Introducción a los métodos de aprendizaje automático

Introducción al curso

1

Objetivos y modalidad del curso

- enfoque aplicado, de “afuera hacia adentro”
- primero veremos las características de algoritmos y modelos, procurando usarlos y discernir entre alternativas
- luego se podrá ir profundizando y analizando en detalle los algoritmos, a efectos de mejorar el rendimiento, precisión, etc.
- queremos generar modelos de ML y construir programas que los integren
- Resultados Esperados del Aprendizaje

2

2

Cronograma

Fecha	Unidad temática - Evento
15/08/2023	UT01 - Introducción al curso y metodología, INTRODUCCIÓN a ML
22/08/2023	UT02 – Tratamiento previo de los datos y fundamentos de los algoritmos de ML
05/09/2023	UT03 – Algoritmos Lineales
26/09/2023	UT04 – Algoritmos No Lineales PARTE 1
14/10/2023	PRIMER PARCIAL, PRESENTACION INDIVIDUAL PORTAFOLIOS, PORTAFOLIOS DE EQUIPO
17/10/2023	UT04 – Algoritmos No Lineales PARTE 2
31/10/2023	UT05 – Clustering y Modelos Jerárquicos
14/11/2023	UT06 – Ensambles y UT07 - Evaluación
28/11/2023	SEGUNDO PARCIAL, PRESENTACIONES PORTAFOLIOS Y POSTERS, CONCURSO

3

3

Dinámica de trabajo

- aprendizaje basado en equipos
- preparación previa básica
- breves introducciones conceptuales
- ejercicios de aplicación – modelos / prototipos
- ejercicios domiciliarios – evaluación de modelos, programas en python
- generación incremental del portafolios y POSTER

4

4

Evaluación

- Aprobación directa – calificación de B o superior
- Componentes
  - pruebas parciales individuales 40%
  - pruebas de aseguramiento de la preparación de cada Unidad Temática 15% (iRAts 12%, tRAts 3%)
  - Portafolio de Machine Learning y POSTER 35% (fase 1 - 15%, fase 2 - 20%)
  - Ejercicios domiciliarios 5%
  - Evaluación de pares 5%

5

5

Unidad temática 1 - introducción

- Objetivos:

Esta primera unidad del curso tiene como objetivo primario el presentar la temática del aprendizaje automático, proveyendo un marco de trabajo aplicado y con enfoque industrial

- Resultados esperados del aprendizaje:

Al culminar esta unidad de aprendizaje serás capaz de:

- Identificar técnicas y herramientas para el tratamiento del aprendizaje automático actualmente disponibles y populares en la industria, y discutir sus ventajas y desventajas
- Comprender, describir y aplicar un proceso estándar de Data Science / Machine Learning
- Discutir a grandes rasgos los tipos de algoritmos de ML existentes y sus aplicaciones.
- Comenzar a construir tu portafolios de ML
- Localizar y utilizar conjuntos de datos públicos disponibles para la práctica de técnicas de ML
- Comenzar a utilizar una herramienta típica industrial de modelado y ejecución de ML
- Utilizar planillas electrónicas para análisis estadísticos básicos

6

6

Enfoque

UCU

- Sistemas aplicados
- algoritmos como “caja negra”, más adelante análisis y ajuste de parámetros
- problemas reales comúnmente atacados con técnicas de Ml
- integración en aplicaciones de software

7

7

Qué es “Machine Learning”

UCU

En **10 MINUTOS:**
hacer una búsqueda rápida y listar al menos 3 definiciones de “Machine Learning” y responder:

1. ¿qué tiene en común y en qué se diferencia de “Inteligencia Artificial”?
2. ¿qué tiene en común y en qué se diferencia de “Análisis Estadístico”?
3. ¿Cómo se diferencia con Data Mining?
4. ¿en qué se aplica?

