

PRÁCTICA 3: Ajuste de Modelos Lineales

*Aprendizaje Automático || Curso 2018-2019*

**Alumno:** José María Sánchez Guerrero

**DNI:** 76067801Q

**Correo:** jose26398@correo.ugr.es

**Grupo:** A3 – Viernes 17:30

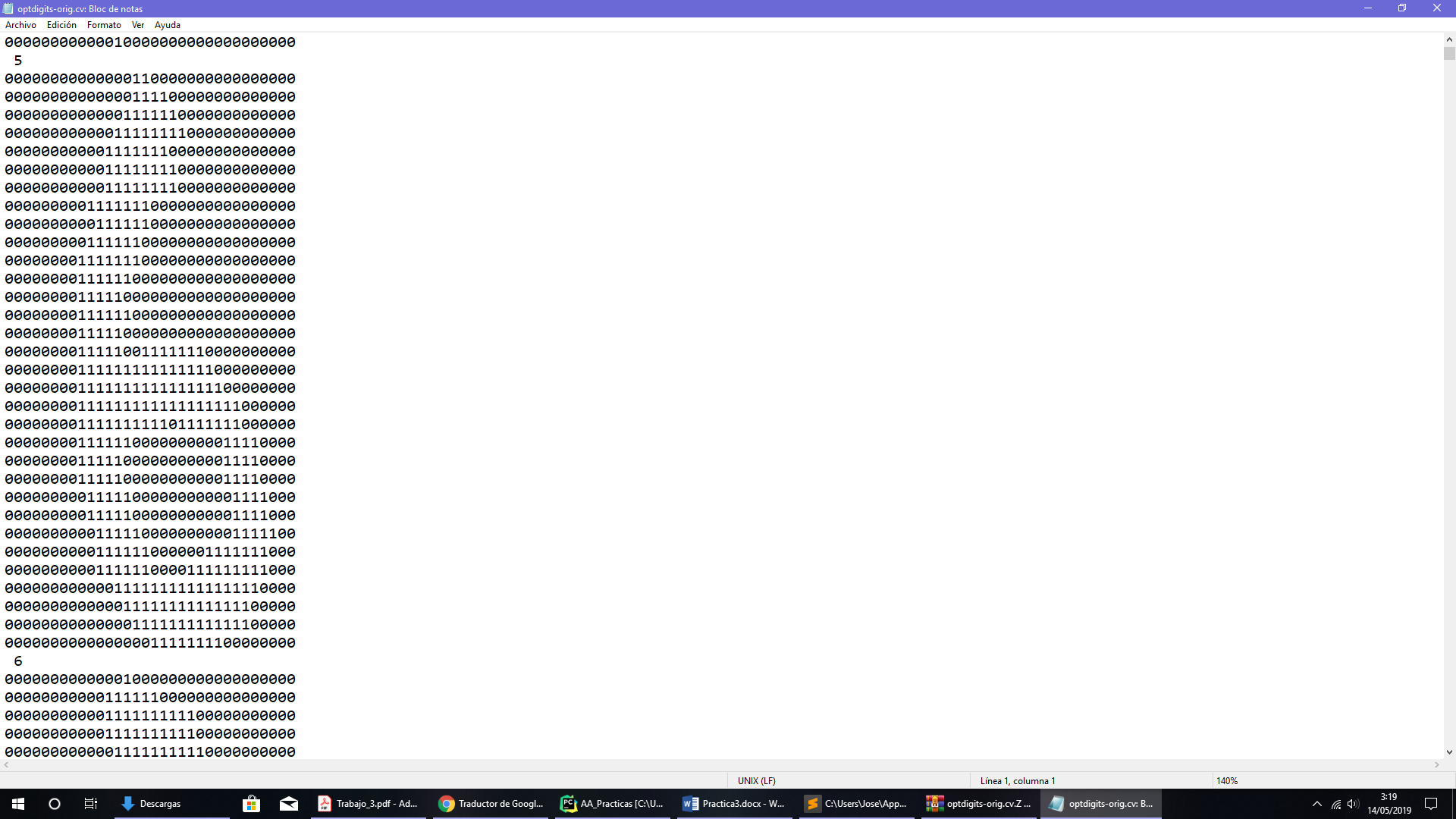
# 1. Comprender el problema a resolver

El objetivo de esta práctica será ajustar y seleccionar el mejor modelo lineal para un problema determinado, y posteriormente discutir los resultados (y el error) de éste modelo elegido. Para llevarlo a cabo, se utilizarán dos conjuntos de datos sacados del *“UCI Machine Learning Repository”*.

