Práctica Final:
Resolución de un problema de Regresión Lineal mediante Algoritmos Genéticos

Computación

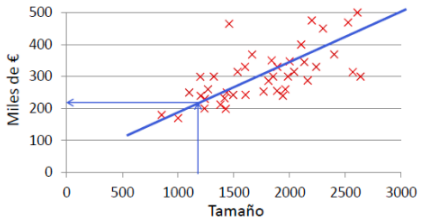
Segundo Semestre. Curso Académico 22/23

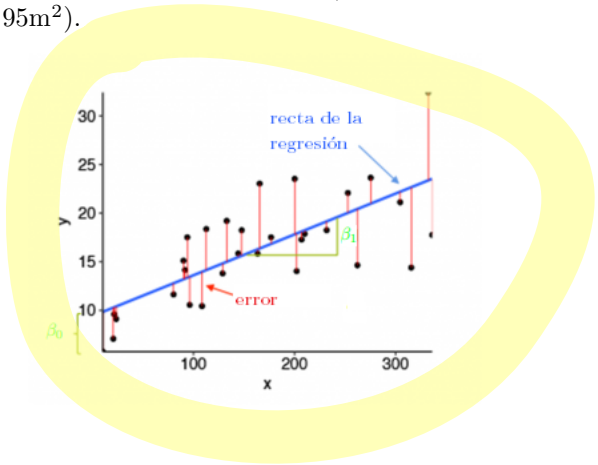
El objetivo de esta práctica final es alcanzar una buena comprensión del funcionamiento de los algoritmos genéticos y, al mismo tiempo, desarrollar la capacidad de analizar un problema desde el punto de vista de la Computación Científica. Más específicamente, se trata de analizar el efecto que la variación de los diferentes parámetros y/o técnicas dentro de un algoritmo genético tiene sobre la resolución de un problema concreto.

1 Regresión Lineal

Una regresión lineal es un modelo de aprendizaje supervisado que tiene como objetivo encontrar la relación lineal entre variables independientes y variable dependiente.

Pongamos un ejemplo simple: predecir el precio de una casa en base a su tamaño. **Como podemos observar en las siguientes figuras, el objetivo de una regresión lineal es obtener una recta (en el caso de una variable independiente) que explique la relación entre variables y minimice el error entre la recta y los valores de entrada.** De esta forma, obtenemos un modelo que podríamos usar para llevar a cabo futuras predicciones (por ejemplo, predecir el precio de una vivienda de 95m^2).





Representación gráfica de la regresión

En este caso, la variable independiente sería el tamaño y la variable dependiente o a predecir, el precio. Para estimar un modelo en el que la variable de salida se expresa linealmente en término de:

- Variables del modelo: x.
- Parámetros a predecir: β_0, β_1

$$\hat{y}(x) = \beta_0 + \beta_1 x$$

Un modelo en el que el precio se obtiene en términos del tamaño de la casa (tam_casa) y en base a los parámetros sería

$$precio(tam\_casa) = \beta_0 + \beta_1 \cdot tam\_casa$$

El que hemos visto, ha sido un ejemplo muy simple, pongamos que tenemos más variables independientes que pueden determinar el precio de una casa:

Tamaño (x_1)	Nº habitaciones (x_2)	Nº pisos (x_3)	Antigüedad (x_4)	Precio (y)
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
...	...	...	...	...

variables del modelo

En este caso tendremos que utilizar un modelo del tipo:

$$\hat{y}(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

ecuación del modelo

Donde $x_1, \ldots, x_4$ son las variables de la tabla anterior y $\beta_0, \ldots, \beta_4$ son los parámetros reales del modelo que hay que predecir.

2 Problema de regresión lineal a resolver

Junto con el enunciado de la práctica se provee un comprimido con nombre "datasets.zip". Dicho comprimido contiene una serie de archivos .dat, cada uno de los cuáles corresponde a un conjunto de datos distinto. El contenido de dichos archivos está dado en formato CSV (Comma Separated Values), y cada fila del documento representa los valores de las n variables del problema a tratar. En concreto, los $n-1$ valores iniciales de la fila corresponden a las variables predictoras (las que vamos a utilizar para entrenar nuestro modelo de regresión lineal), mientras que el n-ésimo valor corresponderá con el valor de la variable a predecir.

La carga de los datos de cada fichero se puede llevar a cabo de manera sencilla usando la librería *pandas*, mediante la función pandas.read_csv. Posteriormente, puede convertirse el objeto *DataFrame* devuelto por *pandas* a un array de *numpy* mediante la función pandas.DataFrame.to_numpy.

A continuación se describen brevemente los conjuntos de datos presentados (todos los datasets han sido extraídos del repositorio KEEL):

1. **Diabetes**: Este conjunto de datos se refiere al estudio de los factores que afectan a los patrones de diabetes mellitus insulinodependiente en niños. El objetivo es investigar la dependencia del nivel de péptido C en suero de los demás factores para comprender los patrones de secreción residual de insulina. La medida de respuesta es el logaritmo de la concentración de péptido C (pmol/ml) en el momento del diagnóstico, y las medidas predictoras la edad y el déficit de bases, una medida de la acidez.

Variable	**Rango**
Age	$[0.9, 15.6]$
Deficit	$[-29.0, -0.2]$
C_peptide	$[3.0, 6.6]$

2. **Quake**: Un conjunto de datos de regresión cuya tarea consiste en aproximar la fuerza de un terremoto dada la profundidad de su punto focal, su latitud y su longitud.