8

8

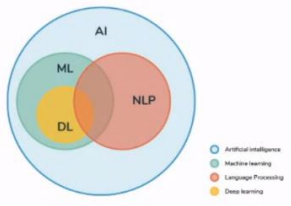
UCU

AI: Enabling machines to think like humans

ML: Training machines to learn from the past
- **Supervised Learning:** Making predictions from labeled data
- **Unsupervised Learning:** Discovering underlying clusters & patterns within unlabeled data

DL: Subset of ML that uses neural networks for complex supervised & unsupervised learning tasks

NLP: Enabling machines to read, understand, and derive meaning from human language



9

9

Plataformas y herramientas

- existen múltiples plataformas y herramientas modernas para generar y ejecutar modelos de ML
  - R y Python con SciKitLearn
  - RapidMiner
  - Microsoft Azure ML Studio
  - Knime
  - Weka
  - Keras, TensorFlow, etc.
  - … y muchas otras más

10

10

Plataformas y herramientas Ejercicio …

utilizando 2 computadoras por equipo, y en **10 minutos**:

- buscar con google diferentes herramientas
- listar los nombres, url donde se la describe y características más importantes

- **Fuera de clase**, como trabajo domiciliario, extender esta "investigación" preliminar y producir un breve documento de resumen

11

11

Plataformas y herramientas Cuadrado de Gartner 2022

Gartner Reprint

Figure 1: Magic Quadrant for Analytics and Business Intelligence Platforms

12

12

Proceso de Data Science….

un posible proceso sistemático, en 5 pasos(1):

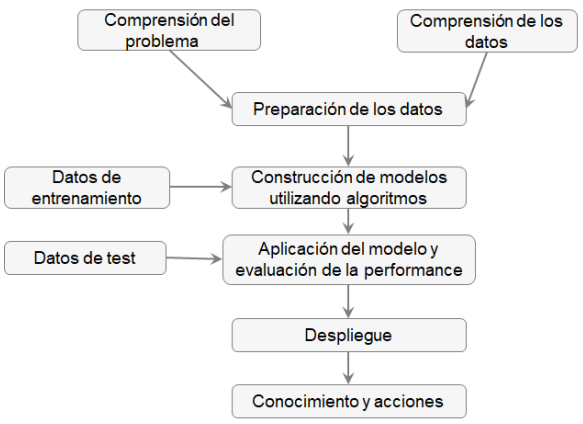
1. definir el problema
2. preparar los datos
3. elegir los algoritmos más apropiados
4. mejorar los resultados
5. presentar los resultados

(1) tomado de http://machinelearningmastery.com/process-for-working-through-machine-learning-problems/

13

13

Proceso….



1. Conocimiento previo

2. Preparación

3. Modelado

4. Aplicación

5. Conocimiento

14

14

el proceso CRISP-DM

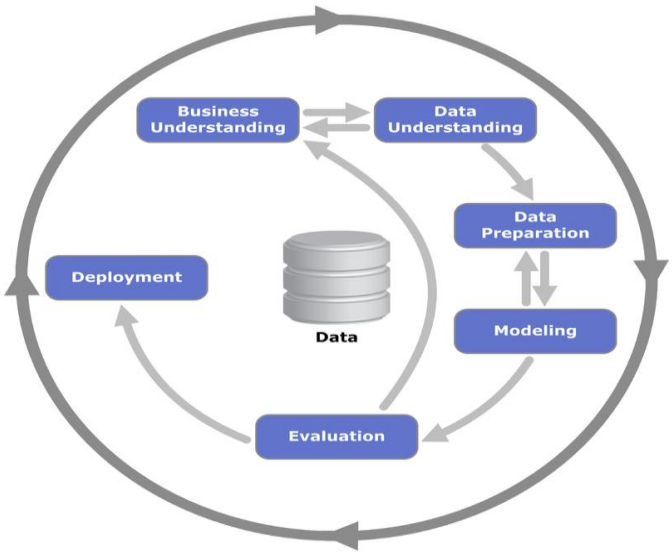
- Comprensión del negocio
- Comprensión de Datos
- Preparación de datos
- Modelado
- Evaluación
- Despliegue

ejercicio domiciliario: buscar información sobre **CRISP-DM** y realizar un resumen con las características más importantes

