Redes Neuronales Artificiales

Trabajo Práctico 2.

1. Introducción

El objetivo de este trabajo práctico es que puedan utilizar su implementación de distintos modelos de redes neuronales artificiales basados en aprendizaje no supervisado sobre el mismo conjunto de datos perteneciente a un problema real. Se espera que, respetando las consignas, se construyan modelos adecuados 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 frecuencias 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. Los datos pertenecen a un problema real y es posible que contengan errores.

El problema a resolver será una reducción de dimensión de los documentos utilizando distintos modelos de aprendizaje hebbiano no supervisado y *opcionalmente* una clasificación automática mediante aprendizaje 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 etiquetar unidades y verificar la clasificación hecha por los modelos.

2.1. Reducción de Dimensiones (Obligatorio)

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 modelos 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 puntos de distintos colores).
- Para apreciar la distribución espacial pueden ser necesarias varias vistas.
- Verificar que se obtengan resultados similares para varias instancias de entrenamiento.

La resolución y entrega de este ejercicio es OBLIGATORIA.

2.2. Mapeo de Características (Opcional)

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.

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).
- Cuál es la diferencia entre los datos de entrenamiento y validación (por ejemplo en distintas figuras).
- Comparar los resultados variando la cantidad de unidades de salida.
- Comparar los resultados variando los parámetros de entrenamiento.

La resolución y entrega de este ejercicio es OPCIONAL.

3. Detalles de la entrega

La entrega deberá consistir de, al menos, un programa ejecutable (script en python), código completo (todos los módulos programados), un modelo entrenado por problema, y un informe escrito.

Al programa se le debe poder indicar un nombre de archivo para el modelo y un nombre de archivo para los datos. Esto puede hacerse por linea de comando, archivo de configuración, o nombres de variables (si son claramente indicadas en el código).

Si el archivo para el modelo no existe se procederá a entrenar un nuevo modelo con las instancias provenientes del archivo de datos. Para este nuevo modelo se puede utilizar una arquitectura y método de entrenamiento predeterminados. Si existe el archivo para un modelo entrenado entonces el conjunto de

datos se utilizará para testeo. En ambos casos el programa deberá mostrar de forma clara el desempeño del modelo. También es posible entregar dos programas distintos, uno para cada problema.

En caso de que existan dificultades en la ejecución del programa para la evaluación, el trabajo puede llegar a ser rechazado. Por esto se recomienda enfáticamente utilizar el lenguaje y librerías vistas en la materia, dependencia mínima de otras librerías externas, explicar claramente su forma de uso, y testear su ejecución en distintas máquinas y/o plataformas antes de enviarlo.

El informe deberá ser breve y conciso (no más de 4 páginas). En el mismo 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 de experimentación que condujo a los resultados finales, en donde se deben justificar las decisiones tomadas y las conclusiones a las que se hubieren llegado.

Todo el material correspondiente a la entrega (informe, código, etc) deberá enviarse en un archivo comprimido a la dirección:

entregas.redneu@gmail.com