## Optical Recognition of Handwritten Digits Data Set

Este conjunto de datos está formado por una serie de filas que representan un número y 65 columnas, formadas por 64 características de un número entero (del 0 al 9) que estará en la última columna de cada fila.

Los datos originalmente son mapas de bits normalizados de dígitos manuscritos de un formulario preimpreso. Estos mapas son de tamaño 32x32 y cada uno de los bits estará activo o no, dependiendo de si en el formulario está pintado o no. Estos mapas de bits se dividirán en bloques no superpuestos de 4x4 y el número de bits activos en este bloque será el número que encontramos en nuestro dataset. Al ser de este tamaño, se generarán 64 valores enteros (número de características del dataset) que irán del 0 al 16. Esto reduce la dimensionalidad y da invarianza a pequeñas distorsiones.

Veamos un ejemplo:

0 0 0 11 9 0 0 0

Si transformamos la primera fila de matrices 4x4 en números enteros como acabamos de explicar, dará este resultado.

Estos 8 valores será los primeros que aparezcan en la fila correspondiente a este número, seguido de los 54 números que completan el mapa de bits.

Obviamente, seguirán el orden lógico de izquierda a derecha y de arriba abajo.

Como hemos dicho antes, el último número será el propio dígito manuscrito (en este caso se ve claramente que es el 6).

Los propios datos del repositorio ya vienen divididos en *training* y *test*. También te proporcionan los datos originales con los mapas de bits completos y sin normalizar, sin embargo, yo voy utilizar los que ya están procesados. Estos dos ficheros se llaman ‘optdigits.tra’ y ‘optdigits.tes’ y los utilizare en un directorio ‘datos’ localizado en el mismo sitio que el código.

Los datos de estos archivos son:

|  |  |  |
| --- | --- | --- |
| *Número de instancias:* |  |  |
| optdigits.tra | Training | 3823 |
| optdigits.tes | Testing | 1797 |

|  |  |  |
| --- | --- | --- |
| *Distribución de clases:* | | |
| Clase | **Nº ejemplos training** | **Nº ejemplos testing** |
| 0 | 376 | 178 |
| 1 | 389 | 182 |
| 2 | 380 | 177 |
| 3 | 389 | 183 |
| 4 | 387 | 181 |
| 5 | 376 | 182 |
| 6 | 377 | 181 |
| 7 | 387 | 179 |
| 8 | 380 | 174 |
| 9 | 382 | 184 |

## Airfoil Self-Noise Data Set

Este segundo conjunto de datos es más sencillo que el anterior. Las filas se componen sólo de 6 columnas, formadas por 5 características que determinan el nivel de presión sonora escalado, en decibelios. Este nivel será el valor de la última columna.

Los datos no están transformados ni modificados como en el ejemplo anterior, simplemente son los datos recogidos por la NASA para diferentes perfiles aerodinámicos NACA 0012. Estos perfiles comprenderán diferentes tamaños a varias velocidades en el túnel de viento y diferentes ángulos de ataque. La duración del perfil aerodinámico y la posición del observador fueron las mismas en todos los experimentos.

Veamos ahora a qué pertenece cada una de las 5 características mencionadas anteriormente, ordenadas por el número de columna al que pertenecen:

1. Frecuencia, en hercios.
2. Ángulo de ataque, en grados.
3. Longitud de la cuerda (trigonométricamente hablando), en metros.
4. Velocidad de flujo, en metros por segundo.
5. Grosor de desplazamiento lateral de aspiración, en metros.

Los datos vienen todos en un fichero ‘airfoil\_self\_noise.dat’ y, a diferencia del anterior sin dividir en *training* y *testing*. Este fichero también estará en el directorio ‘datos’ localizado en el mismo sitio que el código.

Los datos de estos archivos son:

|  |  |  |
| --- | --- | --- |
| *Número de instancias:* |  |  |
| airfoil\_self\_noise.dat | Training y testing | 1503 |

La distribución de clases, al ser tan extensa y tener tantos valores diferentes, no la podremos mostrar en una tabla como el anterior.

# 2. Preprocesado de los datos