15

15

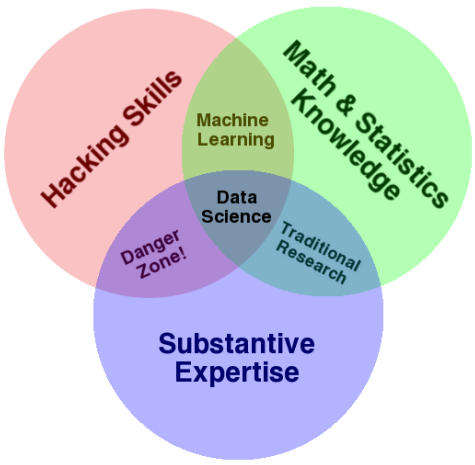
el proceso CRISP-DM



16

16

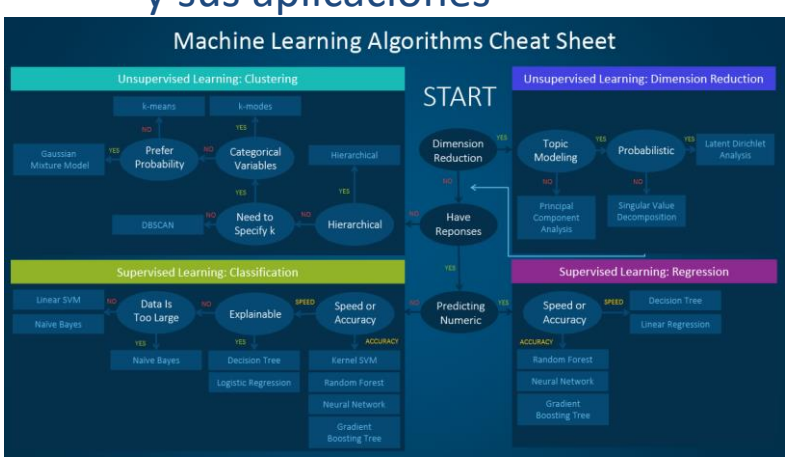
Intersección de conocimientos....



