# Actividad de evaluación a distancia para asignatura: "Ampliación de Sistemas Inteligentes"

curso 2017-18

## 1. Información previa

Es importante que en primer lugar **lea todo este documento completo detenidamente** y hasta el final (se incluye una sección de aclaraciones específicas), ya que proporciona paso a paso toda la información necesaria para realizar la actividad. También es necesario haber estudiado ya el tema 2 (Computación Evolutiva) de la asignatura (ver, en el manual didáctico, el material correspondiente).

### 1.1. Resumen

Se propone realizar los cálculos correspondientes a **un solo paso de una generación** en un problema de optimización de una función dada, resuelto mediante un algoritmo genético sencillo pero para **dos variantes**. Se proporcionan (en fichero aparte) los individuos que forman la población en el instante actual y se piden cuáles serán los individuos que formarán la siguiente generación según cada variante.

### 1.2. Forma de entrega

Esta es una tarea individual, no se permiten grupos. Los datos a utilizar en la actividad **son diferentes para cada estudiante**. El resultado de la actividad debe redactarse en un documento, cuyo contenido y estructura se especifica detalladamente más adelante (sección 3.2), convertido a formato **PDF**. El documento se enviará únicamente a través de la tarea correspondiente **dentro del curso virtual** (en la página de "Actividades Evaluables" del menú izquierdo): `https://2018.cursosvirtuales.uned.es/dotlrn/grados/asignaturas/71014069-18/?page_num=2`

Por favor, **no** envíen la respuesta por ningún otro medio, ni por correo electrónico, para evitar confusiones, extravíos y problemas de gestión. Tampoco se aceptarán entregas fuera del plazo establecido para todos los estudiantes.

En caso de que desee añadir algún otro fichero muy relevante en su respuesta (código fuente, hoja de cálculo, etc.), debe empaquetar todos los ficheros juntos, incluido el documento PDF obligatorio, en un archivo comprimido compatible ZIP o RAR y subir ese archivo comprimido (en lugar del documento) a la tarea correspondiente dentro del curso virtual.

### 1.3. Plazo de entrega

La entrega se debe realizar subiendo el fichero correspondiente en la tarea del curso virtual **antes del día 15 de diciembre de 2017 a las 23:55h.** Se recomienda no esperar hasta el último día, para evitar las saturaciones o problemas puntuales de acceso a internet, etc. Si surge algún problema para el envío, realice lo antes posible una consulta directamente al equipo docente por email (pero no envíe el fichero por ese medio) a <jras@dia.uned.es>.

La respuesta de la actividad se puede enviar en el curso virtual tantas veces como se quiera hasta el plazo indicado, ya que solamente la **última versión** se guarda y es la única que se tendrá en cuenta para la evaluación.

## 2. Fichero de datos personalizados

Para la realización de la práctica se requiere descargar un fichero de hoja de cálculo que permite generar los datos iniciales y los parámetros necesarios de forma específica para cada estudiante. No serán válidas las actividades realizadas con otros datos diferentes de los que corresponden a cada estudiante en este curso.

### 2.1. Obtención y utilización

El fichero de datos tiene el nombre `generador-datos-actividad-AmSI-2017-18.xls` y se descarga de la carpeta "Actividades a Distancia" en el repositorio "Documentos y material" del curso virtual de la asignatura, o en la URL: `https://2018.cursosvirtuales.uned.es/dotlrn/grados/asignaturas/71014069-18/file-storage/view/actividades-a-distancia/generador-datos-actividad-AmSI-2017-18.xls`

El fichero de datos es una hoja de cálculo simple compatible con MS-Excel. Se puede abrir, entre otros muchos programas gratuitos o libres, con: LibreOffice-Calc (=OpenOffice-Calc, ver `http://libreoffice.org/` para obtenerlo), Gnumeric (de escritorio Gnome), Calligra Sheets (=KSpread de escritorio KDE), Google Spreadsheet (de Google Docs online en `https://docs.google.com/spreadsheet/`). Es necesario un programa que permita introducir datos y recalcular, por lo que no sirven programas solo visualizadores (como p.ej. MS.Excel Viewer o similares).

