Control 4

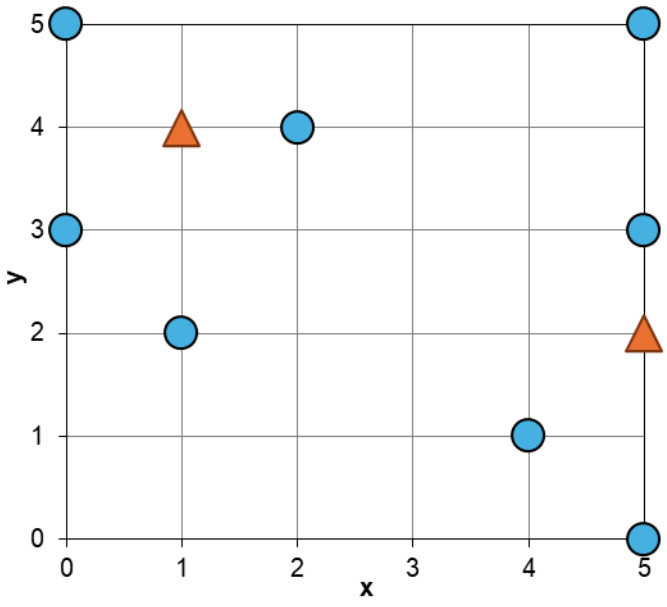
CC5213 – Recuperación de Información Multimedia
Departamento de Ciencias de la Computación
Universidad de Chile
Profesor: Juan Manuel Barrios
Fecha entrega: lunes 07 de julio de 2025

Semana 13 (LSH)

Se tiene el conjunto **R** de 8 vectores (**a**-**h**) y el conjunto **Q** de 2 vectores (**q1** y **q2**) en el espacio bidimensional [0, 5] x [0, 5] según la siguiente tabla y diagrama:

R	x	y
a	0	3
b	0	5
c	1	2
d	2	4
e	4	1
f	5	0
g	5	3
h	5	5

Q	x	y
q1	1	4
q2	5	2



Construya un índice **LSH** (Locality-Sensitive Hashing) mediante los siguientes pasos:

a. (0.5 puntos) Escriba la notación unaria para los vectores de **R** y **Q**.

b. (0.5 puntos) Construya un índice LSH calculando 3 tablas de proyección, cada una de 3 bits al azar sobre los conjuntos **R** y **Q**. La generación de números aleatorios es a su decisión.

c. (0.5 puntos) Localice el vecino más cercano (aproximado) de **q1** y **q2** utilizando el índice LSH que construyó anteriormente.

Semana 13 (PCA)

Se tiene un conjunto de 12 imágenes y a cada imagen se le calculó un descriptor de contenido de 5 dimensiones, creando el conjunto de descriptores **R**. Se desea reducir la dimensionalidad de **R** utilizado PCA para luego hacer búsquedas aproximadas.

El promedio de los descriptores de contenido de **R** es:

$$\bar{x} = \begin{pmatrix} 2 \\ -1 \\ 3 \\ -4 \\ 6 \end{pmatrix}$$

Llamaremos **S** al conjunto de descriptores que se obtiene de restar $\bar{x}$ de los descriptores de **R**, es decir:

$$S = \{y \,|\, \forall x \in R,\ y = x - \bar{x}\}$$

Notar que **S** es un conjunto de descriptores centrados (promedio igual a cero).

Al calcular las covarianzas entre las coordenadas de los vectores de **S** se obtuvo la siguiente matriz de covarianza ***C***:

$$C = \begin{pmatrix} 3.6 & 1.2 & -1.8 & 1.9 & -3.6 \\ 1.2 & 2.7 & -1.8 & 1.9 & -1.5 \\ -1.8 & -1.8 & 4.3 & 0.3 & 2.6 \\ 1.9 & 1.9 & 0.3 & 3.5 & -1.6 \\ -3.6 & -1.5 & 2.6 & -1.6 & 4.1 \end{pmatrix}$$

