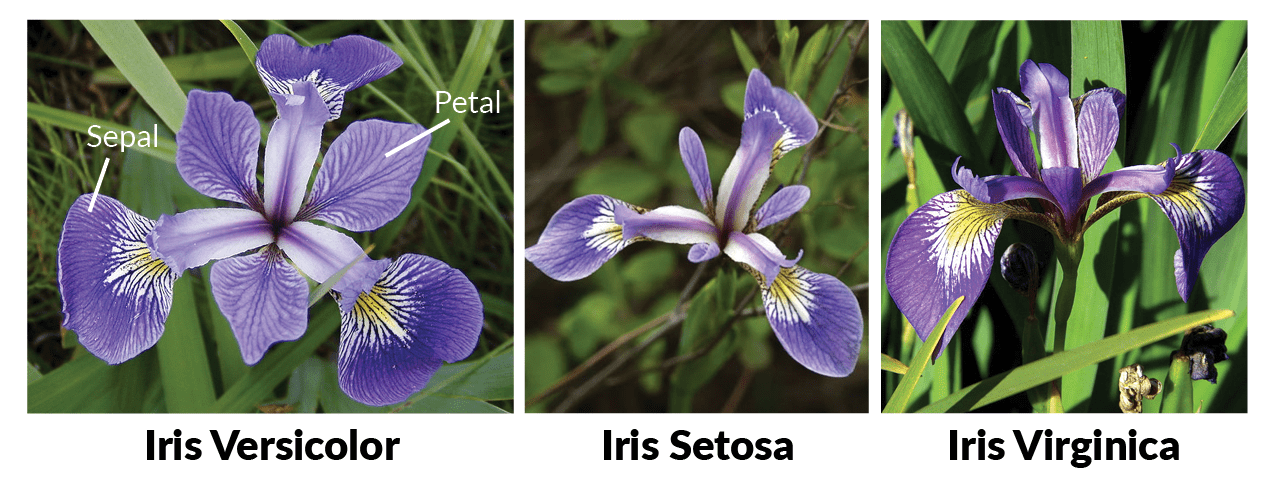
**Búsqueda por Similitud**

Profesor Heider Sanchez

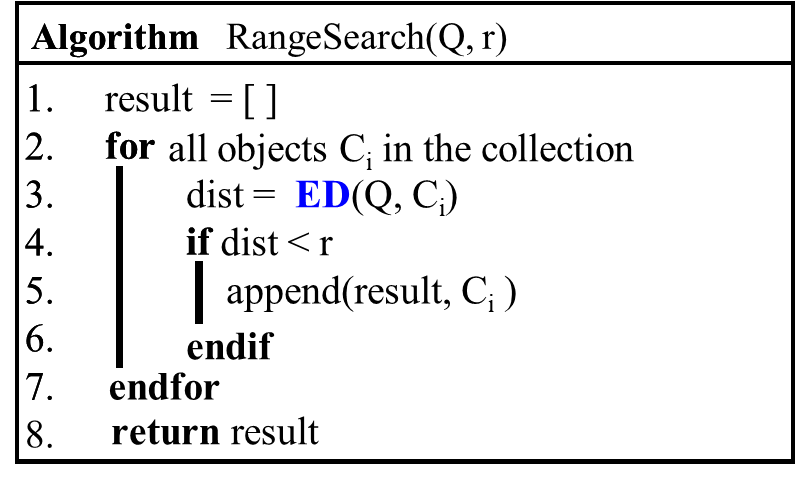
El objetivo del laboratorio es aplicar la búsqueda por rango y la búsqueda de los k vecinos más cercano sobre un conjunto de vectores característicos.

Se toma como referencia la colección de imágenes de flores ***Iris*** *(*[*https://archive.ics.uci.edu/ml/datasets/iris*](https://archive.ics.uci.edu/ml/datasets/iris)*),* en donde cada imagen es representada por un vector característico de 4 dimensiones que recoge información del ancho y largo del sépalo y del pétalo. Además, las imágenes están agrupadas en tres categorías: *versicolor, setosa y virginica*.



**P1. Búsqueda por Rango**

Implementar en cualquier lenguaje de programación el algoritmo lineal de búsqueda por rango, el cual recibe como parámetro el objeto de consulta y un **radio de cobertura**. Luego usando la distancia Euclidiana (ED) se retorna todos los elementos que son cubiertos por el radio.



* Aplique la búsqueda para 3 elementos de la colección (Q15, Q82, Q121) y para tres valores de radio ().
* El objeto de consulta debe ser retirado de la colección antes de aplicar la búsqueda.
* Para saber que valores de radio seleccionar, debe primero realizar un análisis de la distribución de las distancias computando N veces la distancia entre dos elementos aleatorios de la colección.
* Para evaluar la efectividad del resultado se debe usar la medida de Precisión ¿Cuántos de los objetos recuperados pertenecen a la misma categoría de la consulta?:

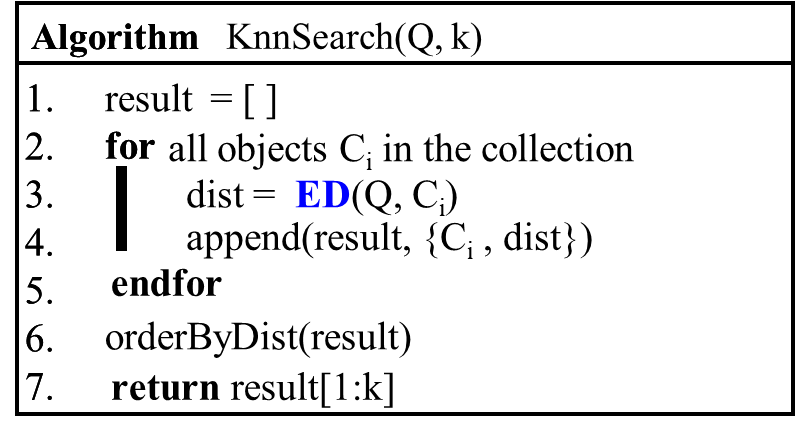
A continuación, se proporciona el cuadro que debe ser llenado por el alumno.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| **1.05** | 1 | 1 | 0.555 |
| **2.55** | 1 | 0.543 | 0.473 |
| **3.76** | 0.549 | 0.333 | 0.471 |

**Los tres radios se obtuvieron tomando n = 150 distancias de cada vector con un vector aleatorio del documento. Cada radio representa el 1er, 2do y 3er cuartil del vector de distancia calculado.**

**P2. Búsqueda KNN**

Usando los mismos objetos de consulta del ejercicio anterior, implementar y aplicar el algoritmo lineal de búsqueda de los vecinos más cercanos (KNN) variando el entre {2, 4, 8, 16, 32}.

****

**\*\* La mejor forma de implementación es gestionando la lista de resultado como una cola de prioridad máxima. Analice la complejidad.**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | 1 | 1 | 1 |
|  | 1 | 1 | 1 |
|  | 1 | 1 | 0.875 |
|  | 1 | 1 | 0.625 |
|  | 1 | 1 | 0.500 |

**Preguntas:**

1. ¿Cuál es la complejidad computacional de ambos métodos de búsqueda en función de cálculos de la ED?

RangeSearch: θ(n). Es necesario hallar la ED entre el query y todos los n elementos del documento.

1. ¿Cuál de los dos métodos de búsqueda usted usaría en un ambiente real de recuperación de la información? Sustente su respuesta.

Consideramos que el knnSearch es más efectivo debido a que puede dar similitudes sin importar la escala de sus valores, caso que el range search si necesita, que es el caso de radio que usa, pues tiene que determinar un radio para el range search