La hoja de cálculo tiene bloqueadas casi todas las celdas para evitar la modificación accidental inadvertida de su contenido al utilizarla. Aún así, es posible que en algunas implementaciones (como por ejemplo, Google Spreadsheet) no se preserve este bloqueo, por lo que se recomienda tener la precaución de no modificar las fórmulas originales de las celdas para evitar errores en los resultados. Sí es posible marcar y copiar los valores mostrados en las celdas para, por ejemplo, "pegar" su contenido en un editor u otra hoja de cálculo a través del portapapeles. Si se quieren copiar los datos para reutilizarlos en otra hoja de cálculo, se debe tener cuidado de utilizar la opción de "pegado especial" para pegar solamente los valores numéricos (no las fórmulas), de forma que no se modifiquen respecto a los valores generados con el identificador de cada estudiante.

**Atención**: si se copia la fórmula de la función de adecuación para usarla en otra hoja de cálculo o en un programa, hay que copiar también el valor (oculto) de la celda `A3`, que se usa en esa fórmula.

### 2.2. Forma de uso del fichero de datos

Una vez abierto el fichero de datos, cada estudiante debe introducir el número completo (sin letras, ni guiones ni otros caracteres) de su DNI, o el de su pasaporte si ha usado ese documento en la matrícula, **dentro de la primera celda** (`A1`, superior-izquierda) en la hoja de cálculo, y después activar una operación de **recalcular** (habitualmente tecla F9 en la mayoría de los programas) en caso de que la actualización no sea automática. En caso de que el bloqueo de celdas no funcione, se debe tener la precaución de no modificar otras celdas de la hoja para evitar errores en los datos.

Una vez introducido el número del documento de identificación del estudiante y recalculada la hoja, se mostrarán los datos que cada estudiante debe utilizar para realizar su actividad. Estos datos consisten en:

- Una tabla (debajo y a la derecha del número de identificación) donde cada fila (de la `5` a la `13` de la hoja de cálculo) representa a un individuo de la población inicial ($m = 9$). La primera columna (`B` de la hoja de cálculo) es el número identificador del individuo y las 7 columnas siguientes (de la `C` a la `I` de la hoja de cálculo) son los valores de sus genes (números reales `x1` a `x7`). La última columna (`J` de la hoja de cálculo) es el valor de la función de evaluación para cada individuo.

- Cinco columnas de números aleatorios que se deben utilizar en orden cuando sea necesario para los cálculos requeridos. La primera columna (`L` de la hoja de cálculo) es entre 0 y 1 (con 6 decimales), las otras cuatro columnas (de la `M` a la `P` de la hoja de cálculo) son de enteros (ambos extremos inclusive): entre 1 y 9; entre 1 y 8; entre 1 y 7; y entre 1 y 6, respectivamente. Las cinco distribuciones son uniformes en su intervalo correspondiente. De cada lista hay 20 números aleatorios, en caso de que de alguna de ellas fuese necesario utilizar más números aleatorios se debe volver a comenzar por el principio en orden.

- Debajo de la tabla de la población actual, se proporciona un mecanismo para calcular la misma función de adecuación de hasta 4 individuos cualesquiera simultáneamente. Al introducir los valores de los genes correspondientes en cada celda marcada (que no están bloqueadas) se obtendrá en la columna de la función de adecuación el valor correspondiente.

Es posible que en algunas implementaciones aparezcan visibles otras celdas (en la 1ª columna debajo del DNI) que deberían estar ocultas, y con números enteros usados internamente para la generación de números aleatorios. Esas celdas se deben *ignorar* en caso de que sean visibles y no se debe modificar su valor ni su fórmula.

## 3. Resultados requeridos

Suponemos que se quiere resolver mediante algoritmos genéticos un problema de optimización para hallar el punto multidimensional en el que una función de varias variables $f(x_1, x_2, x_3, x_4, x_5, x_6, x_7) \in \mathbb{R}$; $x_j \in \mathbb{R}$ alcanza su máximo. La propia función a optimizar se usará como función de adecuación. La función a optimizar está dada mediante parámetros internos en el fichero de datos personalizados (ver apartado 2), que también proporciona los valores de dicha función para los individuos de la población inicial y la posibilidad de calcularla para otros individuos. El rango de valores de la función de adecuación está entre 0 y 6,5.

Dado que los rangos de las variables $x_j$ son todos los números reales, se ha elegido una representación en coma flotante[1]. La solución buscada y los individuos del algoritmo genético serán vectores de 7 números reales representados en coma flotante. Cada número componente del vector será un gen dentro del cromosoma correspondiente a cada individuo. Se utilizará una población de 9 individuos ($m = 9$).

[1] Se considera suficiente precisión la proporcionada por cualquier calculadora de mano, programa de hoja de cálculo o cualquier implementación de coma flotante en lenguajes de programación habituales.

