IA 1

UT 01 – PD1

Felipe Mestre

Ejercicio 1

Machine Learning

Machine learning is the study of computer algorithms that improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence.
https://en.wikipedia.org/wiki/Machine_learning

El ‘machine learning’ –aprendizaje automático– es una rama de la inteligencia artificial que permite que las máquinas aprendan sin ser expresamente programadas para ello. Una habilidad indispensable para hacer sistemas capaces de identificar patrones entre los datos para hacer predicciones.
Te contamos qué es el 'machine learning' y cómo funciona (bbva.com)

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.
What is Machine Learning? | IBM

Inteligencia artificial

Artificial intelligence is intelligence demonstrated by machines, as opposed to the natural intelligence displayed by humans or animals.
https://en.wikipedia.org/wiki/Artificial_intelligence

Análisis Estadístico

Statistical analysis means investigating trends, patterns, and relationships using quantitative data. It is an important research tool used by scientists, governments, businesses, and other organizations. ... Then, you can use inferential statistics to formally test hypotheses and make estimates about the population.
https://www.scribbr.com/category/statistics/

Data mining

Data mining, also known as knowledge discovery in data (KDD), is the process of uncovering patterns and other valuable information from large data sets. ... Data mining has improved organizational decision-making through insightful data analyses.
https://www.ibm.com/cloud/learn/data-mining

¿Qué tienen en común y en qué se diferencia con Inteligencia Artificial?

El objetivo de la inteligencia artificial es que un sistema informático imite la inteligencia humana. Machine Learning se basa en hacer que un sistema aprenda en base a datos que permitan brindar un resultado preciso en tareas más concretas. Por ende, Machine Learning es una rama de la inteligencia artificial; existen otras formas de inteligencia artificial como la clásica, imitar el aprendizaje no es exactamente imitar toda la inteligencia humana.

¿Qué tienen en común y en qué se diferencia con Análisis Estadístico?

Ambos conceptos tienen un objetivo en común que es aprender de los datos. Los algoritmos de machine learning predicen datos y son capaces de aprender de billones de otros datos. Se usa mejor cuando existe una gran cantidad de datos con una gran cantidad de atributos y observaciones. El modelado estadístico usa menos datos con menos atributos y por ende existe una posibilidad de que se produzca un ajuste excesivo.

Los modelos estadísticos requieren un conjunto de suposiciones sobre la generación de datos observados y datos similares de poblaciones más grandes. Machine Learning no requiere suposiciones previas sobre las relaciones subyacentes entre las variables, solo debe ingresar todos los datos que tenga, y el algoritmo procesa los datos y descubre patrones, con estos patrones luego va a hacer predicciones sobre el nuevo conjunto de datos.

¿Qué tienen en común y en qué se diferencia con Data Mining?

El objetivo de data mining es brindar conocimientos en un área donde realmente había poco conocimiento de antemano, o ser capaz de predecir observaciones. La diferencia entre los dos reside en que en machine learning el esfuerzo humano solo está presente cuando se define el algoritmo, además los procedimientos pueden en Data Mining ser no supervisados (no se conoce la respuesta-descubrimiento) o supervisados (se conoce la respuesta).

Ambos utilizan los mismos algoritmos clave para descubrir patrones en los datos. Sin embargo, su proceso y utilidad son diferentes.

¿En qué se aplica machine learning?

En diagnostico médico, motores de búsqueda, reconocimiento del habla, detección de rostros, anti spams, anti-virus, predicciones de clima, etc.

Ejercicio 2

Alteryx

Self-Service Analytics, Data Science & Process Automation | Alteryx

La herramienta proporciona analítica automatizada de todo tipo, ciencia de datos e inteligencia artificial sin codificación.

Su propósito puede ser mejorar las ventas, reducir la tasa de cancelación, automatizar el pago de impuestos y las auditorías, optimizar tu cadena de suministro, pronosticar pagos de seguros, mejorar el desempeño de tus empleados o implementar rápidamente a Alteryx con Azure, AWS, Tableau, Salesforce u otras. Presenta herramientas de aprendizaje automático guiado, NLP y minería de textos visual, o accede a herramientas integradas de R y Python.

