Laboratorio de construcción de software

Trabajo práctico inicial

Universidad Nacional de General Sarmiento

Comisión 1 segundo semestre - año 2023

Docentes

Juan Carlos Monteros

Francisco Orozco De La Hoz

Leandro Dikenstein

Alumnos

Ignacio Barrientos

Tomas Mandril

Lucrecia Verón

Objetivo del documento

En este documento se explicarán los conceptos que se utilizaran en el trabajo práctico como Machine Learning, una subrama del mismo llamada Deep Learning, Big Data o ciencia de los datos y una explicación general del ciclo de vida de los datos.

Inteligencia artificial

Fuente: https://www.ibm.com/topics/artificial-intelligence
Según IBM, la Inteligencia Artificial (IA) es un campo que combina la ciencia del computador con grandes conjuntos de datos con el objetivo de resolver problemas. También incluye campos como “machine learning” y “deep learning”, que son esenciales en campos como la inteligencia artificial. Los algoritmos de IA crean sistemas expertos que hacen predicciones o clasificaciones basándose en una entrada de datos.

Deep learning y machine learning

Deep learning es un subcampo de machine learning, compuesto por “neural networks” (redes neuronales), “deep” se refiere a una red neuronal que tiene más de tres capas que incluye inputs y outputs.
La diferencia entre deep learning y machine learning es la forma en la que cada algoritmo aprende. Deep learning automatiza gran parte del proceso de “feature extraction”, eliminando gran parte de la intervención manual que se requiere y permite el uso de una base de datos más grande. Se puede decir que deep learning es machine learning pero escalable.
Según Lex Fridman, aquel machine learning que no sea deep depende más de la intervención manual para aprender. Expertos humanos determinan la jerarquía de features para entender las diferencias entre los inputs, requiriendo datos más estructurados para aprender.
Así que la principal diferencia es que machine learning usa algoritmos para parsear data, aprender de la misma y hacer decisiones basándose en lo aprendido, mientras que deep learning usa layers (capas) para crear una “red neuronal artificial” que puede aprender y hacer decisiones inteligentes por sí misma.

Modelos de machine learning

Fuente: https://www.javatpoint.com/machine-learning-models

Machine learning tiene tres modelos, y dos de ellos tienen sub-modelos:

- Aprendizaje supervisado:
    - Clasificación
    - Regression
- Aprendizaje sin supervisión
    - Clustering
    - Regla de asociación
    - “Reducción dimensional”
- Reinforcement Learning

Aprendizaje supervisado

Es el más simple de entender, los datos de entrada se llaman “training data” y tienen una etiqueta o un resultado como un output. Funciona con pares de input-output, necesita una función que puede ser entrenada con la training data y se aplica con datos desconocidos y tiene un rendimiento predictivo. Se puede clasificar en dos categorías:

Regresión

Acá el output es una variable continua. Los modelos que se usan con regresión son:

- Regresión linear
- Árboles de decisión
- Random Forest
- Redes neuronales artificiales

Clasificación

Generan conclusiones de valores observados categóricamente. Puede clasificar de las siguientes maneras:

- **Clasificación binaria**: Si el problema solo tiene dos posibilidades (sí o no) se lo llama un clasificador binario.
- **Clasificación de varias clases**: Si tiene más de dos posibilidades es de varias clases.

Algunos algoritmos de clasificación lineal son:

- Regresión logística

- “Support vector Machine”
- Naïve Bayes

Aprendizaje no supervisado

Implementa el proceso de aprendizaje opuesto al aprendizaje supervisado, en otras palabras, puede aprender de datos sin estructurar. Basándose en los datos sin estructura o “etiqueta” el modelo predice cuál va a ser el output, aprende patrones ocultos en los datos sin supervisión alguna.
Los modelos de aprendizaje no supervisado se basan en tres tareas:

1. **Clustering**
   Donde se agrupan los datos en diferentes clusters basándose en las similitudes y diferencias. Los más similares se agrupan en el mismo cluster y no tienen, o tienen muy pocas similitudes, con los otros clusters.
2. Aprendizaje de **Association Rule** (Regla de asociación):
   Las variables que hay en los datos se llama la “dimensionalidad” de los datos, y la técnica que se usa para reducirla se conoce como la técnica de reducción de dimensionalidad

También existe el método de aprendizaje llamado **Reinforcement Learning**, el cual es muy similar al aprendizaje humano, donde un algoritmo aprende acciones basándose en respuestas, si hace una acción positiva recibe una respuesta positiva y si hace una acción negativa recibe una respuesta negativa. La idea es que el algoritmo maximice la cantidad de respuestas positivas para mejorar su rendimiento.

