmo: *Nuevas transacciones de tarjeta de crédito*

Caso de estudio: SmartWatch

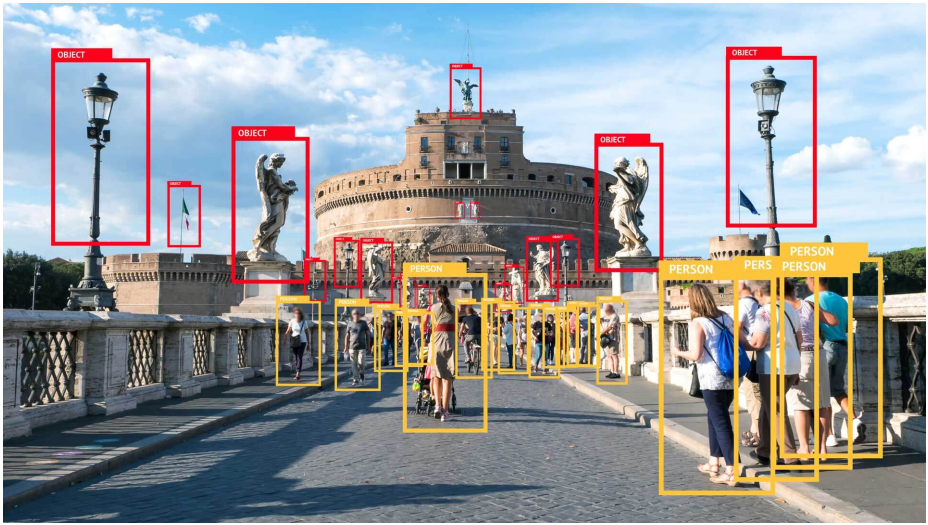
- Pregunta: ¿cuándo está corriendo o caminando el usuario?
- Acelerómetro

Internet de las cosas (IoT)

Se refiere a dispositivos que no son computadores

- Relojes inteligentes
- Sistemas de seguridad
- Peajes eléctricos
- Sistemas de administración de energía

Otro caso de estudio: reconocimiento de imagen



¿Machine Learning?

¿Cómo procedemos?

- La imagen es una cuadrícula de píxeles
- Entonces la imagen tiene representación en una matriz de números donde cada número representa un pixel
- ML falla! Dimensionalidad

Deep learning

- Muchas capas de mini-algoritmos llamados neuronas trabajan de forma conjunta para llegar a conclusiones
- Este tipo de aprendizaje exige muchos (MUCHOS) más datos de entrenamiento que ML tradicional para llegar a conclusiones que ML tradicional no logra desarrollar
- Es usado en problemas más complejos como clasificación de imágenes y aprendizaje/entendimiento del lenguaje

Ejercicio 3: Asignación de proyectos

Daniela dirige un equipo de analítica y tiene algunas tareas que está esperando alcanzar el primer trimestre de 2021 para la planeación corporativa. Las tareas están centradas alrededor de:

- Deep learning
- Machine learning traditional
- IoT/Internet de las cosas

El conocimiento alrededor de la construcción de aplicaciones de aprendizaje de máquinas y aprendizaje profundo está presente en su equipo. Otro equipo de la empresa es†ea especializado en IoT y por esto Daniela quiere saber sobre qué tareas va a solicitar colaboración.

Ejercicio 3: Asignación de proyectos

- Predicción de precios de servicios/viajes compartidos a cierto tiempo y ubicación basado en precios de servicios anteriores
- Alerta de imagen que contenga violaciones de seguridad (contenido)
- Automáticamente resumir textos de nuevos artículos
- Automatización del aire acondicionado de un edificio usando sensores de temperatura
- Agrupar pacientes por síntomas para ayudar a los doctores a elegir tratamiento
- Detección de fallas de maquinas con sensores de vibración

Ejercicio 3: Asignación de proyectos

- **ML** - Predicción de precios de servicios/viajes compartidos a cierto tiempo y ubicación basado en precios de servicios anteriores
- **DL** - Alerta de imagen que contenga violaciones de seguridad (contenido)
- **DL** - Automáticamente resumir textos de nuevos artículos
- **IoT** - Automatización del aire acondicionado de un edificio usando sensores de temperatura
- **ML** - Agrupar pacientes por síntomas para ayudar a los doctores a elegir tratamiento
- **IoT** - Detección de fallas de maquinas con sensores de vibración

