Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Instituto de Ingeniería Matemática y Computacional
Semestre 2020-I
Introducción a la Ingeniería Matemática IMT1001

Profesor: Elwin van 't Wout (e.wout@uc.cl)

Proyecto: Ciencia de Datos

Los algoritmos de Big Data tienen aplicaciones en muchos distintos ámbitos de la ingeniería. Un tema interesante es el reconocimiento automático de carácteres como letras y números. Estos algoritmos son usados, por ejemplo, en lectores de patentes de autos o en la digitalización de documentos escritos a mano. En este proyecto, vamos a investigar métodos de clasificación para este propósito.

Exploración de la metodología

La biblioteca `scikit-learn` tiene una base de datos disponible que contiene números escritos a mano. Dado que la resolución de las imágenes es baja, se puede analizar sus características rápidamente.

Ejercicio 1. Para comenzar, un buen tutorial está disponible en `https://scipy-lectures.org/packages/scikit-learn/auto_examples/plot_digits_simple_classif.html`. Corren el Jupyter Notebook y analizen la base de datos. Expliquen el formato de las imágenes.

Ejercicio 2. Dado la imposibilidad de graficar datos en altos dimensiones, se require una *reducción de dimensionalidad*. En el tutorial, el método de PCA está usado. Expliquen como funciona el método de PCA y su relación con valores singulares.

Ejercicio 3. Hay una variedad a métodos de reducción de dimensionalidad, como por ejemplo t-SNE. Comparen el desempeño de distintos métodos de reducción de dimensionalidad para este base de datos.

Clasificación de números

La base de datos MNIST (ver `http://yann.lecun.com/exdb/mnist/`) incluye un gran cantidad de imágenes de números escritos a mano. Este nos permite comparar el desempeño de distintos clasificadores. Primero, visualizen la base de datos con los métodos de los ejercicios anteriores.

Ejercicio 4. Programen varios clasificadores para este base de datos. Expliquen el fundamento matemático de los métodos usados. Comparen la matriz de confusión y otras medidas de desempeño, y discuten los resultados.

Desbalanceo de letras

La base de datos EMNIST (ver `https://www.nist.gov/itl/products-and-services/emnist-dataset`) contiene imágenes de letras escritas a mano también. Dado que las imágenes tienen el mismo formato que la base de datos MNIST, se puede correr los mismos métodos que en los ejercicios anteriores.

Ejercicio 5. Un problema para letras que no ocurre en números es que algunas letras son más comunes que otras, lo cual genera un desbalanceo en la base de datos (ver `https://arxiv.org/pdf/1702.05373v1.pdf`). Expliquen el impacto que el desbalanceo tiene a la hora de hacer clasificación. Investiguen maneras para mejorar los clasificadores en el caso de datos desbalanceados.

Extensiones

Opcionalmente, se puede extender el proyecto con el siguiente.

Ejercicio 6. Escriben números/letras de mano y convierten las imágenes al estándar de las bases de datos usados. Hagan la clasificación de tus propios letras. ¿Quién del equipo tiene el mejor letra?

Ejercicio 7. Hay muchas bases de datos disponible en internet (ver, por ejemplo, `https://lionbridge.ai/datasets/15-best-ocr-handwriting-datasets/`)), que incluyen otros tipos de símbolos. Investiguen clasificadores para otras bases de datos y relacionen el desempeño con las bases de datos usados en los ejercicios anteriores.