Los cinco valores propios y vectores propios de ***C*** son:

$$\lambda_1 = 2.5 \qquad \lambda_2 = 0.2 \qquad \lambda_3 = 4.3 \qquad \lambda_4 = 11.1 \qquad \lambda_5 = 0.1$$

$$v_1 = \begin{pmatrix} -0.5 \\ 0.7 \\ -0.3 \\ 0.2 \\ 0.4 \end{pmatrix} \quad v_2 = \begin{pmatrix} -0.2 \\ 0.5 \\ 0.5 \\ -0.5 \\ -0.5 \end{pmatrix} \quad v_3 = \begin{pmatrix} 0.1 \\ 0.1 \\ 0.7 \\ 0.7 \\ 0.1 \end{pmatrix} \quad v_4 = \begin{pmatrix} 0.5 \\ 0.3 \\ -0.4 \\ 0.3 \\ -0.6 \end{pmatrix} \quad v_5 = \begin{pmatrix} -0.7 \\ -0.4 \\ -0.2 \\ 0.4 \\ -0.5 \end{pmatrix}$$

Se desea utilizar el método PCA para proyectar los descriptores de **S** desde 5 dimensiones a **una única dimensión** creando el conjunto **T**.

a. (0.25 puntos) Señale la matriz de transformación ***W*** que se debe aplicar sobre **S** para crear el conjunto **T**.

b. (0.25 puntos) Señale la "cantidad de información" que se mantiene (según PCA) al proyectar los descriptores de **S** en **T**.

El conjunto **R** fue proyectado con PCA obteniendo el siguiente conjunto **T** de 12 descriptores de una dimensión:

$$a = (12.1) \quad b = (2.1) \quad c = (9.5)$$
$$d = (-3.9) \quad e = (7.1) \quad f = (-0.4)$$
$$g = (0.2) \quad h = (-17.2) \quad i = (23.2)$$
$$j = (-7.2) \quad k = (16.4) \quad l = (-1.9)$$

Se tiene una imagen de consulta a la que se le calculó su descriptor de contenido ***q***:

$$q = \begin{pmatrix} 8 \\ 3 \\ 3 \\ 2 \\ 7 \end{pmatrix}$$

c. (0.5 puntos) Señale el vector de ***T*** que será el vecino más cercano a ***q*** cuando ***q*** es proyectado a una dimensión con la transformación PCA descrita anteriormente.

Semana 14 (distancias)

1. Se tiene un vector de consulta **q** y tres 3 vectores (**a**, **b** y **c**) de cinco dimensiones:

$$q = \begin{pmatrix} 6 \\ 4 \\ 3 \\ 1 \\ 3 \end{pmatrix} \qquad a = \begin{pmatrix} 4 \\ 2 \\ 5 \\ 3 \\ 5 \end{pmatrix} \qquad b = \begin{pmatrix} 6 \\ 4 \\ 5 \\ 9 \\ 3 \end{pmatrix} \qquad c = \begin{pmatrix} 5 \\ 5 \\ 4 \\ 10 \\ 3 \end{pmatrix}$$

   a. (0.3 puntos) Calcule la distancia de **q** con **a**, **b** y **c** usando distancia L_P con **p=2** y señale el más cercano a **q**.

   b. (0.3 puntos) Calcule la distancia de **q** con **a**, **b** y **c** usando distancia L_P con **p=0.5** y señale el más cercano a **q**.

   c. (0.4 puntos) Calcule la distancia de **q** con **a**, **b** y **c** usando distancia DPF con **p=2** y **m=4** y señale el más cercano a **q**.

$$\mathrm{L_p}(\vec{x}, \vec{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \qquad \mathrm{DPF}(\vec{x}, \vec{y}) = \left(\sum_{i \in \Delta_m} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

2. Dadas dos funciones de distancia d1() y d2() que cumplen las propiedades métricas (no negatividad, reflexividad, simetría y desigualdad triangular).

   a. (0.25 puntos) Demuestre formalmente si la siguiente función d3() cumple o no cumple la propiedad de **desigualdad triangular**:

$$d_3(x, y) = \frac{d_1(x, y) + d_2(x, y)}{2}$$

   b. (0.25 puntos) Demuestre formalmente si la siguiente función d4() cumple o no cumple la propiedad de **desigualdad triangular**:

$$d_4(x, y) = \min\{d_1(x, y),\ d_2(x, y)\}$$

Hint: La desigualdad triangular dice que:

$$\forall x, y, z \quad d(x, z) \leq d(x, y) + d(y, z)$$

Para demostrar formalmente que d3() o d4() cumple la desigualdad triangular, debe asumir que las funciones d1() y d2() cumplen la desigualdad triangular, y luego haciendo operaciones matemáticas concluir que d3() o d4() también la cumple.