Roles en ciencia de datos

Existen 4 roles que se desempeña en ciencia de datos

- Ingeniero de datos: encargado de la primera parte del proceso / Recolección y almacenamiento de datos. Para este rol pensamos en:
  - SQL (almacenar y organizar)
  - Java, Scala, Python (procesamiento)
  - Shell (automatización)
  - Cloud computing (AWS, Azure, GoogleCP)

Roles en ciencia de datos

Existen 4 roles que se desempeña en ciencia de datos

- Ingeniero de datos: encargado de la primera parte del proceso / Recolección y almacenamiento de datos. Para este rol pensamos en:
  - SQL (almacenar y organizar)
  - Java, Scala, Python (procesamiento)
  - Shell (automatización)
  - Cloud computing (AWS, Azure, GoogleCP)
- Analista de datos: análisis y descripción de datos, reportes y dashboards que resume datos, pero antes que todo, limpieza de datos. Se encargan de los dos pasos siguientes de ciencia de datos: preparación de datos y exploración/visualización. Para este rol pensamos en:
  - SQL (recuperar y agregar datos)
  - Hojas de cálculo - Excel/Gsheets (análisis simple)
  - Herramientas BI - Tableau/PowerBI (dashboards y visualización)
  - No excluye uso de Python y R para limpieza y análisis de datos

Roles en ciencia de datos

Existen 4 roles que se desempeña en ciencia de datos

- Científico de datos: encargado de preparación, exploración, visualización, experimentación y predicción (no se hace raro que este cargo absorba el anterior). Rol con habilidades estadísticas para encontrar insights en lo datos más que solo describirlos. ML tradicional para predicción y pronóstico. Para este rol pensamos en:
  - SQL (recuperar y agregar datos)
  - Python ('pandas') y/o R ('tidyverse')

Roles en ciencia de datos

Existen 4 roles que se desempeña en ciencia de datos

- Científico de datos: encargado de preparación, exploración, visualización, experimentación y predicción (no se hace raro que este cargo absorba el anterior). Rol con habilidades estadísticas para encontrar insights en lo datos más que solo describirlos. ML tradicional para predicción y pronóstico. Para este rol pensamos en:
  - SQL (recuperar y agregar datos)
  - Python ('pandas') y/o R ('tidyverse')
- Científico de ML: mismo de arriba especializado en ML (predicción, extrapolación, DL - procesamiento de imágenes,Procesamiento de lenguajes naturales). Para este rol pensamos en:
  - Python y/o R: paquetes como TensorFlow o Spark

Volviendo a ML

Ya vimos lo que hace ciencia de datos.
Si formulamos la misma pregunta para ML: ¿Qué puede hace ML?
Machine learning consiste en aplicar métodos estadísticos y computacionales en datos para:

- Obtener percepciones causales
- Predecir eventos futuros
- Entender patrones en los datos

Volviendo a ML

Ya vimos lo que hace ciencia de datos.
Si formulamos la misma pregunta para ML: ¿Qué puede hace ML?
Machine learning consiste en aplicar métodos estadísticos y computacionales en datos para:

- Obtener percepciones causales
  - ¿Qué causa que los clientes cancelen suscripciones a servicios?
- Predecir eventos futuros
  - ¿Qué clientes son más propensos a cancelar su suscripción el mes entrante?
- Entender patrones en los datos
  - ¿Hay grupos de clientes que son similares y usan nuestros servicios en una forma similar?

Resumiendo...

- Mercadeo
  - Predecir qué clientes van a comparar
  - Predecir valor esperado del ciclo del cliente (target clientes valiosos con servicios premium)
  - Segmentación de clientes: agrupar cliente según sus compras
- Sector financiero
  - Identificar atributos clave de transacciones que indican potenciales fraudes
  - Predecir qué clientes van a hacer default en sus pagos de hipoteca
  - Agrupar transacciones basado en atributos para entender qué segmentos de clientes son más rentables