## 3.1. Variantes a calcular

Se deben realizar los cálculos correspondientes para hallar solo **la siguiente generación de individuos** para cada una de las **dos variantes** de algoritmo genético siguientes:

1. Selección/muestreo: *Proporcional por muestreo estocástico con reemplazamiento (ruleta con 1 puntero)*[2].
   Cruce: con probabilidad 0,4 y, en caso de que se produzca, realizar *cruce simple*.
   Mutación[3]: con probabilidad 0,2 y, en caso de que se produzca, realizar *mutación por intercambio mínimo*.
   Sustitución: *generacional completa con estrategia elitista*.

2. Selección/muestreo: *Selección por torneo binario*.
   Cruce: con probabilidad 0,45 y, en caso de que se produzca, realizar *cruce aritmético*.
   Mutación[3]: con probabilidad 0,25 y, en caso de que se produzca, realizar *mutación por intercambio mínimo*.
   Sustitución: *Sustitución de estado estable (Steady-state)* de $n = 4$ individuos.

Para la interpretación correcta del procedimiento para calcular las dos variantes se deben tener en cuenta las siguientes instrucciones específicas para esta actividad:

- Cada variante se aplica sobre los mismos datos iniciales (individuos de la población inicial).
- Al realizar los cálculos correspondientes, se deben utilizar los números aleatorios proporcionados en la hoja de cálculo de datos personalizados (ver apartado 2). Se deben utilizar en el orden dado y en caso de que fueran necesarios más números, se reutilizarán en el mismo orden desde el principio de la lista correspondiente.
- Los cálculos de cada variante se hacen independientes y utilizando los números aleatorios desde el principio de cada lista al comenzar cada una de las dos variantes (para que la 2 no dependa de los usados en la 1).
- No es obligatorio consumir todos los números aleatorios dados, ni usar todas las columnas, solamente los que el algoritmo requiera. Si se consideran necesarios números en otros rangos distintos de los proporcionados, se deberían construir a partir de los dados mediante una función de conversión adecuada (y explicarla).
- Las tasas de cruce o mutación se dan en forma de probabilidad[4] en tanto por uno (de 0 a 1). La forma correcta de simular una probabilidad es obtener un número aleatorio entre 0 y 1, y si éste es **menor** que la probabilidad a simular, entonces se realiza la acción.
- Si por azar, al realizar alguna operación (selección, mutación, cruce), coinciden o resultan individuos indistinguibles (el mismo o con iguales genes, etc.), se debe utilizar ese caso tal cual, pero explicando porqué el resultado es el mismo individuo y cómo se habría aplicado la operación si hubiesen sido distintos genes. De la misma forma, si por azar no se produce ningún cruce o mutación, se debe incluir una explicación de cómo sería la operación correspondiente que se habría realizado en el caso de sí hubiera salido hacerla.
- En los apartados de ejercicios resueltos del libro base hay ejemplos e información adicional necesaria que también se debe estudiar en cada capítulo. En concreto, el apartado 11.7 contiene información sobre las variantes de los algoritmos en computación evolutiva que pueden ser necesarias para esta actividad.
- La selección por torneo binario requiere generar aleatoriamente grupos de individuos. Para simplificar, se realizará con números aleatorios independientes (con reemplazamiento[5]), que además es lo normal, incluso aunque esto pudiera dar lugar a que un mismo individuo esté repetido dentro de un mismo grupo.
- En ambas variantes los métodos de selección/muestreo deben generar solo la cantidad de individuos necesaria para el método de sustitución correspondiente de cada variante.
- Hay que recordar que las representaciones de los genes más habituales pueden ser binarias (cada gen es un bit) o bien reales (cada gen es un número real en coma flotante). En esta actividad se usan genes en representación real, y por tanto, las operaciones de cruce y mutación se deben realizar sobre genes completos, es decir sobre el número real completo (no sobre los bits individuales de la representación en coma flotante interna de la máquina).
- Dado que la población inicial de los datos de esta actividad está dada en orden aleatorio, es posible realizar los **emparejamientos para el cruce** de forma secuencial (tomándolos de dos en dos en el orden en el que se obtienen[6]) para simplificar. Aunque, naturalmente, también es aceptable (pero no da más puntos) generar emparejamientos aleatorios (con o sin reemplazamiento) para el cruce en ambas variantes (explicándolo).

[2] Para que las parejas ya salgan en orden aleatorio (cosa que no ocurre con ruleta de múltiples punteros).

[3] La probabilidad de mutación suele ser mucho más pequeña, pero para que ocurra algún caso en esta actividad, se han elegido valores mayores a propósito.

