

# Lineamientos para evaluación del curso Inteligencia Artificial Tema No Supervisados

O. L. Quintero

5 de marzo de 2024

## Resumen

Este documento contiene los lineamientos para la presentación de la actividad evaluativa el domingo 17 de marzo segun calendario

## 1. Introducción

El objetivo del curso es proveer elementos teóricos y conceptuales que le permitan a los estudiantes, enfrentar el problema de construir un modelo compacto (learning machine) que permita representar fenómenos del mundo real.

Consecuentemente, los principios de teoría de aprendizaje fueron adelantados en la primera sesión. Se debe cuestionar y NO desviar la tarea de aprendizaje automático, es decir no "presuponer" la naturaleza del mismo. Debe explorarse la construcción de diversos modelos mediante la aplicación de los conceptos.

Esta actividad evaluativa consiste en aplicar los conceptos de aprendizaje no supervisado en un conjunto de datos desconocidos y en un conjunto de datos en los que esta familiarizado.

Los algoritmos que van a explorar son los siguientes:

1. Mountain clustering
2. Subtractive clustering
3. K means clustering
4. Fuzzy c-means clustering
5. Otro algoritmo que les parezca interesante (les enviare varios del estado del arte y sus codigos)

Si bien, el detalle de los anteriores NO fue expuesto en clase, se enviará el material sobre el cuál los estudiantes podran revisar las fórmulas de cada uno de ellos. Los codigos son muy populares y existen muchas implementaciones de los mismos.

Los conceptos generales se pueden revisar directamente del libro "Machine Intelligence for decision making" (en borrador para uso de los estudiantes de este curso y bajo edicion por Springer), y de las diversas fuentes citadas en el libro con artículos científicos y otros libros mas especializados en cada tema.

<https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68>

Para comenzar a realizar su proceso de aprendizaje (me refiero a practicar en datos juguete antes de abordar el problema real), el estudiante puede usar conjuntos de datos sintéticos que haya como ejemplo en cualquier programa o suite, les sugiero el iris dataset pero pueden usar alguno de estos.

UCI Datasets:

<https://archive.ics.uci.edu/ml/datasets.php>

Wisconsin Breast Cancer Database

(small) soybean dataset <http://mlr.cs.umass.edu/ml/datasets/>

Thyroid dataset

<https://archive.ics.uci.edu/ml/datasets/>

Ecoli dataset

<https://archive.ics.uci.edu/ml/datasets/>

Wine dataset

<https://archive.ics.uci.edu/ml/datasets/Wine>

<https://archive.ics.uci.edu/ml/datasets/Iris>

Cuando esten listos, abordar el problema con datos reales.

## 2. Contenidos

El ÚNICO flujo de analisis del espacio de datos es el siguiente:

1. Codigo con un readme, requirements, que automaticamente cree un ambiente, un main que tenga como unico argumento de entrada un archivo con datos de cualquier naturaleza, carpeta de fuente con las funciones que el codigo main requiere ejecutar
2. Que yo pueda ejecutar directamente en la terminal y en el caso que lo acompañen con un cuaderno, que sea completamente legible para ejecutarlo en un ambiente de nube
3. Elegir el conjunto de datos, leerlo, preprocesar los datos categoricos como hot encoding, y caracterizarlo, es decir deben contar en el reporte N, n, m, l y demas cosas.
4. Recuerden que la meta es trabajarlo para poder aplicar los conceptos.
5. Realizar el preprocesamiento de los datos, por ejemplo limpiar los NaN y normalizacion.
6. El proceso de normalizacion puede ser como quieran y si no quieren como les explique en clase entonces pueden hacer lo que quieran. Lo IMPORTANTE es que no induzcan una distribución sobre los datos.

7. Extraer características de los datos. Para ello deben revisar el capítulo número 2 del libro *Machine Intelligence for Decision making* o en su defecto revisar un libro completo de álgebra lineal, repasar análisis, repasar probabilidad, repasar series de tiempo y los libros de 10 lectures on wavelets y a Wavelet tour. La idea es generar un espacio de características de dimensiones superiores al original para que podamos aplicar los conceptos de exploración y poder hacer la comparación con las dimensiones originales y las dimensiones embebidas. Pueden extraer características espaciales, temporales, frecuenciales, estadísticas. Algunos ejemplos de características son medias móviles, cruces por cero, energías, transformadas de Fourier y wavelet.
8. Usar al menos 5 formas diferentes de medir distancias en espacios  $n$  dimensionales (euclidiana, Mahalanobis, Manhattan, Coseno, Norma  $l_p$ ) esto es práctica de análisis numérico
9. Calcular la matriz de distancias dependiendo del método de no supervisión 2.1
10. Calcular los subconjuntos candidatos a ser clases de supervisión

## 2.1. Maquinas de aprendizaje

Deben aplicarlo teniendo en cuenta que deben:

- Técnicas de agrupamiento exploratorias como Mountain y Subtractive.
- Variar los parámetros asociados con el radio y seleccionar un número factible de grupos. Realizar las tablas para que puedan evaluar la información que obtienen del espacio haciendo los cambios en los parámetros.
- Evaluar los algoritmos de  $k$ -means y FC-means.
- Evaluar otro que les guste.
- Evaluar los índices de validación intra cluster y extraccluster PARA LOS DATOS JUGUETE porque en la vida real es complicado.

## 3. Para la entrega final

1. Aprendizaje 1: Aprender los datos usando técnicas de agrupamiento en el espacio de altas dimensiones aplicando los métodos de la sección 2.1.
2. Aprendizaje 2: Aprender los datos usando técnicas de agrupamiento en el espacio original aplicando los métodos de la sección 2.1.
3. Aprendizaje 3: Aprender los datos usando técnicas de agrupamiento en el espacio de los embebimientos aplicando los métodos de la sección 2.1.

4. Utilizar el algoritmo de embebimiento UMAP (ver la descripción del algoritmo en el libro y si no quieren, entonces en el paper original e interiorizarlo y no si quieren, pues no importa) para generar un espacio de dimensiones 3D o 2D para aprender sobre el. El algoritmo está disponible en varios lenguajes de programación.
5. Realizar la comparación de los modelos en los tres espacios de dimensiones.

Deben argumentar por qué razones el flujo de trabajo que han definido es el adecuado, usando argumentos de los que se han trabajado en clase y los que se encuentran en la literatura. Para ello es que van a trabajar en los conjuntos de datos y luego llegar a un acuerdo.

Se debe entregar:

- Un solo Documento informe tipo paper en formato IEEE
- El código final aplicado en el conjunto de datos reales