UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Inteligencia Artificial - Laboratorio 1 -

Semestre I - 2024

Instrucciones:

- Esta es una actividad en grupos de no más de 3 integrantes.
  - Recuerden **unirse al grupo de canvas**
- No se permitirá ni se aceptará cualquier indicio de copia. De presentarse, se procederá según el reglamento correspondiente.
- Tendrán hasta el día indicado en Canvas.
  - No se confíen, aprovechen el tiempo en clase para entender todos los ejercicios y avanzar lo más posible.

Task 1 - Regresión Lineal

Responda a cada de las siguientes preguntas de forma clara y lo más completamente posible.

1. Considera un modelo de regresión lineal con dos características, X_1 y X_2, y sus pesos correspondientes w_1 y w_2. Si el modelo predice una salida y mediante la ecuación $y = 2w_1X_1 + 3w_2X_2 + 1$, ¿cuál es la interpretación del coeficiente $3w_2$ en el contexto del modelo?
2. Explica el concepto de **multicolinealidad** en el contexto de la regresión lineal. ¿Cómo afecta la multicolinealidad a la interpretación de los coeficientes de regresión individuales?

Task 2 - Clasificación de Sitios de Phishing Regresión Logística y KNN

Como bien se sabe, los sitios web de phishing siguen siendo de las formas más efectivas para los cibercriminales para robar información. Por ello, aprender a identificar de forma proactiva aquellos sitios sospechosos para poder bloquearlos es una tarea importante. Bajo este contexto, se le ha solicitado que cree modelos para la identificación de sitios. Para ello:

- Usará el dataset proporcionado en Kaggle en el siguiente enlace
  - Recuerden que pueden descargar el código directamente con llaves generadas desde Kaggle o bien pueden ingresar al enlace y descargar el archivo como usualmente lo hace con cualquier otro documento
- La especificación de las columnas la encuentran en el siguiente enlace
  - Las columnas se especifican a partir de la página 6
- Deben hacer una breve exploración con los datos. Esto implica, pero no está limitado a:
  - Hacer encoding de las variables que se necesiten
  - Revisar si el dataset está balanceado, caso no estarlo, aplicar alguna técnica para balancearlo lo más y mejor posible
  - Escalar las variables si considera necesario
  - Selección de variables
- Recuerden hacer el split para training, testing y si consideran necesario para validation
  - 80% training
  - 20% testing
    - 10% validation si lo necesitan
- Recuerde definir de forma clara y razonada (es decir, diga el por qué de su elección) de una métrica de desempeño principal

Task 2.1 - Regresión Logística

Implemente desde cero el algoritmo de Regresión Logística. Para ello considere lo siguiente

- Recuerde implementar el algoritmo de gradiente descendente, tomando en cuenta parámetros como el learning rate y épocas
- Utilice el dataset proporcionado para mostrar el funcionamiento de su algoritmo
- Provea una métrica de desempeño, justificando su elección
- Grafique los grupos encontrados

    - Puede usar solamente dos variables para mostrarlos en un plano cartesiano
- Mencione, como comentario las consideraciones extras que tuvo que tomar en cuenta durante la realización de su implementación

Para este task **no usen librerías**, sino implementen el algoritmo por ustedes mismos. Además, **evite el uso de herramientas de AI generativas (ChatGPT)**.

Luego:

- Repita los pasos para entrenar su modelo, pero ahora usando librerías, y compare los resultados.
- Para esta parte sí puede usar herramientas de AI generativas.
- Responda:
    - ¿Cuál implementación fue mejor? ¿Por qué?

Task 2.2 - K-Nearest Neighbors

Implemente desde cero el algoritmo de K-Nearest Neighbors. Para ello considere lo siguiente

- La distancia entre puntos debe ser la dada por la forma de la distancia Euclidiana
- Utilice el dataset proporcionado para mostrar el funcionamiento de su algoritmo
- Provea una métrica de desempeño, justificando su elección
- Grafique los grupos encontrados
    - Puede usar solamente dos variables para mostrarlos en un plano cartesiano
- Mencione, como comentario las consideraciones extras que tuvo que tomar en cuenta durante la realización de su implementación

Para este task **no usen librerías**, sino implementen el algoritmo por ustedes mismos. Además, **evite el uso de herramientas de AI generativas (ChatGPT)**.

Luego:

- Repita los pasos para entrenar su modelo, pero ahora usando librerías, y compare los resultados.
- Para esta parte sí puede usar herramientas de AI generativas.
- Responda:
    - ¿Cuál implementación fue mejor? ¿Por qué?

Nota: Para las partes donde se solicita que usen librerías, consideren usar Sckit-learn. Este tiene implementados tanto regresión logística como KNN. Por favor, **asegúrese de leer la documentación** de estas librerías para entender mejor los hiperparametros. Siempre recuerde, si tiene alguna duda o consulta, por favor comuníquese con el catedrático.

Entregas en Canvas

1. Jupyter Notebook respondiendo a cada task.
    a. Incluyendo las preguntas del task 1
    b. Dentro del Jupyter Notebook deben colocar el link a su repositorio en GitHub. Este puede permanecer como privado hasta la fecha de entrega.

Evaluación

1. [0.5 pt] Task 1
    a. [0.25 pts] Cada pregunta
2. [4.5 pts.] Task 2
    a. [1.5 pts] Cada modelo implementado desde cero (total 3pts)
    b. [0.75 pts] Cada modelo implementado usando librerías (total 1.5pts)