VC4 – Descenso del Gradiente

03MIAR – Algoritmos de Optimización

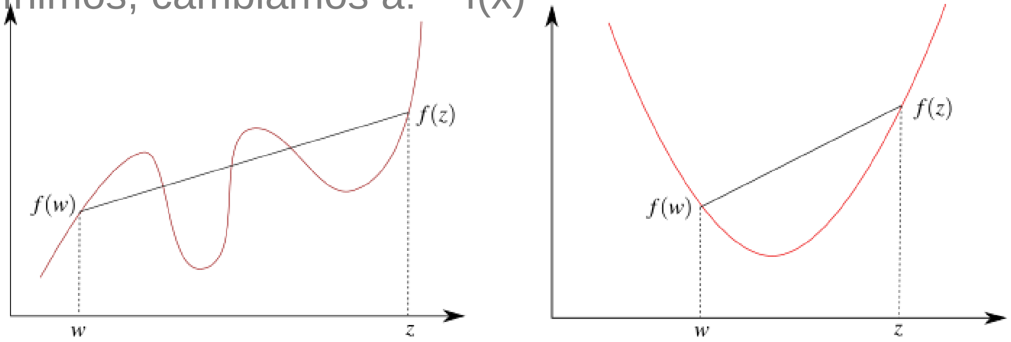
Universidad Internacional de Valencia

Descenso del Gradiente – AGD

importante

- Es un algoritmo **iterativo** de optimización para encontrar valores mínimos para funciones multivariables **convexas** y **diferenciables (y por tanto continuas).**

- Convexidad
  - Permitir asegurar que la óptimo que encontremos es global frente a que solo sea local
  - ¿El segmento que une dos puntos está siempre por encima de la función?
  - Si es cóncava y buscamos mínimos, cambiamos a: –f(x)



a) Función no convexa

b) Función convexa

Descenso del Gradiente – AGD

- **Diferenciable**: derivable en n variables(dimensiones)

- **Gradiente**: Es un vector(diferente para cada punto = campo vectorial) cuyas coordenadas son las derivadas parciales:

$$\nabla f(\mathbf{r}) = \left(\frac{\partial f(\mathbf{r})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{r})}{\partial x_n} \right)$$

Ejemplo: Dada la función **f(x,y) =** $x^2 + 2x + y^2 + y^3 + xy$
el gradiente será:

$$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right) \qquad (2x + y + 2, 2y + 3y^2 + x)$$

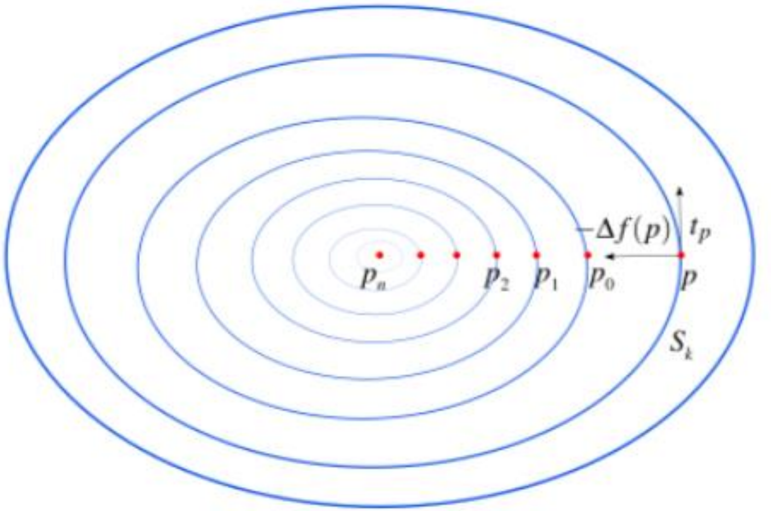
En el punto (x,y)

Universidad Internacional de Valencia

Descenso del Gradiente – AGD

El procedimiento parte de un punto **p como solución aproximada(inicial)** y da pasos en el sentido opuesto(si minimizamos) al gradiente de la función en dicho punto.

$$p_{t+1} = p_t - \alpha_t \cdot \Delta f(p_t)$$



La elección de α_t estará condicionada para que :

- p_{t+1} sea solución factible
- mejore el valor de f respecto a p_t : $f(p_{t+1}) <= f(p_t)$

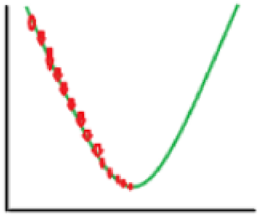
Descenso del Gradiente – AGD

La elección del parámetro **α_t** , llamado **tasa de aprendizaje (learning rate),** es importante para hacer efectivo el proceso de acercamiento a la solución óptima.