17

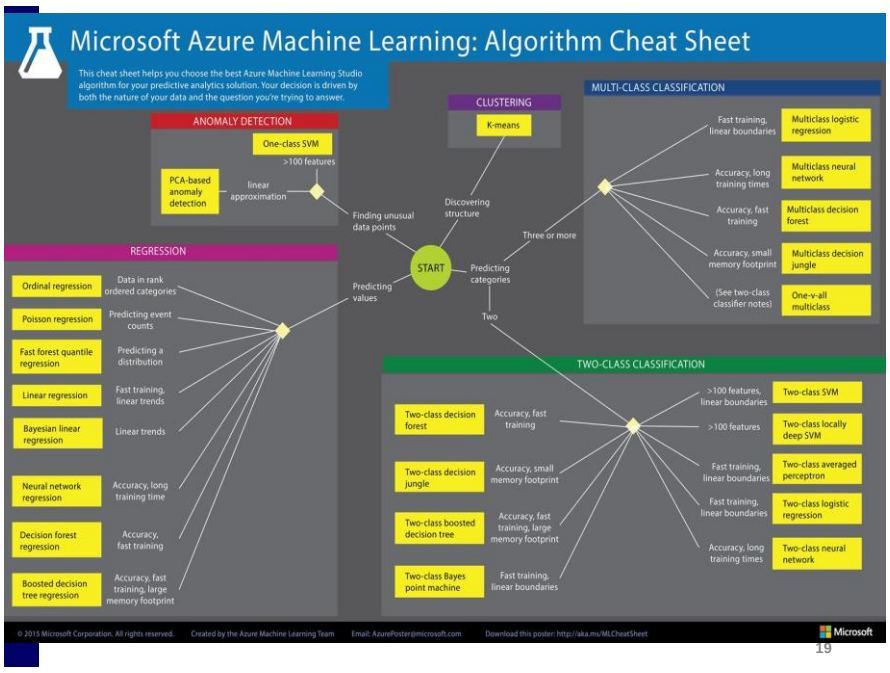
17

Tipos de algoritmos y sus aplicaciones



18

18



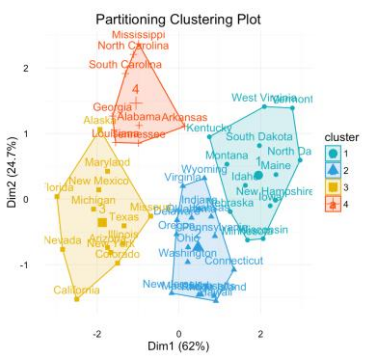
19

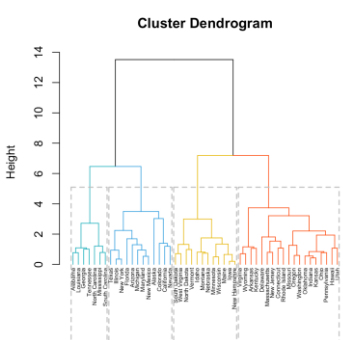
20

Clustering

UCU

Permite agrupar datos en función de las características de sus atributos





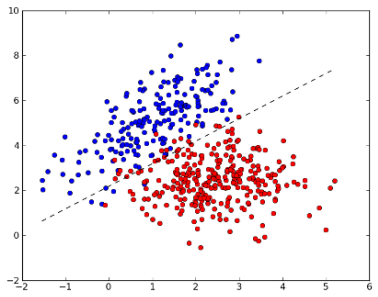
22

Aprendizaje Supervisado Clasificación

UCU

Identificar a qué categoría pertenece una muestra

Ej: clasificación binaria a partir de un modelo lineal

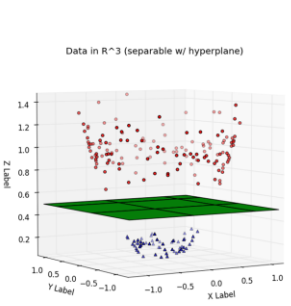


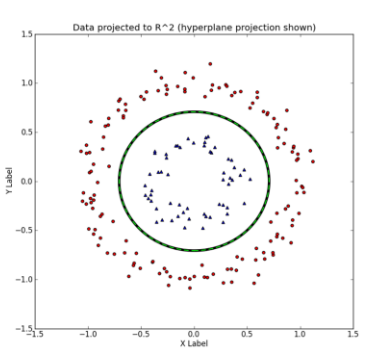
23

Aprendizaje Supervisado Clasificación

UCU

Ej: clasificación binaria mediante “kernel trick” (SVM)



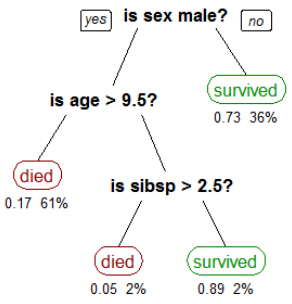


24

Aprendizaje Supervisado Clasificación

UCU

Ej: clasificación basada en árbol de decisión



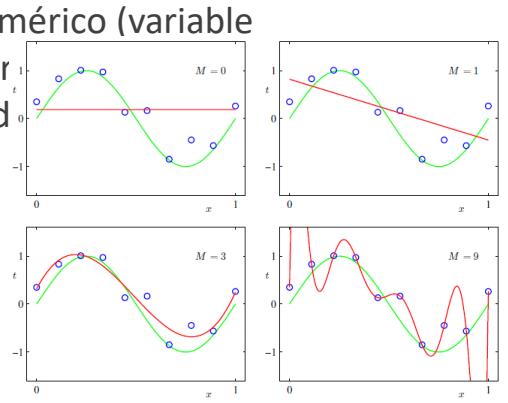
Modelo Explicable!!

25

Aprendizaje Supervisado Regresión

UCU

Predecir un valor numérico (variable dependiente) a par variables independ



Ej: regresión mediante polinomios

26

Conjuntos de datos

UCU

- Fuentes
- Características deseables
- Tipos de problemas a resolver

27

Ejercicio: Conjunto de datos y Tipos de Algoritmos

UCU

Utilizando los datasets del repositorio UCI:
http://archive.ics.uci.edu/ml/datasets.html
en **15 minutos**,

- Analizar 3 ejemplos e identificar
  - cuál es el problema que abordan
  - qué tipos de algoritmos de Machine Learning pueden aplicarse para resolverlo.

