Lineamientos para evaluación del curso Inteligencia Artificial Tema No Supervisados

O. L. Quintero

28 de septiembre de 2023

Resumen

Este documento contiene los lineamientos para la presentación de la actividad evaluativa el domingo 1 de octubre 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. Cajas de Juan
2. Vecinos de Sofia

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 (PARA EL AVANCE):

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 y demas cosas. Recuerden que la meta es trabajarlo para poder aplicar los conceptos.

4. Realizar el preprocesamiento de los datos, por ejemplo limpiar los NaN y normalizacion. 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.

5. Realizar el analisis estadistico descriptivo. Pacho y Nico 1 y 2. Con el fin de identificar si el problema es tratable o no con las tecnicas de modelado que vieron durante su carrera asi no gastamos polvora en gallinazos.

6. Usar al menos 5 formas diferentes de medir distancias en espacios n dimensionales (euclidiana, mahalanobis, manhattan, coseno, norma lp) esto es practica de analisis numerico

7. Calcular la matriz de distancias

8. Como salida deben entregarme al menos dos tipos de agrupacion artesanal que vimos en clase: cajitas y vecinos, cada uno elige el criterio para armar las cajitas y los vecinos

Para la entrega final

1. Usar un autoencoder para aumentar el espacio de caracteristicas que llamaremos **.ªltas dimensiones"**

2. Usar un autoencoder para aumentar el espacio de caracteristicas que llamaremos **"bajas dimensiones"**

3. Utilizar el algoritmo de embebimiento UMAP (ver la descripcion 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 esta disponible en varios lenguajes de programación.

4. Aprendizaje 1: Aprender los datos usando tecnicas de agrupamiento en el espacio de altas dimensiones aplicando los metodos de la seccion 2.1.

5. Aprendizaje 2: Aprender los datos usando tecnicas de agrupamiento en el espacio original aplicando los metodos de la seccion 2.1.

6. Aprendizaje 3: Aprender los datos usando tecnicas de agrupamiento en el espacio de los embebimientos aplicando los metodos de la seccion 2.1.

7. Realizar la comparación delos modelos en los tres espacios de dimensiones.

2.1. Maquinas de aprendizaje

Deben aplicarlo teniendo en cuenta que deben:

- Tecnicas de agrupamiento exploratorios como mountain y subtractive.

- Variar los parametros asociados con el radio y seleccionar un numero factible de grupos. Realizar las tablas para que puedan evaluar la inforamcion que obtienen del espacio haciendo los cambios en los parametros.
- Evaluar los algoritmos de k-means y FC-means.
- Evaluar otro que les guste.
- Evaluar los indices de validacion intra cluster y extracluster PARA LOS DATOS JUGUETE porque en la vida real es complicado.

Deben argumentar por que 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
- Los Codigos elaborados por cada uno de los estudiantes para los datasets juguete.
- El codigo final aplicado en el conjunto de datos reales