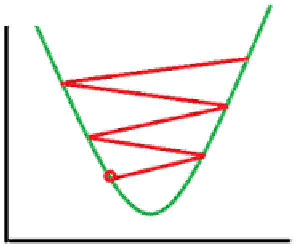
- Un valor demasiado pequeño puede provocar exceso de iteraciones(Figura 1)

- Un valor demasiado alto puede provocar que el proceso no se acerque los suficiente(Figura 2)

$$p_{t+1} = p_t - \alpha_t \cdot \Delta f(p_t)$$



learning rate demasiado pequeño
(Figura 1)



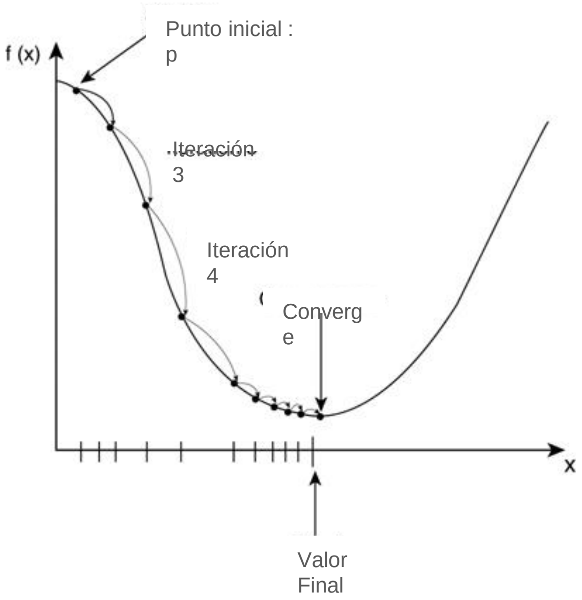
learning rate demasido grande
(Figura 2)

Descenso del Gradiente – AGD

En general es buena estrategia ir reduciendo el valor de α_t dinámicamente a medida que nos aproximamos a la solución

- ¿Como sabemos que nos acercamos?:
  - la magnitud del gradiente
  - cantidad de iteraciones que hemos realizado

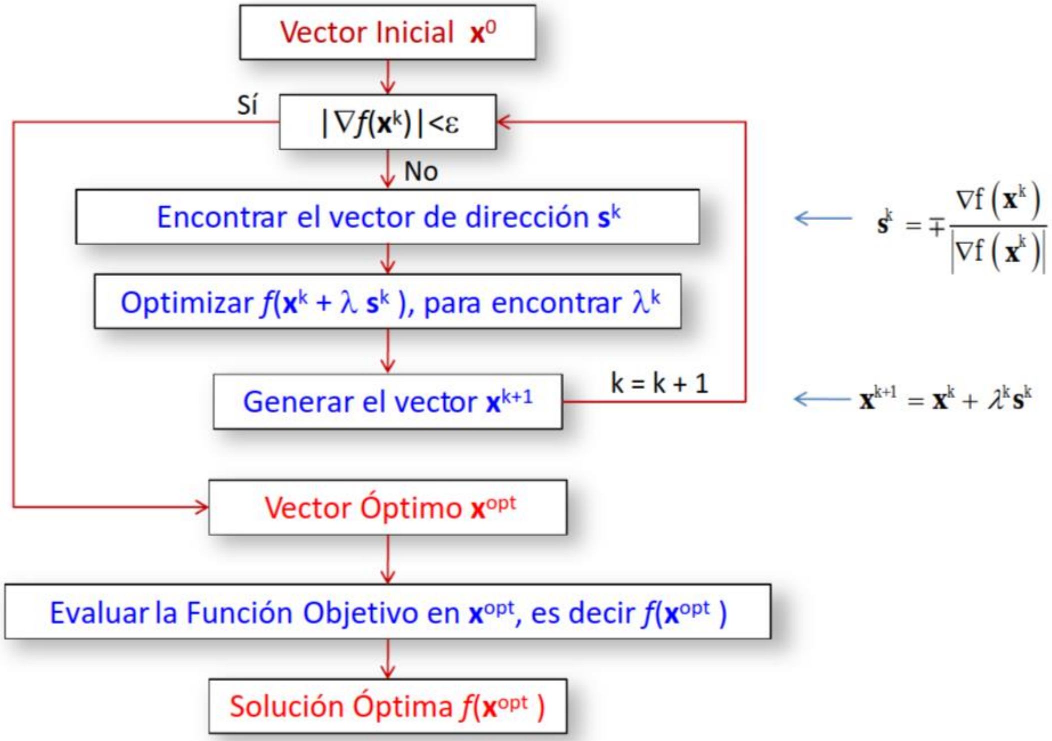
$$p_{t+1} = p_t - \alpha_t \cdot \Delta f(p_t)$$