SAS

SAS Machine Learning | SAS

Proporciona programación a demanda para acceder a los algoritmos de machine learning en la nube, una experiencia para generar modelos, acceder a los resultados y obtener conocimientos. Además, combina la preparación de datos, ingeniería, estadística moderna y técnicas de machine learning en un ambiente de procesamiento en memoria para desarrollar, testear y desplegar modelos.

IBM

Machine Learning (ibm.com)

Ofrece varios productos con opciones como ejecución en la nube pública y privada o en local y plataformas cooperativas o locales para explorar, modelar y desplegar soluciones de datos. Sus herramientas permiten usar algoritmos de aprendizaje supervisados, no supervisados, Deep learning además de algoritmos de NLP planning y reasoning.

RapidMiner

RapidMiner.com

Es una herramienta de análisis y minería de datos que permite desarrollar procesos de análisis de datos usando distintos operadores en un entorno gráfico. Incluye 500 operadores para hacer análisis, preprocesamiento y visualización de datos.

Weka

Weka

Da acceso a una colección de algoritmos, herramientas de visualización para modelado predictivo y análisis de datos mediante una interfaz gráfica. Provee clasificación, agrupamiento, preprocesamiento, regresión, selección de características y visualización de datos.

Ejercicio 3

CRISP-DM

Son las siglas de cross industry standard process for data mining, es un método para orientar trabajos de minería de datos.

La metodología contiene seis fases relacionadas sobre las cuales se pueden hacer ciertas iteraciones según los requerimientos del proyecto. Los pasos son comprensión empresarial, comprensión de datos, preparación de datos, modelado, evaluación y despliegue.

La comprensión empresarial tiene como objetivo contextualizar las metas y datos para brindar este conocimiento a los desarrolladores y así poder reconocer la relevancia de los datos en el negocio y poder hacer análisis acertados. Para ello se deberán llevar a cabo lecturas de documentación, reuniones, aprendizaje de campo entre otras actividades

La compresión de datos es el segundo paso, se basa en establecer un objetivo sobre que se puede lograr a partir de los datos disponibles. Se debe verificar la calidad, integridad y distribución de los datos y sus valores. En esta parte del proceso se especifica su viabilidad y la confiabilidad de los resultados finales.

El tercer paso se basa en preparar los datos mediante procesos que preparan los datos para que sean utilizables por los algoritmos a usar.

El cuarto paso es el modelado y es núcleo de los proyectos de aprendizaje automático. En este paso está en juego si los resultados podrán satisfacer o no los requerimientos.

La evaluación es el quinto paso del proceso. Aquí se verificará la validez y correctitud de los resultados. Si los resultados carecen de calidad la metodología permite volver al primer paso del proceso y reflexionar acerca de porque los resultados son erróneos.

El sexto paso se llama despliegue y consiste en presentar los resultados de una forma óptima, comprensible y útil para su consumidor final.

CRISP-DM - Data Science Process Alliance (datascience-pm.com)

Una metodología similar a CRISP-DM es TDSP. TDSP es un ciclo de vida que estructura el desarrollo de un proyecto de ciencia de datos en 5 etapas:

1- Comprensión empresarial
2- Adquisición y comprensión de datos
3- Modelado
4- Despliegue
5- Aceptación del cliente

Esta metodología está catalogada como ágil y más orientada hacia una filosofía iterativa a diferencia de CRISP-DM.

What is the Team Data Science Process? - Azure Architecture Center | Microsoft Docs

Ejercicio 4

RapidMiner

- Aprendizaje supervisado: Regresión
  - Árboles de decisión: Decision Tree - RapidMiner Documentation
  - Regresión Lineal: Linear Regression - RapidMiner Documentation
  - Bosque Aleatorio: Random Forest - RapidMiner Documentation
  - Red Neuronal: Neural Net - RapidMiner Documentation
  - Gradient Boosted Tries: Gradient Boosted Trees - RapidMiner Documentation
- Aprendizaje no supervisado: Reducción de dimensiones
  - Análisis de componente principal: Principal Component Analysis - RapidMiner Documentation
  - Descomposición en valores singulares: Singular Value Decomposition - RapidMiner Documentation
  - Análisis de Dirichlet: No lo tiene
- Aprendizaje no supervisado: Clustering
  - DBSCAN: DBSCAN - RapidMiner Documentation
  - Hierarchical: Agglomerative Clustering - RapidMiner Documentation
  - K-Medoids: k-Medoids - RapidMiner Documentation
  - K-Means: k-Means - RapidMiner Documentation
  - Gaussian Mixture Model: Gaussian Process - RapidMiner Documentation
- Aprendizaje supervisado: Clasificación
  - Kernel SVM: Support Vector Machine - RapidMiner Documentation
  - Random Forest: Random Forest - RapidMiner Documentation
  - Red Neuronal: Neural Net - RapidMiner Documentation
  - Gradient Boosted Tries: Gradient Boosted Trees - RapidMiner Documentation
  - Árboles de decisión: Decision Tree - RapidMiner Documentation
  - Regresión Logística: Logistic Regression (SVM) - RapidMiner Documentation
  - Naive Bayes: Naive Bayes - RapidMiner Documentation
  - Linear SVM: Weight by SVM - RapidMiner Documentation