Data science / Big data

¿Qué es Big Data?

Fuente: https://www.oracle.com/ar/big-data/what-is-big-data/
Oracle describe Big Data como conjuntos de datos de mayor tamaño y más complejos, los cuales proceden en gran mayoría de nuevas fuentes de datos. Como hay una gran cantidad de estos datos el software de procesamiento de datos convencional no los puede gestionar, pero aún así se puede usar para abordar problemas empresariales que antes no era posible de solucionar.

Big data tiene **tres V**:

1. Volumen: La cantidad de datos

2. Velocidad: El ritmo al que se reciben los datos y al que se les aplica alguna acción
3. Variedad: Los diversos tipos de datos que existen

El ciclo de vida de los datos

Fuente: https://hdsr.mitpress.mit.edu/pub/577rq08d/release/4

El ciclo de vida de la data science, tiene seis etapas y una que es opcional:

1. Generación de los datos
2. Colección de los datos
3. Procesamiento de los datos
4. Almacenamiento de los datos
5. Gestión de los datos
6. Análisis de los datos
7. (Opcional) Visualización de los datos

Generación de los datos

Las personas generan millones de datos con cada acción que hacen, los sensores de las infraestructuras también generan datos, todo lo que esté relacionado a Internet of Things también genera datos. Pero no todos estos datos son guardados o recolectados.

Colección de los datos

Como se mencionó antes, no todos los datos que se generan son recolectados, ya que hay mucha información que no es relevante para el proyecto, o también porque la información que se recibe es más de la que se puede procesar.

Procesamiento de los datos

En este paso ocurre la limpieza de los datos que se recolectaron, su formateo y encriptación. Se hacen todas estas acciones para que los datos estén ordenados, seguros y no consuman tanto espacio.

Almacenamiento de los datos

En esta etapa los bits de la información procesada se guardan en cintas magnéticas, discos duros o algún otro tipo de almacenamiento.

Gestión de los datos

Acá se guardan datos de manera estructurada o no, también se usa la metadata para maximizar la velocidad de acceso y modificación de la información.

Análisis de los datos

Se aplican técnicas para analizar los datos, se emplean algoritmos y métodos relacionados a la inteligencia artificial, machine learning, y se aplican para clasificar o predecir. Esta etapa es la más importante en Data Science

Visualización de los datos (opcional)

Toda la información se guarda para que sea más fácil para la computadora, no para las personas. La etapa de visualización no es necesaria pero permite ver los datos de una manera que sea fácil de interpretar y visualizar para las personas. Le provee al lector una explicación de qué significan los datos que fueron procesados.

Objetivos

Fuente: https://datos.gob.ar/dataset/salud-vigilancia-enfermedades-por-virus-dengue-zika
Se propone crear un modelo de IA que, en base a los datos obtenidos de la fuente mencionada, permita predecir de forma aproximada la cantidad de casos de las enfermedades Dengue y Zika en determinadas regiones, y clasificar dichos resultados por cada región por rango etario.
Para esto se implementará un modelo de machine learning que se encargue de realizar esta clasificación de forma automática luego de suministrarle los datos necesarios para su entrenamiento y funcionamiento. A nivel implementación, se consideran como opciones principales los algoritmos de machine learning SVM (Support vector machine) y Naive Bayes. Se investigará más a fondo la funcionalidad de cada uno para luego decidir cual utilizar.
Se utilizara el siguiente repositorio en Gitlab para llevar control de version del proyecto: https://gitlab.com/mandrilTom/pp1-tp-inicial