http://www.bdhammel.com/learning-rates/

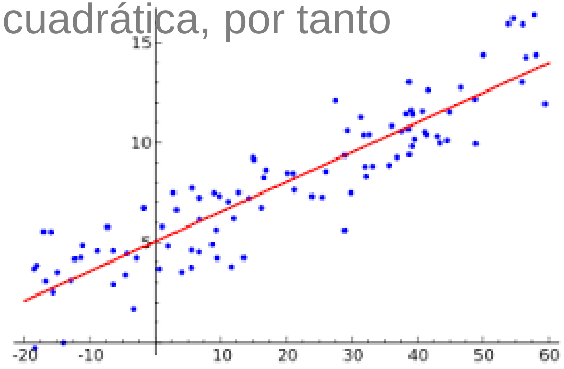
Descenso del Gradiente – AGD

- Diagrama de resolución



Descenso del Gradiente – AGD

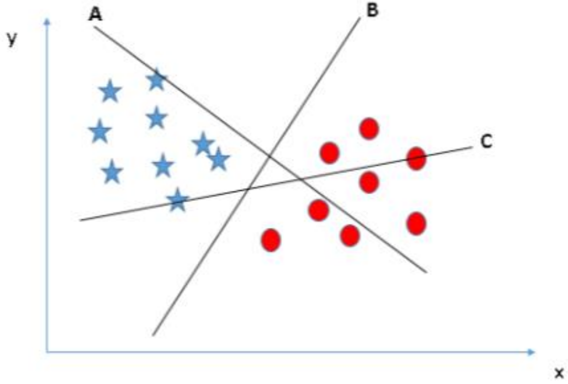
- **Aprendizaje supervisado y la regresión lineal**
  - ✓ En problemas de clasificación tratamos de clasificar los puntos en sus categorías correspondientes de tal manera que se minimice el error.
  - ✓ Para la medida de este error solemos utilizar el **error cuadrático medio**(regresión lineal) pero es posible usar otras medidas(por ejemplo: distancia Euclidea, Manhattan, ...)
  - ✓ La función que acumula los errores para los valores conocidos la llamamos función de coste. En el caso de la regresión líneal es cuadrática, por tanto diferenciable y podemos obtener el gradiente.

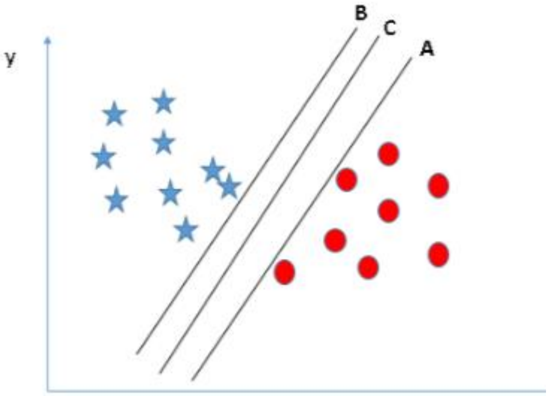


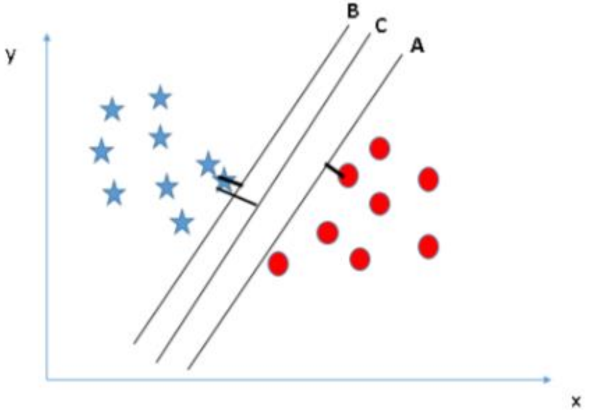
Descenso del Gradiente – AGD

- **Aprendizaje supervisado y clasificación**

¿Cómo identificar la mejor recta (hiperplano en N-dimensión) que separa los conjuntos de datos?