Alteryx

- Aprendizaje supervisado: Regresión
  - Árboles de decisión: Regression Tool | Alteryx Help
  - Regresión Lineal: Regression Tool | Alteryx Help
  - Bosque Aleatorio: Regression Tool | Alteryx Help
  - Red Neuronal: Neural Network Tool | Alteryx Help
  - Gradient Boosted Tries: Boosted Model Tool | Alteryx Help
- Aprendizaje no supervisado: Reducción de dimensiones
  - Análisis de componente principal: Principal Components Tool | Alteryx Help
  - Descomposición en valores singulares: No lo tiene
  - Análisis de Dirichlet: No lo tiene
- Aprendizaje no supervisado: Clustering
  - DBSCAN: No lo tiene
  - Hierarchical: No lo tiene
  - K-Medoids: No lo tiene
  - K-Means: K-Centroids Cluster Analysis Tool | Alteryx Help

    - Gaussian Mixture Model: No lo tiene
- Aprendizaje supervisado: Clasificación
    - Kernel SVM: Support Vector Machine Tool | Alteryx Help
    - Random Forest: Neural Network Tool | Alteryx Help
    - Red Neuronal: Neural Network Tool | Alteryx Help
    - Gradient Boosted Tries: Boosted Model Tool | Alteryx Help
    - Árboles de decisión: Decision Tree Tool | Alteryx Help
    - Regresión Logística: Logistic Regression Tool | Alteryx Help
    - Naive Bayes: Naive Bayes Classifier Tool | Alteryx Help
    - Linear SVM: Support Vector Machine Tool | Alteryx Help

IBM Watson Studio

- Aprendizaje supervisado: Regresión
    - Árboles de decisión: IBM Watson Studio Documentation
    - Regresión Lineal: IBM Watson Studio Documentation
    - Bosque Aleatorio: IBM Watson Studio Documentation
    - Red Neuronal: IBM Watson Studio Documentation
    - Gradient Boosted Tries: IBM Watson Studio Documentation
- Aprendizaje no supervisado: Reducción de dimensiones
    - Análisis de componente principal: IBM Watson Studio Documentation
    - Descomposición en valores singulares: IBM Watson Studio Documentation
    - Análisis de Dirichlet: No lo tiene
- Aprendizaje no supervisado: Clustering
    - DBSCAN: IBM Watson Studio Documentation
    - Hierarchical: IBM Watson Studio Documentation
    - K-Medoids: No lo tiene
    - K-Means: IBM Watson Studio Documentation
    - Gaussian Mixture Model: No lo tiene
- Aprendizaje supervisado: Clasificación
    - Kernel SVM: IBM Watson Studio Documentation
    - Random Forest: IBM Watson Studio Documentation
    - Red Neuronal: IBM Watson Studio Documentation
    - Gradient Boosted Tries: IBM Watson Studio Documentation
    - Árboles de decisión: IBM Watson Studio Documentation
    - Regresión Logística: IBM Watson Studio Documentation
    - Naive Bayes: IBM Watson Studio Documentation
    - Linear SVM: IBM Watson Studio Documentation

SAS

- Aprendizaje supervisado: Regresión
    - Árboles de decisión: Sas Viya Documentation
    - Regresión Lineal: Sas Viya Documentation
    - Bosque Aleatorio: Sas Viya Documentation
    - Red Neuronal: Sas Viya Documentation
    - Gradient Boosted Tries: No figura
- Aprendizaje no supervisado: Reducción de dimensiones
    - Análisis de componente principal: Sas Viya Documentation
    - Descomposición en valores singulares: Sas Viya Documentation

