

# APRENDIZAJE AUTOMÁTICO

Grado en Ingeniería Informática

Grupo 1

## Práctica 3



UNIVERSIDAD  
DE GRANADA

# ÍNDICE

Normas de desarrollo

Fecha de entrega

Definición del problema

Tareas

Bases de datos

Optical Recognition of Handwritten Digits Data Set

Airfoil Self-Noise Data Set

Pasos Machine Learning

Desarrollo de los puntos

1. Comprensión del problema
2. Preprocesamiento
3. Clases de funciones
4. Training, Test y Validación
5. Regularización
6. Modelos
7. Selección de modelo y ajuste
8. Métrica

# NORMAS DE DESARROLLO

- ▶ El código de cada ejercicio/apartado de la práctica se debe estructurar en un script Python incluyendo las funciones que se hayan definido. Cada script debe incluirse en un fichero distinto.
- ▶ Todos los resultados numéricos o gráficas serán mostrados por pantalla, parando la ejecución después de cada apartado. EL código NO DEBE escribir nada a disco.
- ▶ El path que se use en la lectura de cualquier fichero auxiliar de datos debe ser siempre "datos/nombre\_fichero". Es decir, se espera que el código lea de un directorio llamado "datos", situado dentro del directorio donde se desarrolla y se ejecuta la práctica.
- ▶ Un código es apto para ser corregido si se puede ejecutar de principio a fin sin errores.
- ▶ NO ES VÁLIDO usar opciones en las entradas. Para ello fijar al comienzo los parámetros por defecto que considere que son los óptimos.
- ▶ El código debe estar obligatoriamente comentado explicando lo que realizan los distintos apartados y/o bloques.
- ▶ Poner puntos de parada para mostrar imágenes o datos por consola.
- ▶ Todos los ficheros (\*.py, \*.pdf) se entregan juntos dentro de un único fichero zip, sin ningún directorio que los contenga.
- ▶ ENTREGAR SOLO EL CODIGO FUENTE, NUNCA LOS DATOS.
- ▶ Forma de entrega: Subir el zip al Tablón docente de CCIA.

FECHA DE ENTREGA

20 de MAYO

# DEFINICIÓN DEL PROBLEMA

Este ejercicio se centra en el ajuste de un modelo lineal a conjuntos de datos dados con el objetivo de obtener el mejor predictor posible. En todos los casos los pasos a desarrollar serán aquellos que nos conduzcan al ajuste y selección del mejor modelo y a la estimación del error  $E_{out}$  del modelo final.

# TAREAS

Cómo mínimo se habrán de analizar y comentar los siguientes pasos sobre un problema de clasificación y otro de regresión:

1. Comprender el problema a resolver: Conocer la base de datos con la que se está trabajando.
2. Preprocesado de datos: datos categóricos, normalización, proyección, etc.
3. Selección de clases de funciones a usar.
4. Definición de los conjuntos de training, validación y test usados en su caso.
5. Discutir la necesidad de regularización y en su caso la función usada para ello.
6. Definir los modelos a usar y estimar sus parámetros e hyperparámetros.
7. Selección y ajuste modelo final.
8. Discutir la idoneidad de la métrica usada en el ajuste
9. Estimación del error  $E_{out,ccvc}$  del modelo lo más ajustada posible.
10. Discutir y justificar la calidad del modelo encontrado y las razones por las que considera que dicho modelo es un buen ajuste que representa adecuadamente los datos muestrales.

# BASES DE DATOS

## Optical Recognition of Handwritten Digits Data Set

Acceso a los datos e información sobre ellos: [Aquí](#)

43 personas contribuyen a generar esta base de datos (30 training y 13 test) escribiendo dígitos numéricos.

- ▶ Tipo de atributos: enteros
- ▶ Número de instancias: 5620 (3823 training, 1797 test)
- ▶ Número de atributos: 64

# BASES DE DATOS

## Airfoil Self-Noise Data Set

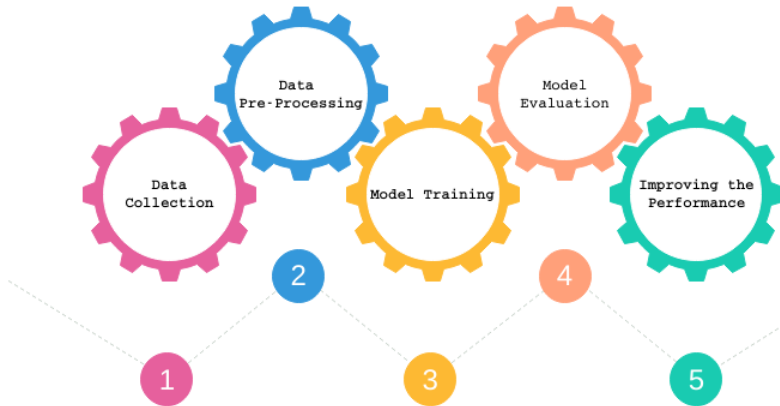
Acceso a los datos e información sobre ellos: [Aquí](#)

Datos obtenidos de una serie de pruebas aerodinámicas y acústicas de secciones de palas aerodinámicas realizadas en un túnel de viento.