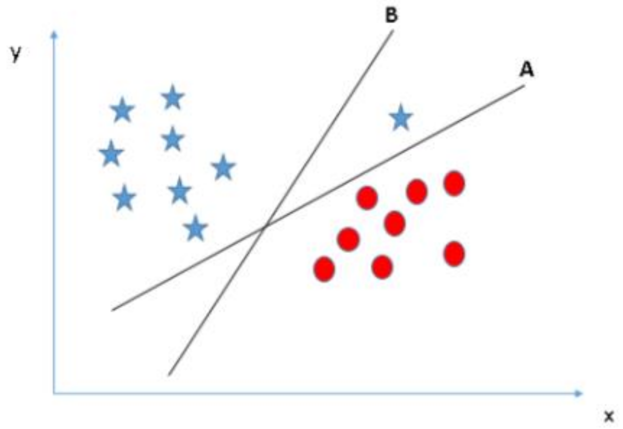


Seleccionar el que esté
a mayor distancia de ambos conjuntos

Descenso del Gradiente – AGD

- **Aprendizaje supervisado y la clasificación**

¿Cómo identificar la mejor recta (hiperplano en N-dimensión) que separa los conjuntos de datos?



1º . Seleccionar el que mejor clasifique

2º . Seleccionar el que está a más distancia

Descenso del Gradiente – AGD

- **Aprendizaje supervisado y la regresión lineal**

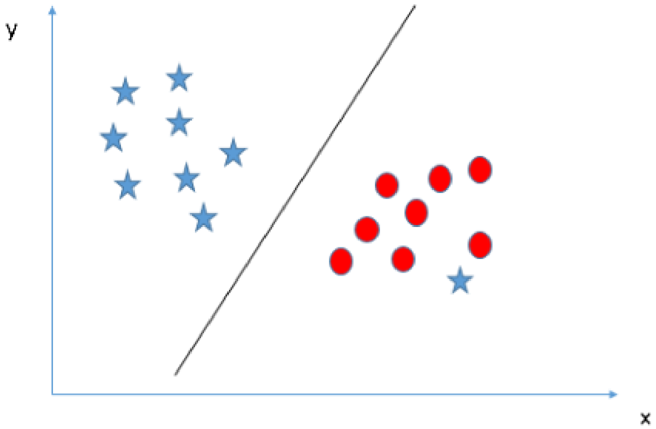
¿Cómo identificar la mejor recta (hiperplano en N-dimensión) que separa los conjuntos de datos?

- *No siempre es posible clasificar 100%*
- *¿valores atípicos?*

Descenso del Gradiente – AGD

- **Aprendizaje supervisado y la clasificación**

¿Cómo identificar la mejor recta (hiperplano en N-dimensión) que separa los conjuntos de datos?

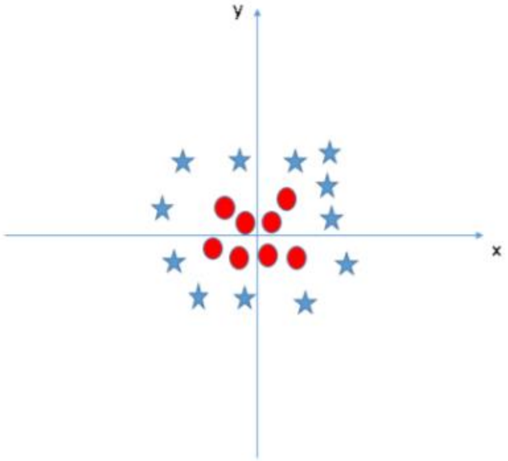


- *Ignorar valores atípicos*

Descenso del Gradiente – AGD

- **Aprendizaje supervisado y la clasificación**

¿Cómo identificar la mejor recta (hiperplano en N-dimensión) que separa los conjuntos de datos?