28

Uso de hojas de cálculo y estadística

- referencias celdas
- operadores
- suma(rango),
- sumaProducto
- cuentas
- condicionales (if)
- generación de números aleatorios

29

29

Ejercicio - uso básico de hoja de cálculo (1) (15 mins)

1. en la columna C1: 10 filas, generar valores aleatorios entre 1 y 1000, con distribución uniforme. Los valores deben ser de tipo entero.
2. calcular la suma de todos los elementos de C1
3. C2: generar 10 valores aleatorios, enteros, entre 1 y 5
4. hallar el valor mínimo y el máximo de la columna C1
5. en columna C3, calcular los valores **normalizados** de C1
6. usando "count", indicar cuántas celdas de la columna C1 tienen valores menores a 200, y cuántas tienen valores mayores a 700
7. calcular el cuadrado, la raíz cuadrada y la potencia cúbica de las suma de valores de C1, multiplicados por la constante PI
8. generar un gráfico de barras con los valores de las sumas de las columnas C1y C3
9. utilizando la función/ fórmula "if", en las celdas de la columna C7 poner
   a. un texto "inferior" si el valor de la columna C1 es menor que 100,
   b. el valor del logaritmo en base 2 de C1 si éste está entre 100 y 500, o
   c. "superior" si es mayor que 500

30

30

Estadística descriptiva básica, conceptos para el ejercicio

REVISAR LOS CONCEPTOS DE:

1. media
2. desvío estándar , varianza
3. moda
4. mediana
5. rango, mínimo, máximo

6. distribución normal, media, varianza

7. números aleatorios,
   a. con distribución uniforme
   b. con distribución gaussiana

31

31

Ejercicio - estadísticas básicas en hoja de cálculo (1) (10 mins)

Instalar ¨Analysis Toolpack¨ en Excel
File / options / Addins

Utilizando la hoja de cálculo generada en el anterior ejercicio, calcular para los valores de la columna C1::

- media
- varianza
- desvío estándar
- moda
- mediana

En una segunda hoja, en la columna C1, generar 100 valores aleatorios con distribución normal (gaussiana), y luego calcular, para este conjunto de datos:

- mínimo
- máximo
- media
- desvío estándar
- varianza

Revisión de Rapid Miner

1. ayuda
   a. tutoriales, foro, documentos
   b. doc en línea de cada bloque

2. repositorio
   a. paleta de ejemplos (datos, procesos, templates y tutoriales)
   b. repositorio local (datos y procesos)

Revisión de Rapid Miner (2)

3. Operadores
   a. Data Access
   b. Blending
   c. Cleansing
   d. Modelling
   e. Scoring
   f. Validation
   g. Utility
   h. Extensions

Ejercicio - datos en RM

Abrir el conjunto de datos “Iris” (carpeta “Samples/data”

- observar los atributos, la variable objetivo (“label”)
- revisar estadísticas
- gráficos
- editor de datos

EJ Domiciliario: descargar el dataset “Iris” de UCI, analizarlo con hoja de cálculo y compararlo con el dataset de muestra Iris de RM

35

35

Portafolio de ML

Colección de proyectos independientes, los cuales utilizan machine learning de alguna forma.

- Propiedades de un buen portafolio de proyectos:
  - Accesible: repositorio público
  - Acotado: alcance concreto
  - Completo: objetivo alcanzable
  - Independiente: tiene sentido por si mismo
  - Entendible: propósito y resultados claros
- Referencia: http://machinelearningmastery.com/build-a-machine-learning-portfolio/

36

36

PORTAFOLIOS DE MACHINE LEARNING

IMPORTANTE:

- seleccionar la(s) plataformas, bosquejar una estructura inicial y comenzar a crear el Portafolios
- Enviar la URL en la tarea que se ha de publicar en la sección “Portafolios” de la webas

37

37

Preguntas...