Redes Neuronales Artificiales

Trabajo Práctico 2.

1. Introducción

El objetivo de este trabajo práctico es utilizar distintos modelos de redes neuronales artificiales sobre el mismo conjunto de datos para una tarea de clasificación. Todos modelos estarán basados en variantes de aprendizaje no supervisado. Los datos pertenecen a un problema real y es posible que las categorías no estén claramente diferenciadas. Se espera que respetando las consignas se construyan modelos adecuados para cada problema, que puedan aprender sobre las instancias de entrenamiento y que tengan una capacidad aceptable para generalizar sobre otras. Tanto la arquitectura de los modelos, la elección de parámetros, las técnicas de entrenamiento y testeo, y los métodos para presentar los resultados deberán ser documentados y entregados en un informe en donde se deberán justificar las decisiones tomadas y explicados los resultados obtenidos.

2. Problema

El conjunto de datos consiste en documentos con descripciones de texto correspondientes a compañías Brasileñas clasificadas en nueve categorías distintas. Los textos originales fueron preprocesados para obtener un tipo de representación conocida como Bolsa de Palabras (Bag-of-Words, o simplemente BoW). En este tipo de formato cada documento es representado mediante un vector en donde cada dimensión corresponde a una palabra específica y su valor está dado por la cantidad de apariciones de esa palabra en el documento. Para mejorar la representación las palabras más comunes (artículos, preposiciones, etc) no son tenidas en cuenta. Notar que al representar un documento de esta forma no se está teniendo en cuenta el orden de las palabras, solo su frecuencia de aparición.

El conjunto de datos se encuentra en formato CSV sin encabezado y contiene 900 entradas distribuidas uniformemente entre las 9 categorías. Cada entrada representa un documento y consiste en el número de categoría (1 a 9) más 850 atributos correspondientes a apariciones de palabras. Tener en cuenta que el conjunto de datos es disperso (casi todos los valores son ceros) y que el número de categoría corresponde a la actividad principal de la empresa, no necesariamente la única.

El problema a resolver será clasificar automáticamente los documentos utilizando distintos modelos de aprendizaje hebbiano no supervisado y competitivo. Notar que por tratarse de aprendizaje no supervisado para el entrenamiento solo se deben utilizar los valores de las frecuencias de palabras, la información de la categoría solo será utilizada para verificar la clasificación hecha por los modelos.

2.1. Reducción de dimensiones

Construir un modelo de red neuronal artificial con una arquitectura que permita reducir la alta dimensionalidad de las entradas en el conjunto de datos a solo 9 dimensiones. Para efectuar la reducción de

dimensión entrenar al modelo mediante las reglas de aprendizaje de Oja y de Sanger. Una vez realizado el entrenamiento por cada método representar gráficamente las respuestas utilizando 3 figuras en el espacio $\mathbb{R}^3$ en donde cada documento esté representado por un punto. Por ejemplo, si la respuesta de la red es el vector $Y \in \mathbb{R}^9$, la figura 1 tendrá por ejes a Y_1, Y_2, Y_3, la figura 2 a Y_4, Y_5, Y_6, y la figura 3 a Y_7, Y_8 e Y_9.

Para realizar la representación gráfica de los resultados tener en cuenta que se debe diferenciar claramente cada categoría (por ejemplo con distintos colores).

Además tener en cuenta que es posible que no todas las categorías aparezcan claramente diferenciadas en todas las figuras, y que distintas instancias de entrenamiento deberían producir resultados similares.

2.2. Mapeo de características

Construir un modelo de mapeo de características auto-organizado que clasifique automáticamente los documentos en un arreglo de dos dimensiones. Tener en cuenta que para poder realizar una buena clasificación se debe contar con suficientes unidades de salida y con suficientes instancias de datos por unidad.

Se recomienda experimentar entrenando distintos modelos, por ejemplo variando la cantidad de unidades de salida y los parámetros adaptativos utilizados en el aprendizaje, para tratar de obtener una buena solución.

Una vez realizado el entrenamiento representar gráficamente en un mapa de características los resultados, señalando para cada unidad de salida cuál es la categoría que más la activa (por ejemplo con distintos colores).

3. Detalles de la entrega

La entrega deberá consistir de un informe general o uno por problema, en formato *jupyter-notebook*/*colab,* que contenga el código completo demostrando el entrenamiento y la evaluación de los modelos, como también una descripción de cómo se implementó la solución, las decisiones tomadas, y su justificación.

Esto significa que en el informe se debe describir qué modelo de red neuronal artificial fue adoptado como solución para cada problema, especificando la arquitectura elegida, el método de entrenamiento utilizado, tipo de procesamiento a los datos, y los resultados obtenidos. Es importante también que esté documentado el proceso que condujo a los resultados finales, justificando las decisiones tomadas y las conclusiones a las que se hubieren llegado. Este informe deberá ser breve y conciso, es decir, que no debe contener explicaciones demás, pero tampoco de menos. Por ejemplo, bloques de código copiados sin ninguna descripción no son aceptables.

Como la evaluación del desempeño de estas soluciones depende en gran parte de una representación visual, es importante que se incluyan figuras que muestren claramente los resultados junto a simples explicaciones de cómo interpretarlas.

Para el desarrollo de estas soluciones se recomienda utilizar únicamente las librerías *numpy* y *matplotlib*. Adicionalmente se debe recurrir a la librería *requests*, junto al enlace provisto con el enunciado, que permite cargar los datos desde un origen externo.

```
import requests
r = requests.get('https://...url.../datos.csv')
data = numpy.loadtxt( r.iter_lines(), delimiter=',')
```

La inclusión injustificada de otras librerías puede tener un impacto negativo en la calificación del trabajo.

Se espera también que al volver a ejecutar todas las celdas del informe se obtengan resultados similares a los presentados. En caso de que existan dificultades en la ejecución del código para la evaluación, el trabajo puede llegar a ser rechazado.

El material correspondiente a la entrega deberá enviarse a la dirección:

entregas.redneu@gmail.com

La fecha de entrega será confirmada en la presentación de este trabajo y/o por la lista de correo de la materia.