[4] Esta forma es equivalente a las tasas en forma de tanto por ciento usadas en otros textos, teniendo en cuenta que la probabilidad 1 equivale a tasa del 100 % de que la mutación o el cruce ocurra y, por ejemplo, probabilidad 0,5 equivale a tasa del 50 %, etc.

[5] En procesos estocásticos se entiende "con reemplazamiento" cuando el número extraído (por ejemplo de un bombo) se vuelve a insertar (se reemplaza) y puede volver a salir en otra extracción. Es decir, cada extracción es independiente de resultados anteriores.

[6] Siempre que el muestreo estocástico universal se haga con ruleta de 1 puntero como se pide, y no con múltiples punteros.

- La probabilidad de cruce indicada en cada variante es aplicable a cada pareja. Cada torneo binario selecciona un individuo descendiente, habrá que emparejar los seleccionados de dos torneos sucesivos (que ya son aleatorios) para ver si se cruzan, incluso aunque ambos individuos sean copias del mismo (aunque, evidentemente, en este caso el resultado del cruce sería el mismo individuo).
- En el ejemplo 11.2.4 del libro base, se aplica la probabilidad de mutación a cada gen de cada individuo cuando se trata de mutación simple. En cambio, la mutación por intercambio mínimo (equivalente al intercambio recíproco), que se detalla en el algoritmo 11.6 del apartado 11.7 del libro base, ya incluye un mecanismo de selección aleatoria de los genes a mutar, por lo cual la probabilidad de mutación indicada en esta actividad se debe aplicar solamente a cada individuo seleccionado para decidir si hay mutación o no (no a cada gen, ni a cada posible pareja de genes).

### 3.2. Contenidos y estructura del documento de respuesta

El documento de respuesta puede redactarse en cualquier editor, pero se debe **exportar a PDF** para enviarlo. Por ejemplo, se puede usar el editor de LibreOffice que permite exportar a PDF, es gratuito y está disponible para todas las plataformas (ver `http://libreoffice.org/` para obtenerlo). También es muy recomendable el programa LyX (ver `http://lyx.org/`) que puede generar PDF de calidad fácilmente y está disponible en la mayoría de distribuciones de Linux, y también para MS-Windows o Mac.

En el documento de respuesta (indicado en el apartado 1.2) se deben especificar los resultados, y las explicaciones de cómo se han obtenido éstos, ordenados según el siguiente **esquema**:

1. Datos del estudiante: Nombre, Apellidos, DNI o pasaporte y email de contacto. También se puede incluir opcionalmente un número de teléfono de contacto.
2. Información sobre el entorno y programas usados para esta actividad (incluyendo números de versiones): Sistema operativo, entorno de escritorio, programa para hoja de cálculo y editor de texto para generar el documento PDF. Adicionalmente se puede incluir información de otros programas o medios utilizados para la realización.
3. Copia de los datos de individuos iniciales (solo es necesario incluir las componentes de cada vector).
4. Respuestas de cada una de las dos variantes del apartado 3.1 claramente separadas y etiquetadas. Para cada variante se debe incluir lo siguiente:
   *a*) Qué valores se han usado (qué lista de números aleatorios y su valor, etc) para elegir los padres y cuáles son los padres seleccionados para formar la siguiente generación en cada caso.
   *b*) En qué casos se ha realizado el cruce, porqué y cuáles son los descendientes.
   *c*) En qué casos se ha realizado la mutación, porqué y cuál es el resultado.
   *d*) Cómo se ha realizado la sustitución y cuáles serían los individuos que finalmente formarían la siguiente generación.
5. Comentarios y opiniones: Dificultades o problemas encontrados, programas o ayudas utilizadas, comentarios sobre la realización de la actividad, etc.
6. Bibliografía o fuentes de información: Documentos, páginas web, libros, etc. consultados para la realización (con datos de título y editorial o URL de localización, etc.).

Es muy importante para la evaluación seguir el esquema dado e **incluir explicaciones** de cómo se han realizado los cálculos (qué lista de datos, operación, resultado, etc.) para que se pueda calificar correctamente.

