Tarea 1

Tarea equivalente a un 10% de la nota del curso.

Esta tarea se entrega en línea, en un archivo comprimido que deberá incluir:

- URL del sitio de Internet pedido en este documento
- Archivos con el código usado y documentado.

Objetivo General

Conocer y practicar el algoritmo de cálculo de distancia euclidiana para medir diferencias y similitudes entre ejemplos de personas o cosas.

Para alcanzar ese objetivo, cada estudiante cumplirá estos objetivos secundarios, es decir, que llevan a conseguir el anterior o general:

Objetivos Secundarios

1. Crear un sitio de Internet (puede usar cualquiera de los servidores de la Sede del Atlántico, sea Turrialba, Guayabo o el de Paraíso o bien puede utilizar otro fuera de la universidad) en PHP (puede usar el servidor MySQL de Turrialba o el de Paraíso, si lo considera necesario). Ese sitio tendrá una página de ingreso con un menú para navegar por diferentes opciones.

2. Poseer un menú permitirá escoger varios formularios donde se podrán ingresar datos de personas o cosas para encontrar los más parecidos o los más diferentes de un conjunto de datos proporcionado más adelante. Este menú nunca debe desaparecer de la vista del usuario y deberá permanecer ubicado en la misma posición de la página o pantalla.

3. Los formularios que el sitio incluirá son
    3.1. El mismo formulario que Ud. llenó en el URL http://multi.ucr.ac.cr/estilo.htm, para detectar el estilo de aprendizaje, puede copiarlo y variar la presentación cuanto considere pertinente. De ese formulario aprovechará el cálculo que da por resultado los valores de las cuatro columnas. Elimine el resto del algoritmo de cálculo incluido en el código JavaScript, porque para determinar el estilo de aprendizaje usará la fórmula de cálculo de distancia de Euclides en lugar del algoritmo original, tomando en cuenta los resultados de las cuatro columnas (CA, EC, EA, OR).

    3.2. Otro formulario para adivinar el recinto de origen de un estudiante (Paraíso o Turrialba), donde el usuario podrá seleccionar su estilo de aprendizaje de los cuatro usados (divergente, convergente, asimilador, acomodador), su último promedio para matrícula y su sexo.

    3.3. Otro formulario para adivinar el sexo de un estudiante, donde el usuario podrá seleccionar su estilo de aprendizaje de los cuatro usados (divergente, convergente, asimilador, acomodador), su último promedio para matrícula y su recinto (Paraíso o Turrialba).

    3.4. Otro formulario para adivinar el estilo de aprendizaje de un estudiante (divergente, convergente, asimilador, acomodador), donde el usuario podrá seleccionar su recinto

(Paraíso o Turrialba), su último promedio para matrícula y su sexo.

3.5. Otro formulario para determinar el tipo de profesor (beginner, intermediate, advanced), a partir de los siguientes criterios que el usuario podrá definir gracias a la interfaz.

Demographic

A. Age.
   - 1= teacher's age <= 30
   - 2= teacher's age > 30 AND <= 55
   - 3= teacher's age > 55

B. Gender.
   - F= female
   - M= male
   - NA= not available

Background

C. Teacher's self-evaluation of his skill or experience teaching the selected subject.
   - B= beginner
   - I= intermediate
   - A= advanced

D. Number of times the teacher has taught this type of course.
   - 1= never
   - 2= 1 to 5 times
   - 3= more than 5 times

E. Teacher's background discipline or area of expertise.
   - DM= decision-making
   - ND= network design
   - O= other

F. Teacher's skills using computers.
   - L= low
   - A= average
   - H= high

G. Teacher's experience using Web-based technology for teaching.
   - N= never
   - S= sometimes
   - O= often

H. Teacher's experience using a Web site.
   - N= never
   - S= sometimes
   - O= often

3.6. Otro formulario para determinar la clasificación de redes (clases A o B), a partir de los siguientes criterios que el usuario podrá definir gracias a la interfaz

a. The network reliability → Reliability (Re): 2 to 5.
b. The number of links → Number of links (Li): 7 to 20.
c. The total network capacity → Capacity (Ca): low, medium, high.
d. The network cost → Cost (Co): low, medium, high.

4. Para cada uno de los formularios del objetivo 3, use el algoritmo de cálculo de distancia visto en clase y que está en el archivo “Similiarityv5.pdf” como ecuación (1):

$$dist(x_i, x_j) \equiv \sqrt{\sum_{r=1}^{r=n} \left[a_r(x_i) - a_r(x_j) \right]^2} \qquad (1)$$

De hecho, esa distancia se puede apreciar como las distancias entre las coordenadas cartesianas de dos puntos P1 y P2, es decir, como la acumulación o suma de diferencias entre puntos en un cuadrante, como está en la ecuación siguiente (2).

$$d_E(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Esta ecuación se refiere solo a un punto dado en comparación con otro, mientras que la ecuación (1) considera la sumatoria de muchas diferencias (r = n) entre dos puntos o ejemplos en comparación, como puede estudiarse en la siguiente ecuación (3).

$$d_E(P, Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}.$$

Fórmulas tomadas de http://es.wikipedia.org/wiki/Distancia_euclidiana

Esta otra versión ayuda a comprender cómo distintos criterios se pueden acumular. Esta versión se refiere a la comparación de topologías de redes

$$Dissimilarity(a_1, a_2) = C_1\sqrt{(G_1 - G_2)^2} + C_2\sqrt{(R_1 - R_2)^2} + C_3\sqrt{(T_1 - T_2)^2} + C_4\sqrt{(L_1 - L_2)^2}$$

donde:

- *a1* and *a2* are the topologies whose dissimilarity is to be calculated.
- *G* is global cost.
- *R* is average reliability.
- *T* is the total capacity.
- *L* is the number of links.
- *C1*, *C2*, *C3*, and *C4* are weights to tune the importance of each of the four criteria in the dissimilarity calculation.