    - Análisis de Dirichlet: Sas Viya Documentation
- Aprendizaje no supervisado: Clustering
    - DBSCAN: Sas Viya Documentation
    - Hierarchical: Sas Viya Documentation
    - K-Medoids: No figura
    - K-Means: Sas Viya Documentation
    - Gaussian Mixture Model: No lo tiene
- Aprendizaje supervisado: Clasificación
    - Kernel SVM: Sas Viya Documentation
    - Random Forest: Sas Viya Documentation
    - Red Neuronal: Sas Viya Documentation
    - Gradient Boosted Tries: No figura
    - Árboles de decisión: Sas Viya Documentation
    - Regresión Logística: Sas Viya Documentation
    - Naive Bayes: Sas Viya Documentation
    - Linear SVM: Sas Viya Documentation

Weka

- Aprendizaje supervisado: Regresión
    - Árboles de decisión: The Weka Workbench Data Mining: Practical Machine Learning Tools and Techniques capítulo 2.1.3.
    - Regresión Lineal: The Weka Workbench Data Mining: Practical Machine Learning Tools and Techniques capítulo 2.4.4
    - Bosque Aleatorio: The Weka Workbench Data Mining: Practical Machine Learning Tools and Techniques capítulo 2.4.2
    - Red Neuronal: The Weka Workbench Data Mining: Practical Machine Learning Tools and Techniques capítulo 2.4.5
    - Gradient Boosted Tries: No figura
- Aprendizaje no supervisado: Reducción de dimensiones
    - Análisis de componente principal: The Weka Workbench Data Mining: Practical Machine Learning Tools and Techniques capítulo 2.3.1.3
    - Descomposición en valores singulares: No figura
    - Análisis de Dirichlet: No lo tiene
- Aprendizaje no supervisado: Clustering
    - DBSCAN: No lo tiene
    - Hierarchical: The Weka Workbench Data Mining: Practical Machine Learning Tools and Techniques capítulo 2.5
    - K-Medoids: No figura
    - K-Means: The Weka Workbench Data Mining: Practical Machine Learning Tools and Techniques capítulo 2.5
    - Gaussian Mixture Model: No lo tiene
- Aprendizaje supervisado: Clasificación
    - Kernel SVM: The Weka Workbench Data Mining: Practical Machine Learning Tools and Techniques capítulo 3.2
    - Random Forest: The Weka Workbench Data Mining: Practical Machine Learning Tools and Techniques capítulo 2.4.2
    - Red Neuronal: The Weka Workbench Data Mining: Practical Machine Learning Tools and Techniques capítulo 2.4.5

- Gradient Boosted Tries: No figura
- Árboles de decisión: The Weka Workbench Data Mining: Practical Machine Learning Tools and Techniques capítulo 2.1.3
- Regresión Logística: The Weka Workbench Data Mining: Practical Machine Learning Tools and Techniques capítulo 2.4.4
- Naive Bayes: The Weka Workbench Data Mining: Practical Machine Learning Tools and Techniques capítulo 2.4.1
- Linear SVM: The Weka Workbench Data Mining: Practical Machine Learning Tools and Techniques capítulo 3.2

Ejercicio 5

1- Elegí el dataset UCI Machine Learning Repository: Congressional Voting Records Data Set. El dataset tiene datos sobre que leyes votaron los 435 legisladores del congreso de los estados unidos en el año 1984.
2- El objetivo del dataset es lograr clasificar si un legislador es demócrata o republicano según la información sobre qué leyes votó.
3- Atributos
    a. El primer atributo es el partido del legislador que puede ser democrat o republican. Esta es la variable para predecir.
    b. El resto de las 16 columnas refieren a leyes y su valor es y (yes) o n (no) que refiere a si el legislador votó o no la ley. Básicamente son valores booleanos. Lógicamente van a haber leyes que sean votadas en mayoría por los legisladores de un partido que de otro, estás leyes serán más determinantes para hacer la predicción.
4- Algoritmos para resolver:
    a. Este es un problema de clasificación binaria por ende se podría utilizar un algoritmo de regresión logística. Otra opción es utilizar algoritmos de clustering como arboles de decisión.