## 4. Aclaraciones sobre Algoritmos Genéticos para esta actividad

### 4.1. Cantidad de individuos a generar en la selección

La cantidad de individuos en cada generación se suele mantener constante, por lo que al diseñar un algoritmo genético se debe elegir el método de selección de los individuos a copiar (reproducción) y el método de sustitución de la generación anterior por los descendientes. Es muy habitual, como en los ejemplos del libro base, que se use sustitución completa, por lo cual en ese caso el mecanismo de copia debe producir tantos individuos nuevos como los que hay en la población. Pero en cambio, todos los métodos en los que no se reemplazan todos los individuos de la generación anterior, **solo deben fabricar los individuos nuevos necesarios**. Este es el caso de las dos variantes de la actividad (se seleccionan menos individuos de los que forman la población inicial).

El método de selección para copia ya debe incorporar la aleatoriedad suficiente según un criterio determinado. No tiene sentido, y es poco eficiente, seleccionar aleatoriamente individuos para copiar en función de un criterio, y después tener que descartar los innecesarios por otro criterio. Por lo tanto al diseñar el algoritmo genético, cuando se decide el método de sustitución, la cantidad de individuos necesarios condiciona cuántos individuos se producen para copia en el algoritmo de selección.

## 4.2. Selección proporcional

Entre los mecanismos de selección proporcional hay dos muy parecidos:

- Selección proporcional por *muestreo estocástico con reemplazamiento* o rueda de ruleta (con un puntero), que se aplica tantas veces como individuos se necesiten copiar. Éste método produce copias en un orden aleatorio directamente.
- Selección proporcional por *muestreo estocástico universal* o ruleta de varios punteros equidistantes (tantos como individuos se necesiten copiar), que solo se aplica una vez y produce de un golpe todos los individuos necesarios. Es más eficiente (solo hace falta un número aleatorio) pero a cambio produce las copias en orden correlativo de ranking.

En la actividad evaluable a distancia se pide específicamente usar el primero para evitar tener que realizar un paso extra de emparejamiento aleatorio para cruce. Así se pueden emparejar simplemente tomando parejas consecutivas en el orden en el que se obtienen las copias. En cambio, en el segundo habría que forzar una aleatoriedad adicional en el emparejamiento para que hubiera más diversidad respecto a la función de adecuación en los cruces.

## 4.3. Torneo entre grupos de individuos iguales

Los grupos para la selección por torneo se podrían realizar aleatoriamente sin repetir individuo, es decir sin reemplazamiento (que se puede simular descartando los números ya seleccionados si se repiten), pero aún así podría ocurrir que resultaran en el grupo algunos individuos con los mismos valores de sus genes (p.ej. porque esos genes son más abundantes en la población). Por lo tanto, y para simplificar la actividad no es necesario hacer los grupos sin reemplazamiento.

En el caso de que hubiera más de un individuo con la misma adecuación máxima dentro de un grupo, se puede elegir el ganador de ese torneo simplemente tomando el primero en orden, puesto que su orden ya es aleatorio por la manera de formar los grupos.

## 4.4. Aplicación de probabilidad de cruce

Una vez que se han copiado los individuos seleccionados y que se han realizado los emparejamientos, entonces se comprueba si en cada pareja se debe realizar el cruce o no, simulando la probabilidad decidida en el diseño del algoritmo. No debe afectar a la probabilidad el hecho de que los valores de los genes de ambos individuos coincidan (pueden provenir de dos selecciones del mismo individuo por casualidad, o bien porque los dos individuos originales ya tenían valores iguales), eso forma parte del algoritmo. En esta actividad solo se pide que en caso de que no toque cruzar ninguno por la probabilidad o que los únicos cruzados tengan los mismos genes, se indique esta situación y se explique cómo se hubiera hecho el cruce.

Evidentemente, los descendientes de las parejas a las que no les ha tocado cruzarse son copias directas.

## 4.5. Mutación de genes iguales

Si una vez simulada la probabilidad de mutación (solo por cada individuo en esta actividad al ser por intercambio de genes) y elegidas correctamente dos posiciones distintas de genes a intercambiar, resulta que los valores de esos genes son iguales, entonces se considera que la mutación ya se ha realizado y no es necesario repetirla con otros genes. En la actividad, en caso de que no toque mutar ninguno por probabilidad o que los únicos mutados fueran en genes con valores iguales, se debe indicar la situación y explicar cómo se hubiera hecho la mutación.

## 4.6. Mejora de adecuación media

Al ser este un ejercicio puramente académico muy simplificado, es posible que la adecuación media de la población resultante después de una sola generación no mejore. Por tanto, si de los resultados obtenidos se observa que la adecuación no mejora, esto no implica que el ejercicio esté mal (salvo que se haya cometido algún error). De igual forma, evidentemente, el mero hecho de que la adecuación media sea mejor, no implica que el ejercicio se haya realizado correctamente.