Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Instituto de Ingeniería Matemática y Computacional
Semestre 2023-I
Introducción a la Ingeniería Matemática IMT1001

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

Proyecto: Aprendizaje Automatizado

Los algoritmos de Aprendizaje Automatizado (*machine learning*) tienen aplicaciones en muchos distintos ámbitos de la ingeniería. En general, los algoritmos de aprendizaje supervisado aprenden una tarea específica a través de un gran cantidad de ejemplos. Una aplicación interesante es el reconocimiento automático de carácteres como letras y números. Estos algoritmos son usados en tecnologías tales como lectores de patentes vehiculares 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 librería `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 analicen la base de datos. Expliquen como se puede almacenar y analizar imágenes en Python.

Ejercicio 2. Dado la imposibilidad de graficar datos en altas dimensiones, se requiere una *reducción de dimensionalidad*, como por ejemplo el método de PCA. Expliquen la matemática detrás del método PCA.

Ejercicio 3. Existe una variedad a métodos de reducción de dimensionalidad disponible en `scikit-learn`. Comparen el desempeño de distintos métodos de reducción de dimensionalidad para esta base de datos.

Clasificación de números

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

Ejercicio 4. Implementen varios clasificadores para esta 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 el hecho que algunas letras son más comunes que otras, lo cual genera un desbalance en la base de datos (ver `https://arxiv.org/pdf/1702.05373v1.pdf`). Expliquen el impacto del desbalance al desempeño de los clasificadores. Investiguen maneras para mejorar los algoritmos para datos desbalanceados y revisen si efectivamente mejoren el desempeño.

Extensiones

Opcionalmente, se puede extender el proyecto con el siguiente.

Ejercicio 6. Escriben números/letras a mano en un papel. Tomen una foto, convierten las imágenes al estándar de las bases de datos usados, y clasifiquen tu propia escritura.

Ejercicio 7. Hay otras bases de datos disponible en internet que incluyen, por ejemplo, letras de otras idiomas o símbolos matemáticos. Buscen otras bases de datos en internet (palabra clave: *Optical Character Recognition*) e investiguen clasificadores para estos datos.