UBA – Facultad de Ciencias Exactas y Naturales – Departamento de Computación

Algoritmos y Estructura de Datos I

Primer cuatrimestre de 2018

Versión: 21 de abril de 2018

TPE - "Extracción de características principales"

Fecha de entrega: Viernes 27 de Abril (hasta las 17hs)
Devolución y coloquio: Viernes 4 de Mayo
Recuperatorio: Viernes 18 de Mayo

Figura 1: A "11" le hubiese venido bien hacer este TP

CORRECCIONES (21/4):

- Ejercicio 9: -1.66 en vez de 1.66 en el ejemplo.
- Ejercicio 10: "huecos" en vez de "missings".

1. INTRODUCCIÓN

En los últimos años, a habido un gran interés técnicas de estadística y aprendizaje automático (machine learning) aplicado a problemas provenientes de todo tipo de áreas. Estas técnicas se basan en "aprender" patrones a partir de ejemplos. Por ejemplo, tareas como la de determinar a partir de una grabación si una persona está alegre o triste puede abordarse a partir de recolectar miles de ejemplos de personas alegres hablando y miles de ejemplos de grabaciones de personas tristes. Luego, procesar estas grabaciones en busca de características que las definan para después identificar patrones en estas características que permitan clasificar entre una u otra condición.

En este paradigma, un paso casi esencial es el de extraer características a partir de los datos que se deseen clasificar. Por ejemplo, siguiendo con el mismo ejemplo, es posible que características interesantes sean, la intensidad media de la voz, la cantidad de palabras por minuto, el tono de voz máximo y mínimo, la calidad de la voz (por ejemplo para saber si la persona está susurrando o hablando de manera clara), la cantidad de veces que dice cada una de las palabras que dice, etc. Como se puede observar, algunas características son más sencillas de extraer que otras y crear algoritmos que obtengan estos atributos a partir de la grabación parece una tarea no sencilla, pero realizable.

En este trabajo, nos dedicaremos a facilitar el trabajo de las personas interesadas en aplicar aprendizaje automático sobre problemas que utilicen señales temporales (como la de la voz). Para ello, especificaremos e implementaremos algoritmos sencillos del área del procesamiento de secuencias temporales con el objetivo de extraer características básicas a partir de ellas. Una señal para nosotros será simplemente una secuencia no vacía de números enteros que podrán contener valores en el rango $[-2^{15}, 2^{15} - 1]$. Por ejemplo, $s_1 =< 1, 2, -10, 0, 0, 0, 0, 1, 2, 100 >$ será una señal en este trabajo.

2. EJERCICIOS

Especificar los siguientes problemas dados el **renombre** de tipos:

type *señal*= $seq\langle\mathbb{Z}\rangle$

Ejercicio 1 *:* **proc esValida(in s: *señal*, out result :** Bool**)**
Que dada una señal compruebe si es válida.

Ejercicio 2 *:* **proc media(in s: *señal*, out media :** $\mathbb{R}$**)**
Que dada una señal calcule su promedio.

Ejercicio 3 *:* **proc ctrlf(in s: *señal*, in x :** $\mathbb{Z}$**, out indices :** $seq\langle\mathbb{Z}\rangle$**)**:
Que busque las apariciones de x en s y devuelva en que índices aparece.

Ejercicio 4 *:* **proc máximo(in s: *señal*, out posicion :** $\mathbb{Z}$**, out result :** $\mathbb{Z}$**)**
Que dada una señal calcule su único valor máximo y la posición en que este aparece.

Ejercicio 5 *:* **proc medianaEntera(in s: *señal*, out latencia :** $\mathbb{Z}$**, out mediana :** $\mathbb{Z}$**)**
Que dada una señal, calcule un valor similar a su mediana. La definición de medianaEntera que utilizaremos es la siguiente: dado un conjunto de números, la medianaEntera está definida como el valor en la posición central luego de ordenar los números de menor a mayor. En caso que la cantidad de elementos sea par, la mediana estará definida como el elemento en el primero de los dos puntos centrales. La latencia, será la posición en que aparezca dicho valor por primera vez en la señal.

