Tarea 4

Problema 1

Considere al conjunto de datos Ideology Data, consistente en la autoidentificación ideológica de 5916 personas, incluyendo además 6 variables demográficas.

Se pide:
a) Describa numérica y gráficamente las principales características del conjunto de datos, incluyendo, por ejemplo, la distribución de valores dentro de cada atributo, valores faltantes, distribución de la clase objetivo. Esto tiene como objeto entender el conjunto con el que se está trabajando.
b) Muestre cómo se agrupan las instancias, distinguiendo según su identificación ideológica. Intente con cada atributo, y luego sus combinaciones.
c) Aplique k-means para agrupar las instancias en 3 grupos. ¿Están estos clusters vinculados a alguna agrupación ideológica? Comente.
d) Intente predecir la identificación ideológica a partir de los atributos, utilizando K-nearest neighbors. ¿Cómo se relacionan los resultados obtenidos con lo visto en el punto anterior? ¿Qué puede concluirse sobre el conjunto de datos?

Problema 2

Un modelo de lenguaje es un método para asignar la probabilidad a una secuencia de caracteres o palabras. Una forma de construir un modelo de lenguaje orientado a caracteres, utilizando aprendizaje automático, es entrenar una Red Neuronal Recurrente que intente predecir el siguiente caracter a partir de la secuencia de caracteres anteriores. Esta red es entrenada sobre un corpus de texto suficientemente grande.

Se pide:
a) Reproduzca el trabajo descrito en [1], y en particular su implementación en PyTorch [2].
b) Describa su funcionamiento, incluyendo la arquitectura, entradas y salida, y parámetros
c) Utilice el método para generar texto de acuerdo al modelo de lenguaje, a partir de un corpus de entrenamiento de su elección. Presente ejemplos del texto generando con diferentes parámetros de sampleo.
d) Grafique las modificaciones en la función de pérdida (tanto en el conjunto de entrenamiento como en el de validación), en función de las épocas de entrenamiento. ¿Es posible que su modelo esté sobreajustando?
e) Presente ejemplos de textos generados a partir de los modelos intermedios producidos durante el entrenamiento. Comente.

[1] "The Unreasonable Effectiveness of Recurrent Neural Networks". A. Karpathy
[2] "Character-Level LSTM in PyTorch". F. Paulin.

Observaciones

- El objetivo principal de esta tarea es el análisis de datos y resultados obtenidos, más que la construcción de algoritmos. Por lo tanto, se evaluará especialmente la claridad de los análisis presentados (que **no** es proporcional al largo del informe).
- Para tarea no es necesario implementar algoritmos que ya existan en biliotecas, pero debe presentarse muy claramente cómo son utilizados.
- Para el problema 2, sugerimos ejecutar los notebooks en la plataforma Google Colab, o utilizar una máquina con GPU, para reducir los tiempos de ejecución.

Entregables para ambos problemas

- Informe con las pruebas realizadas y los resultados obtenidos.
- El informe a entregar debe ser un Jupyter Notebook.

Fecha límite de entrega

Lunes 8 de junio (inclusive)