Para demostrar formalmente que d3() o d4() no cumple la desigualdad triangular, basta con encontrar un caso donde no se cumple, es decir, mostrar que:

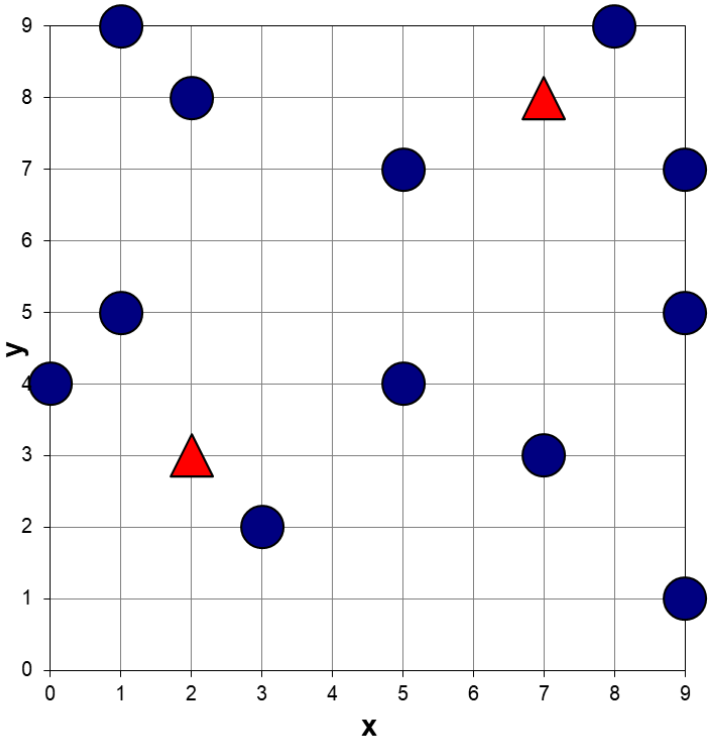
$$\exists x, y, z \;\; d(x, z) \nleq d(x, y) + d(y, z)$$

Semana 14 (LAESA)

El conjunto **R** contiene 12 vectores de dos dimensiones (**a**-**l**) y el conjunto de consulta **Q** contiene dos vectores (**q1** y **q2**):

R	x	y
a	5	4
b	3	2
c	2	8
d	5	7
e	9	5
f	0	4
g	8	9
h	1	5
i	9	1
j	9	7
k	7	3
l	1	9

Q	x	y
q1	2	3
q2	7	8



Se desea localizar para cada objeto de **Q** su vecino más cercano en **R** de acuerdo a la distancia **Manhattan** o $\mathbf{L_1}$ utilizando el enfoque métrico.

Suponga que se selecciona al elemento **i** como pivote, es decir, **P = { (9,1) }**

a. (0.5 puntos) Calcule la Tabla de Pivotes para **P** usando la distancia $\mathbf{L_1}$.

b. (0.5 puntos) Resuelva la búsqueda exacta del vecino más cercano en **R** para **q1** y para **q2** usando cotas inferiores según **P**.

c. (0.5 puntos) Calcule la complejidad interna y externa del índice al resolver ambas búsquedas. ¿Se realizaron más o menos cálculos que en un linear scan?

Semana 15

Durante la última parte del curso se vieron varios ejemplos de indexamiento sobre dos datasets distintos (llamados “A” y “B”). Ambos datasets tenían la misma cantidad de vectores y la misma dimensionalidad, por lo que el linear scan tomaba el mismo tiempo en ambos datasets.

(0.5 puntos) Señale posibles razones de porqué las búsquedas aproximadas con los índices estudiados obtienen distinto comportamiento en cada dataset, es decir, en “A” la búsqueda aproximada con cada índice lograba una relación de calidad vs tiempo, que era bastante distinta a lo que lograba el mismo índice en “B”. Sea conciso en la explicación (máximo 3 líneas).

Entrega

- Puede desarrollarlo en papel y enviar una foto (.jpg, .png), o puede desarrollarlo en formato digital en una planilla (.xlsx .ods), un documento (.docx) u otro formato exportado a .pdf.

- El plazo máximo de entrega es el **lunes 07 de julio de 2025** hasta las 23:59. Existirá una segunda fecha (por definir) para entregar su respuesta sin descuentos en la nota.

El control es *individual* y debe ser de su autoría, es decir, no pueden ser resueltos por otro estudiante, no se pueden copiar respuestas de Internet, no se permite usar ChatGPT ni similares. En caso de detectar copia o plagio se asignará nota 1.0 a las o los estudiantes involucrados.