Ejercicio 6 *:* **proc histograma(in s: *señal*, in bins:** $\mathbb{Z}$**, out cuentas :** $seq\langle\mathbb{Z}\rangle$**, out limites :** $seq\langle\mathbb{R}\rangle$**)**
Que dada una señal calcule su histograma. Dado un conjunto de números, un histograma puede calcularse dividiendo el rango en que se mueven los valores en segmentos de igual ancho (también llamados bins) y se cuenta cuántos valores caen en cada bin.

- el parámetro **cuentas** deberá corresponderse con la cantidad de elementos en cada bin.
- el parámetro **limites** deberá contener los límites de cada uno de los bins en orden como se muestra en los siguientes ejemplos:
  - si $s =< 33, 24, 24, 1, 62, 88, 94, 79, 25, 24 >$ y $bins = 4$, el resultado debería ser $cuentas =< 4, 2, 1, 3 >$, limites$=< 1.00, 24.25, 47.50, 70.75, 94.00 >$.
  - si $s =< 0, 1, 2, 10 >$ y $bins = 5$, el resultado debería ser $cuentas =< 2, 1, 0, 0, 1 >$, $limites =< 0, 2, 4, 6, 8, 10 >$

Aclaración: como puede verse en el ejemplo, los intervalos se consideran cerrado-abierto, salvo el caso del último intervalo que será cerrado-cerrado.

Ejercicio 7 *:* **proc distanciaEuclidena(in p: *señal*, in q: *señal*, out distancia :** $\mathbb{R}$**)**
Que calcule la distancia euclidiana entre dos señales p y q de la misma longitud. Esta distancia está definida como:

$$d_E(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

Ejercicio 8 *:* **proc distanciaAcordeon(in s: *señal*, in q: *señal*, out distancia :** $\mathbb{R}$**, out asignaciones :** $seq\langle\mathbb{Z} \times \mathbb{Z}\rangle$**)**
Calcular la distancia que hay entre dos señales no siempre es trivial. Un caso conocido es en la tarea de reconocimiento del habla en donde es común comparar sonidos grabados contra prototipos de cómo suena cada letra. Por ejemplo, se le puede pedir a una persona que grabe el sonido de una vocal y luego comparar contra señales prototípicas de cada vocal para determinar cuál fue la pronunciada. El problema que surge es que nunca diremos una vocal dos veces de la misma manera, alguna veces las estiramos, otras veces las acortamos. Es por problemas como estos, que es necesario calcular una distancia más inteligente que la distancia punto a punto.

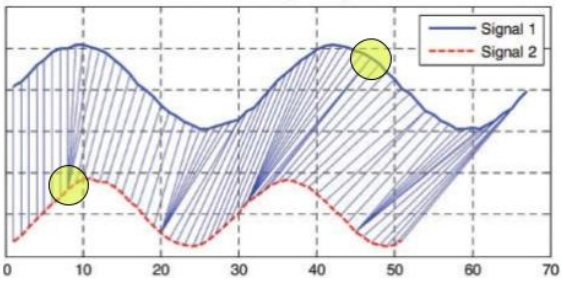

Figura 2: Distancia acordeón

Para este ejercicio, necesitaremos especificar un algoritmo que que calcule la distancia mínima que pueda lograrse al estirar las señales, es decir, decidiendo qué punto de la señal s compararemos con qué punto/s consecutivos de la señal q y viceversa. Veamos por ejemplo la Figura 2. Como puede verse, hay varios puntos (algunos marcados con un círculo) en donde se eligió comparar un punto de una de las señales contra varios consecutivos de la otra. Una vez asignados los puntos, se calculará la distancia euclidiana antes definida. Es importante notar que las asignaciones no pueden cruzarse y que todo punto debe quedar asignado a algún otro.

El parámetro *asignaciones* será una lista de pares que contenga todas las asignaciones de puntos entre las dos señales. La primera componente se referirá a posiciones de la secuencia s y la segunda componente posiciones de la secuencia q. Por ejemplo,

- si $s =< 33, 25, -1, 3, 1 >$ y $q =< 33, 24, 26, 1 >$, el resultado debería ser
  $asignaciones =< (0,0), (1,1), (1,2), (2,3), (3,3), (4,3) >$
  $distancia = 3.16$.

