

Instrucciones Práctica Inteligencia de Negocio

Base de Datos: MNIST

January 15, 2021

1 Introduction

En este documento recogemos las instrucciones para los alumnos de la asignatura de Inteligencia de Negocio para la realización de la práctica alternativa al examen.

2 Conjunto de datos: MNIST

La práctica consiste en un trabajo de clasificación sobre un conjunto de datos. En vuestro caso es sobre la base de datos MNIST.

Esta base de datos presenta la siguiente estructura:

- Contiene 10 clase de números manuscritos en blanco y negro del 0 al 9.
- Presenta un total de 70000 imágenes de las cuales 60000 son para entrenamiento y 10000 para test.

En [1] tenéis toda la información sobre este conjunto de datos. Además, en [2] se encuentran las diferentes características de este conjunto de datos y resultados sobre este conjunto de datos en función al clasificador y al preprocesamiento realizado.

Además, en dicha dirección os indica cómo leer los datos en el caso de que uséis Python como lenguaje de programación. Existen diferentes lenguajes de programación para usar estos datos, pero Python es uno de los más usados. Lo primero que tendréis que hacer es hacer la lectura de los datos o bien importando los datos usando una librería de Python o bien usando otros métodos como podéis ver en el siguiente enlace [4], pero no tendríais que tener ningún problema, pues es sencillo.

Esta base de datos es sencilla para obtener buenos resultados, por lo que os recomiendo que hagáis algunas propuestas alternativas y que apostéis por conceptos más innovadores. En [3] se proporcionan varias técnicas de ensemble y analizan varios esquemas usando como hilo conductor MNIST. Finalmente, se propone un nuevo método de fusión de información con la base de datos MNIST.

La realización de este estudio ha hecho que se obtenga el récord mundial de menos imágenes mal clasificadas en este problema, por lo que esta referencia es bastante interesante para podáis ver diferentes formas de afrontar este problema.

Algunas ideas que podéis desarrollar son las siguientes:

- Estudio de diferentes técnicas de preprocesamiento para el conjunto de datos.
- Estudio de algoritmos para ver cuál de ellos clasifica mejor este conjunto.
- Aplicar los dos conceptos anteriores de forma conjunta.
- Desarrollar métodos de ensemble para intentar superar los resultados actuales.

References

- [1] Yann LeCun, Corinna Cortes, and CJ Burges. “MNIST handwritten digit database”. In: *ATT Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist> 2 (2010).
- [2] *MNIST Information*. <https://www.tensorflow.org/datasets/catalog/mnist>. Accessed: 2021-01-15.
- [3] “MNIST-NET10: A heterogeneous deep networks fusion based on the degree of certainty to reach 0.1% error rate. Ensembles overview and proposal”. In: *Information Fusion* 62 (2020), pp. 73–80.
- [4] *Read MNIST*. <https://stackoverflow.com/questions/40427435/extract-images-from-idx3-ubyte-file-or-gzip-via-python>. Accessed: 2021-01-15.