- ▶ Tipo de atributos: reales
- ▶ Número de instancias: 1503
- ▶ Número de atributos: 6



# PASOS MACHINE LEARNING



<https://towardsdatascience.com/machine-learning-a-gentle-introduction-17e96d8143fc>

# DESARROLLO DE LOS PUNTOS

## 1. Comprensión del problema

- ▶ ¿Qué base de datos estamos viendo?
- ▶ ¿Qué representan las columnas? ¿Son numéricas o categóricas?
- ▶ ¿Qué hay en la variable de clase?
- ▶ ¿Se trata de un problema de aprendizaje supervisado o no supervisado?
- ▶ ¿Es un problema de regresión o de clasificación?

0.5 puntos

# DESARROLLO DE LOS PUNTOS

## 2. Preprocesamiento

### **¿Por qué se preprocesan los datos?**

Para eliminar impurezas y reducir la probabilidad de aprender de manera errónea de los datos. Causas:

- ▶ Datos incompletos (Valores perdidos)
- ▶ Datos con ruido
- ▶ Datos inconsistentes

1.5 puntos

# DESARROLLO DE LOS PUNTOS

## 2. Preprocesamiento

### Tareas:

- ▶ Colección, integración y transformación
  - Obtención de los datos, de una o más fuentes
  - Decodificación
  - Integración de datos de distintas bases de datos
  - Generación nuevo conocimiento
- ▶ Limpieza
  - Modificación de datos con conflicto
  - Eliminación de outliers
  - Tratamiento de valores perdidos y problemas de ruido
- ▶ Reducción
  - Selección de características
  - Selección de instancias
  - Discretización

# DESARROLLO DE LOS PUNTOS

## 3. Clases de funciones

Combinaciones lineales, cuadráticas, etc... de las observaciones.  
Justificar su uso o por qué no se consideran necesarias.

1 punto

# DESARROLLO DE LOS PUNTOS

## 4. Training, Test y Validación

**TRAINING** → Subconjunto de los datos que se estudia, se visualiza y a la que se le aplican los modelos.

**VALIDACIÓN** → Subconjunto de los datos que indica cuál es el mejor modelo.

**TEST** → Subconjunto de los datos que proporciona el error cometido.

### Posibles particiones:

- ▶ Si se decide usar el conjunto Validación: 50% training, 25% Validación y 25% test.
- ▶ Si no se decide usar el conjunto de Validación: 70% training y 30% test u 80% training y 20% test.

1 punto

# DESARROLLO DE LOS PUNTOS

## 5. Regularización

La regularización se trata del método que penaliza la complejidad del modelo, al usar función de coste. Produciendo modelos más simples que generalizan mejor.

- ▶ **L1 (Regularización Lasso)** → Interesante cuando se observa que algunas de las características no influyen demasiado en el modelo. Al dar coeficientes a cada atributo para generar la combinación de ellas, ciertos coeficientes tenderán a 0. Funciona mejor cuando los atributos no están correlados entre sí.
- ▶ **L2 (Regularización Ridge)** → Útil cuando parezca que varios de los atributos están correlados entre ellos. Hace que los coeficientes sean pequeños. Funciona mejor cuando la mayoría de los atributos son relevantes.

1.5 puntos

# DESARROLLO DE LOS PUNTOS

## 6. Modelos

Posibles modelos a usar:

- ▶ **Regresión lineal**
- ▶ **Regresión logística**
- ▶ **Perceptrón + Pocket**

1.5 puntos



# DESARROLLO DE LOS PUNTOS

## 7. Selección de modelo y ajuste

Seleccionar el que se considera el mejor de los modelos y argumentar por qué se elige.

Ajustar los datos de validación (o test) usando el modelo seleccionado.

1 punto

# DESARROLLO DE LOS PUNTOS

## 8. Métrica

Elegir la métrica a usar y discutir su elección. Teniendo en cuenta si se trata de un problema de regresión o de clasificación, así como el tipo de problema a tratar.

- ▶ **Regresión** Aquí y Aquí
- ▶ **Clasificación** Aquí

1 punto

# DESARROLLO DE LOS PUNTOS

## 9. Estimación del error

Especificar el error que se produce al ajustar el modelo.

1 punto

# DESARROLLO DE LOS PUNTOS

## 10.Justificación

Responder y argumentar:

- ▶ ¿Representa el modelo de manera adecuada los datos?
- ▶ ¿Consideras que la calidad del modelo es buena?

2.5 puntos