Ejercicio 9 *:* **proc slidingWindows(in s: *señal*, in tamaños :** $seq\langle\mathbb{Z}\rangle$ **, out promedios :** $seq\langle\mathbb{R}\rangle$**, out ventanas:** $seq\langle\mathbb{Z}\times\mathbb{Z}\rangle$**)**

Este problema se basa en calcular promedios dentro de ventanas (intervalos de índices) que irán recorriendo la señal. Para ello, recibiremos como parámetro los distintos tamaños de ventana que utilizaremos.

Cada ventana deberá deslizarse por la señal con un paso constante (igual al tamaño de la ventana) hasta llegar al final. En caso de no coincidir el final de la ventana con el final de la señal, deberá extenderse la señal con el último valor tantas veces como sea necesario.

El parámetro *promedios* contendrá cada uno de los promedios calculados. El parámetro *ventanas* indicará, cada promedio, a qué intervalo de índices (cerrado-cerrado) se corresponde. Tener en cuenta que al ser posible extender las señales, las ventanas pueden contener intervalos de índices que no existan en la señal original. Por ejemplo:
Si $s =< 33, 25, -1, 3, 1 >$ y $tamaños = < 3, 2 >$. Una solución válida es $ventanas =< (0,1), (2,3), (4,5), (0,2), (3,5) >$ y $promedios =< 29, 1, 1, 19, -1.66 >$

Ejercicio 10 *:* **proc completar(inout s: *señal*, in huecos:** $seq\langle\mathbb{Z}\rangle$**)**

A veces las señales tienen valores faltantes (o huecos). Para este ejercicio, se contará con una señal incompleta en donde en las posiciones faltantes (pasadas por parámetro en la variable *huecos*) hay ceros. Se pide completar la señal de la siguiente manera: donde haya un hueco, se buscan los puntos más cercanos que contengan valores y se calcula su promedio. El hueco deberá ser completado con el entero más cercano a dicho promedio.

Ejercicio 11 *:* **proc sacarOutliers(inout s: *señal*, out borrados :** $seq\langle\mathbb{Z}\rangle$**)**:

Para limpiar una señal, es muy común el borrado de *outliers* (definidos como puntos que superan al percentil-95[1] de los datos). En este ejercicio, se pide que se reemplace los outliers de la señal por ceros y en la variable *borrados* se indique cuáles fueron los indices en que han sido eliminados.

[1] `https://en.wikipedia.org/wiki/Percentile`

Términos y condiciones

El trabajo práctico se realiza de manera grupal, pero su aprobación será individual. Para aprobar el trabajo se necesita:

- Que todos los ejercicios estén resueltos.
- Que las soluciones sean correctas.
- Que el lenguaje de especificación esté bien utilizado.
- Que las soluciones sean prolijas: evitar repetir especificaciones innecesariamente y usar adecuadamente las funciones y predicados auxiliares.
- Que no haya casos de sub-especificación ni sobre-especificación.
- Que demuestren en el coloquio que entienden la solución de cualquiera de los ejercicios y puedan explicarlas con sus palabras.

Pautas de Entrega

Se debe enviar un e-mail conteniendo informe a la dirección algo1-tt-doc@dc.uba.ar. Dicho mail debe cumplir con el siguiente formato:

- El título debe ser `[ALGO1;TPE]` seguido inmediatamente del nombre del grupo.
- En el cuerpo del email deberán indicar: Nombre, apellido, libreta universitaria de cada integrante.
- El informe deberá estar adjuntado en el email con formato .pdf.

Importante: se admitirá un único envío, sin excepción alguna. Por favor planifiquen el trabajo para llegar a tiempo con la entrega.

Fecha de entrega: Viernes 27 de Abril (hasta las 17hs)
Devolución y coloquio: Viernes 4 de Mayo
Recuperatorio: Viernes 18 de Mayo