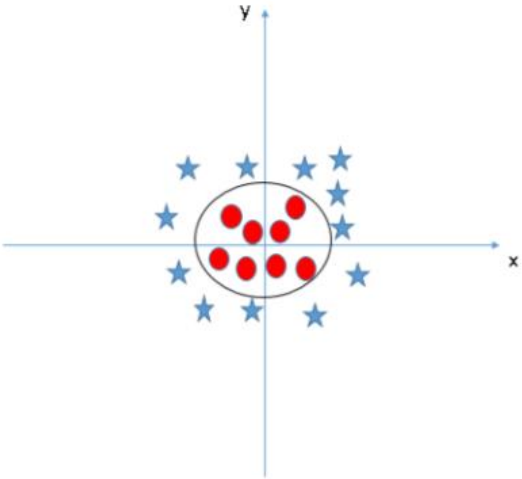


- *No parecen valores atípicos*
- *Parece que si hay una clasificación*
- *Es imposible con hiperplanos lineales*

Descenso del Gradiente – AGD

- **Aprendizaje supervisado y la clasificación**

Podemos extender el procedimiento a funciones no lineales!



Una función como $f(x,y) = x^2 + y^2$ *lo resuelve*

Descenso del Gradiente – AGD

- **Otra opción para función de coste(no cuadrática)**

- **Entropía cruzada(cross-entropy)**
    - Para problemas con soluciones binarias
    - **Definición**: Número de bits diferentes. Mide como de diferentes son dos elementos.
    - Usada en algoritmos de clasificación de imágenes

Descenso del Gradiente – AGD

- **Dependiendo del Volumen de Datos**

✓ Descenso del gradiente por lotes (**batch gradient descent**)

Calcula la desviación para todos los puntos en cada iteración!!!

✓ Descenso del gradiente estocástico(**stochastic gradient descent**)

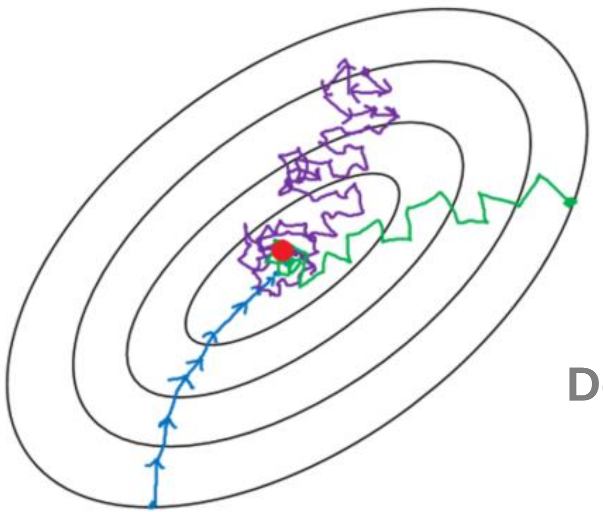
Calcula la desviación para un punto en cada iteración!!!

✓ Descenso del gradiente por lotes reducido(**mini-batch gradient descent**)

Mezcla de ambos conceptos

Descenso del Gradiente – AGD

- Dependiendo del Volumen de Datos



Batch gradient descent (directo pero muy lento)

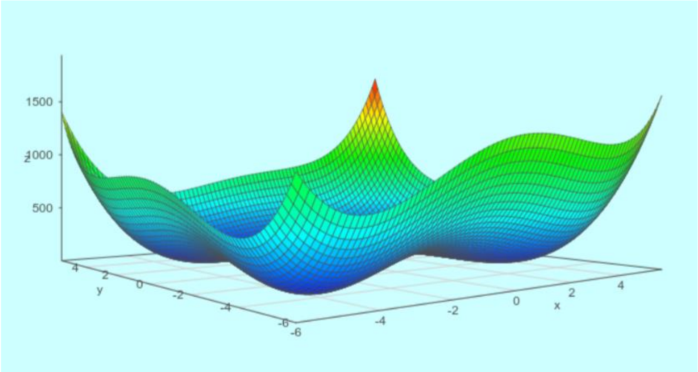
Stochastic gradient descent(rápido pero muy disperso)

Mini-batch gradient descent(equilibrio)