Variable	**Rango**
Focal_depth	$[0, 656]$
Latitude	$[-66.49, 78.15]$
Longitude	$[-179.96, 180]$
Richter	$[5.8, 6.9]$

3. **Stock prices**: Los datos proporcionados son los precios diarios de las acciones de diez empresas aeroespaciales desde enero de 1988 hasta octubre de 1991. La tarea consiste en aproximar el precio de la décima empresa a partir de los precios del resto.

Variable	**Rango**
Company1	$[17.219, 61.5]$
Company2	$[19.25, 60.25]$
Company3	$[12.75, 25.125]$
Company4	$[34.375, 60.125]$
Company5	$[27.75, 94.125]$
Company6	$[14.125, 35.25]$
Company7	$[58, 87.25]$
Company8	$[16.375, 29.25]$
Company9	$[31.5, 53]$
Company10	$[34, 62]$

NOTA: Todos las variables del problema se han reescalado al rango $[0, 1]$ para facilitar el desempeño del modelo de regresión. En concreto, se ha hecho uso de la normalización "min-max", dada por la fórmula:

$$\hat{x_i} = \frac{x_i - \min x_i}{\max x_i - \min x_i}$$

Normalización usada

Por lo tanto, se podrían recuperar los valores originales desnormalizando las variables (teniendo en cuenta los valores mínimos y máximos del rango original) mediante la siguiente función:

$$x_i = \hat{x_i}(\max x_i - \min x_i) + \min x_i$$

Como desnormalizar

Esto puede resultar beneficioso a la hora de generar visualizaciones en la escala original.

3 Indicaciones para la solución

3.1 Indicaciones generales

1. El número de atributos del conjunto de datos con el que estemos trabajando condicionará el tamaño de los cromosomas de nuestro problema.
2. En el documento a entregar debe describirse tanto la forma escogida para los cromosomas como la función o funciones de fitness utilizadas.
3. El objetivo de la práctica es estudiar cómo varía el comportamiento de los algoritmos genéticos empleados para resolver este problema cuando se varían los diferentes parámetros. En particular, debe analizarse:
   - Si se alcanza o no una solución.
   - El coste temporal y de iteraciones.
   - Cualquier otro aspecto que se considere relevante.

3.2 Variaciones

Más que el código en sí, que debe ser correcto, se valorará la calidad del informe de resultados, con especial énfasis en el número de pruebas realizados, la justificación de los cambios de parámetros llevados a cabo, el análisis del efecto de dichos cambios sobre el comportamiento del algoritmo genético y las conclusiones obtenidas a partir de las pruebas realizadas.

Entre los parámetros que se pueden tener en cuenta están:

1. El método de asignación de probabilidades para la selección de progenitores.
2. El método de selección de progenitores: al menos deben implementarse el método de la ruleta y el método del torneo con k contrincantes, siendo k uno de los parámetros a estudiar.
3. El método de cruzamiento. Deben implementarse al menos dos métodos de cruzamiento diferentes.

4. El método de selección de supervivientes.
5. La función de fitness
6. El efecto de la presencia o ausencia de mutaciones.

Además, en cada uno de estos puntos debe tenerse en cuenta la posible variación de los parámetros que los definen (probabilidades altas o bajas, etc)

4 Documentación a entregar

Se entregará a través de las tareas de MiAulario los siguientes dos archivos:

1. `Codigo_Apellido1Apellido2.zip`: Un comprimido .zip donde se encuentren todos los archivos de código de Python necesarios. Los archivos deberán estar comentados para que el código resulte comprensible.

   En el caso de que se programe en un Notebook (.ipynb), obligatoriamente, el programa principal deberá recibir el nombre de `main.ipynb` y deberá indicarse claramente y ejecutarse el programa completo en la última celda del Notebook.

   En el caso de que no se programe en un Notebook, obligatoriamente, uno de estos archivos recibirá el nombre de `main.py`. Al ejecutar dicho archivo, se deberá ejecutar el programa completo automáticamente hasta llegar a una solución.

2. `Informe_Apellido1Apellido2.pdf`: Un informe en PDF que incluya:

   (I) Descripción de cómo se ha planteado y resuelto el problema. En particular, **debe explicarse la codificación de los cromosomas y cuál es la función de coste** (o las funciones de coste, si se han hecho pruebas con diferentes funciones de coste) **y su justificación** en la utilización.

   (II) Experimentos:
   - Descripción de los experimentos llevados a cabo: diferentes valores de los parámetros utilizados y resultados de los mismos para cada problema.
   - Un estudio de cómo ha afectado al comportamiento del algoritmo las diferentes variaciones de los parámetros llevadas a cabo, es decir, **interpretación y conclusión de los resultados** de cada problema.

   (III) **Conclusiones** a las que se pueda llegar tras la realización de una gran variedad y combinaciones de parámetros.

   (IV) En general, sobre la presentación del trabajo:
   - Identificación: nombre, asignatura, título del trabajo, etc.

- Elementos del informe: índice, numeración de páginas, sección de introducción y objetivos, organización en secciones indicando el objetivo específico de cada sección.
- Presentación de resultados: tablas y gráficos con el mismo formato (incluyendo pie de tabla o gráfico), interpretables y claros (gráficos con etiquetas y leyendas si son necesarios).
- Redacción gramatical correcta y comprensible.

Ejemplo: si el estudiante tiene los apellidos Pérez García, deberá entregar los siguientes dos ficheros:

- `Codigo_PerezGarcia.zip`
- `Informe_PerezGarcia.pdf`

La práctica ponderará el 85% de la nota de las prácticas de la asignatura (50%). El 15% restante de las prácticas se obtendrá a través de la media de las prácticas entregadas.

La fecha límite para la entrega de la práctica será el día **31 de mayo de 2023** (día del examen ordinario).