Profesor: Yasmany Prieto

Tarea 1: Aprendizaje supervisado. Diseño y evaluación de clasificadores.

Objetivos:

- Diseñar y evaluar clasificadores para problemas reales.
- Analizar y presentar un artículo científico del tema.

Tiempo de desarrollo de la tarea:

20 días

Equipos:

De **hasta 3 estudiantes**

Enunciado:

Problema 1.

Los mensajes de spam son mensajes no solicitados y no deseados que se reciben ya sea en una dirección de correo electrónico o una cuenta de mensajería instantánea. Los mensajes spam representan uno de los mayores problemas en la comunicación electrónica y hasta el 50% de todas las comunicaciones por correo electrónico. Frecuentemente son fuente de publicidad que se envía de forma automática a un número masivo de direcciones. Sin embargo, también pueden formar parte de los pasos iniciales de un esquema de estafa. Dado lo anterior es relevante la detección de estos mensajes nocivos para evitarlos.

La base de datos reportada en (https://archive.ics.uci.edu/dataset/94/spambase) contiene una descripción de 4601 muestras de correos. De cada muestra se obtuvieron 57 características y una variable binaria objetivo que denota si la muestra representa spam (1) o no-spam (0). Se desea construir un clasificador capaz de identificar correos spam, de manera que los identifique antes de que lleguen al usuario.

i) Describa el dataset, identificando las características y la variable de clasificación, que tipo de datos es cada una, y que valores toma la variable de clasificación. ¿El dataset presenta datos faltantes?

ii) Para el clasificador que propone como solución modifique los parámetros que tiene disponibles en la implementación de *sklearn*. Construya una tabla como la que aparece a continuación y determine cuáles son los valores óptimos de los parámetros (pruebe al menos con 5 valores de los parámetros). Tome uno de los parámetros y grafique su relación con la métrica de desempeño. Esta gráfica debe obtenerse tanto para los datos de entrenamiento, como los de evaluación. Interprete estas gráficas de acuerdo con el posible *underfitting* u *overfitting* del modelo.

Parámetro 1	…	Parámetro n	*accuracy promedio por clase*
…			
…			

iii) Del clasificador que propone como solución del problema entregue la matriz de confusión, así como los valores de las métricas de *recall* y *precision* de cada clase. Interprete el valor de *recall* y *precisión* de cada clase. Obtenga las métricas de *accuracy* y *accuracy promedio por clase,* comente la diferencia entre estos valores.

iv) Ocupe alguno de los métodos vistos en clase para reducir la dimensión de las características del problema y evalúe su impacto en las soluciones obtenidas (compare las métricas de desempeño). Pruebe al menos con la mitad de las características y con dos características.

v) Ocupe algún método de balanceo de clases de ser necesario y evalúe su impacto en las soluciones obtenidas (compare las métricas de desempeño y la matriz de confusión).

Problema 2.

Escoja un artículo científico donde se emplee al menos un método de reconocimiento de patrones con aprendizaje supervisado, para resolver algún problema práctico. De este artículo extraiga la siguiente información:

i) ¿Cuál es el problema práctico que se busca resolver? ¿Por qué es relevante? ¿Por qué se propone resolverlo mediante métodos de reconocimiento de patrones?

ii) ¿Qué métodos de clasificación se emplean en el trabajo? ¿Por qué éstos y no otros?

iii) ¿Cuál es la metodología empleada en el trabajo? ¿Cómo se diseñan los experimentos? ¿Qué tipos de datos se emplean? ¿Qué métricas se usan para medir los resultados?, etc.

iv) ¿Cuáles fueron los resultados del trabajo y que se concluyó? Discuta gráficos, comparaciones, etc.

v) ¿Qué hubiese hecho usted diferente? ¿Por qué?

Incluya en el informe la referencia al artículo.

Entrega:

i) Se debe entregar un informe que contemple las respuestas a los Problemas 1 y 2.

ii) El análisis del artículo incluye una presentación donde se comunique la información extraída. La fecha de presentación es el **30 de mayo de 2024**.

Adicionalmente, se puede incluir de manera opcional un resumen sobre que se aprendió de la experiencia o si se experimentó alguna dificultad específica.

¿Cómo se evaluará mi trabajo?

Aspectos para evaluar	**Ponderación**
Descripción del dataset	0.2
Selección del clasificador óptimo. Ajuste de parámetros de clasificador.	1.2

Interpretación de la relación entre el valor de los parámetros del clasificador y el valor de la métrica de desempeño.	0.4
Obtención de matriz de confusión e interpretación de *recall* y *precision*.	0.3
Interpretación de diferencia entre *accuracy* y *accuracy promedio por clase.*	0.3
Uso de métodos para reducción de dimensionalidad. Interpretación de los resultados obtenidos.	0.4
Uso de métodos de balanceo de clases o justifique que no es necesario su uso. Interpretación de los resultados obtenidos.	0.4
El artículo científico escogido está relacionado con el uso de clasificadores y el aprendizaje supervisado.	0.5
Sintetiza correctamente toda la información que se solicita.	1.0
Propone alguna mejora a los resultados del artículo a partir de los conocimientos del curso.	0.3
Presentación del artículo en clase.	0.5
El informe es claro y organizado. Se usa vocabulario de la asignatura.	0.5
Total	6.0