Decisión: Elección de mini-batch size

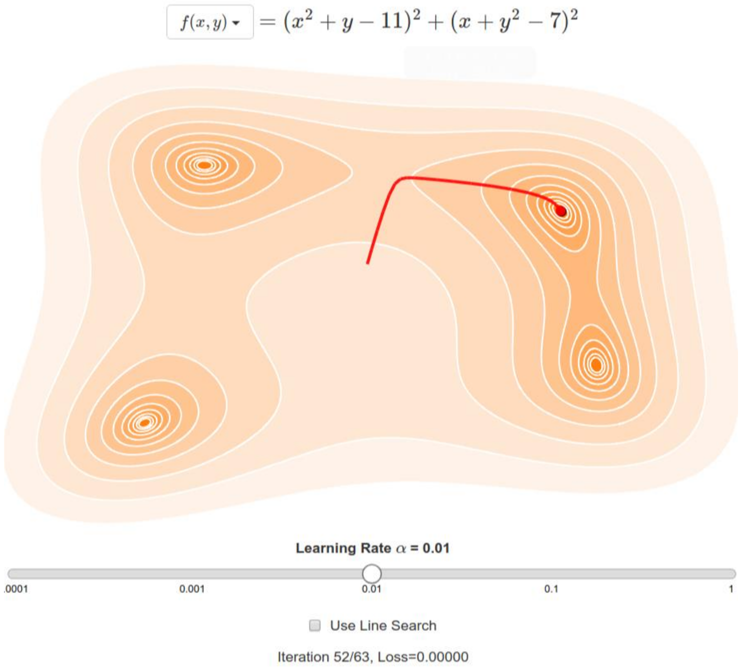
Descenso del Gradiente – AGD

Visualización



(x^2+y-11)^2+(x+y^2-7)^2

http://al-roomi.org/3DPlot/index.html

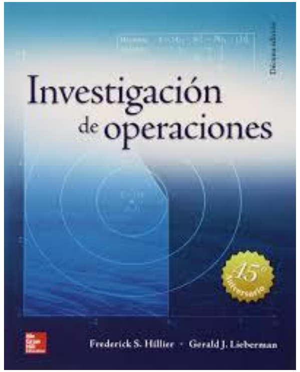


https://www.benfrederickson.com/numerical-optimization/

Ampliación de conocimientos

- ¿Qué es el Descenso del Gradiente? Algoritmo de Inteligencia Artificial | DotCSV
https://www.youtube.com/watch?v=A6FiCDoz8_4

- Hillier, F. S., y Lieberman, G. J. Investigación de Operaciones. Ciudad de México



Ampliación de conocimientos y habilidades

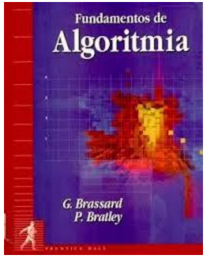
- Bibliografía

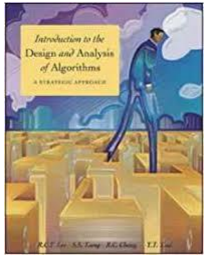
-Brassard, G., y Bratley, P. (1997). Fundamentos de algoritmia. *ISBN 13: 9788489660007*

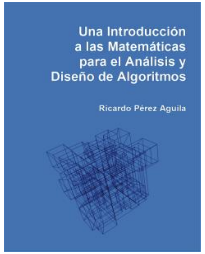
-Guerequeta, R., y Vallecillo, A. (2000). Técnicas de diseño de algoritmos. http://www.lcc.uma.es/~av/Libro/indice.html

-Lee, R. C. T., Tseng, S. S., Chang, R. C., y Tsai, Y. T. (2005). Introducción al diseño y análisis de algoritmos. *ISBN 13: 9789701061244*

- *Pérez Aguila, R. (2012).* Una introducción a las matemáticas para el análisis y diseño de algoritmos. *ISBN 13: 9781413576474. https://tinyurl.com/yzlt5oed*







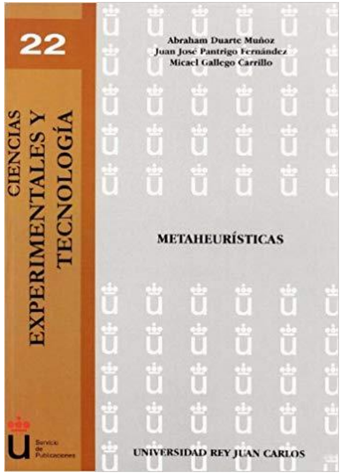
Universidad Internacional de Valencia

Ampliación de conocimientos y habilidades

- Bibliografía

- Duarte, A. (2008). Metaheurísticas. Madrid: Dykinson.

https://elibro-net.universidadviu.idm.oclc.org/es/ereader/universidadviu/35696



¿Preguntas?

Gracias

juanfrancisco.vallalta@campusviu.es

Universidad Internacional de Valencia