C1, C2, C3, y C4 son valores importantes en este ejemplo, porque dichos pesos ayudan a normalizar los valores de los diferentes criterios, de manera que unos no queden con valores de miles y otros con centésimas o milésimas.

En relación con el formulario 3.1, su sitio Web debe determinar el estilo de aprendizaje tomando en cuenta los resultados de las cuatro columnas (CA, EC, EA, OR). En lugar del algoritmo original (el que está en JavaScript y que termina obteniendo las coordenadas vertical CA-EC y horizontal EA-OR), su tarea usará la fórmula de cálculo de distancia de Euclides. Tomará los valores CA, EC, EA, OR generados a partir de los datos llenados por el usuario y los comparará, uno por uno, con cada registro de una una base de datos que incluye esos valores de unos 200 estudiantes (hoja “RecintoEstilo” del libro Excel

"DatosTarea1.xls"). Al final, su algoritmo de cálculo de distancias encontrará el registro más parecido (con la menor diferencia o la más cercana a cero) y el estilo asignado a ese registro será el estilo del registro propuesto en el formulario. Su interfaz mostrará el estilo asignado al registro que fue evaluado a partir del formulario.

En relación con los formularios de los puntos 3.2, 3.3 y 3.4, su tarea debería aprovechar el código del algoritmo de cálculo de distancias usado en el punto 3.1 con el fin de calcular la similitud entre los datos especificados por el usuario en cada uno de esos formularios y 77 registros que incluyen datos para sexo, recinto, último promedio de matrícula, valores para CA, EC, EA, OR y estilo de aprendizaje (eso también está en una hoja "EstiloSexoPromedioRecinto" del libro Excel "DatosTarea1,xls"). Lo que pasa es que con estos tres formularios no se usan los valores CA, EC, EA, OR sino el recinto de origen de cada estudiante (Paraíso o Turrialba), el estilo de aprendizaje de los cuatro usados (divergente, convergente, asimilador, acomodador), el último promedio para matrícula de cada alumno en la base y el sexo de cada quien.

Cuando el usuario detalla sus datos en cada formulario y los envía, el sitio de Internet calcula la distancia de esos datos con los alojados en la base y encuentra el o los más parecidos o cercanos (vecino más próximo k-nn) y de allí se toma la información buscada, por ejemplo, el recinto, el sexo o el estilo de aprendizaje que se desplegará en la interfaz.

En el formulario del punto 3.5, también se calculará la diferencia a partir de los ocho criterios usados en el respectivo formulario. Los datos aportados desde la interfaz se comparan con los de la base de profesores (teachers) que incluye tres clases o tipos de educadores. La respuesta es la clase o tipo de maestro al que pertenece o en el que su aplicación clasifica al registro dado desde el formulario.

En cuanto al formulario 3.6, el usuario escogerá los valores de los cuatro atributos de una supuesta red y su sitio buscará el o los vecinos más próximos (K-NN) de entre una base de configuraciones de redes. Con base en la red o redes más parecidas, su algoritmo de cálculo de distancias determinará si el ejemplo de red dado desde el formulario pertenece a la clase o tipo A o si se puede clasificar como clase B y así lo desplegará en la interfaz.

5. El código que Ud. escriba deberá estar documentado con detalle y claridad para que el profesor pueda comprender y revisar lo que Ud. hizo, paso a paso. En particular, debe quedar claro como su algoritmo se adapta para calcular distancias en relación con los formularios de los punto de 3.1 a 3.6. y cómo compara el ejemplo dado registro por registro.

Si Ud. quiere, puede agregar trabajo extra que el profesor tomará en cuenta: evaluar la eficiencia del algoritmo de cálculo de similitud mediante un método llamado "10-fold *cross-validation"*. Eso lo puede hacer con la base de 200 registros, caso del formulario 3.5. Eso implica incluir el código, documentado, que hace las pruebas y los resultados por cada bloque y el porcentaje final de eficiencia. Eso debería verse mediante un link en el menú general del sitio.

"Cross-validation is a procedure that consists of dividing the training dataset data into a number k of blocks or partitions. Testing is performed by extracting one partition from the dataset; each example of the extracted block is classified to measure the error rate of the classifier. After finishing with a partition, it is reinserted into the dataset and another block of examples is tested. These operations are repeated for each partition, and finally, an average of the number of errors is calculated. The inverse of this

average shows the degree of accuracy of the classifier”.

Datos proporcionados para la tarea

En el sitio en línea se facilita un archivo Excel (extensión xls) con los siguientes datos:

- Cerca de 200 registros para lo referente al formulario de estudiantes 3.1.
- Cerca de 80 registros para lo referente a formularios de estudiantes 3.2, 3.3, 3.4.
- Cerca de 20 registros para lo referente a formularios de profesores 3.5.
- Cerca de 35 registros para lo referente a redes formulario 3.6.