Práctica 1. Caso Práctico: Sistema (preliminar) de clasificación de instancias utilizando el método k-nn

Propósito: Desarrollar en Java, usando el paradigma de orientación a objetos, un sistema software preliminar para la clasificación de instancias usando el método *k*-nn.

Enunciado: El método *k*-nn (*k*-nearest neighbours o *k*-vecinos más cercanos) es una algoritmo de *minería de datos* basado en "*vecindad*", que se puede utilizar para la tarea (predictiva) de *clasificación*.

Se dispone de un *dataset*, conjunto de *casos* o *instancias* ya clasificados y almacenados en un fichero. A este conjunto se le denomina *conjunto de entrenamiento*.

Cada fila del conjunto de entrenamiento tiene el mismo número de atributos y corresponde a una instancia o caso. Los atributos describen la instancia y cada instancia pertenece a una sola clase.

El número de clases es finito y conocido a partir de las instancias almacenadas en el dataset y ya etiquetadas (asignadas a una clase).

Cuando el número de clases es 2, tenemos un problema de *clasificación binaria*. Cuando el número de clases es superior a 2 la tarea es de *multiclasificación*.

La clase de una instancia etiquetada viene dada por el valor que toman uno o varios atributos. Para esta práctica utilizaremos un solo atributo de clase que, por simplicidad, consideraremos que será el último de cada instancia.

El resto de atributos de la instancia se utilizarán, en principio, para predecir la clase de instancias futuras y son todos de tipo numérico.

En la versión básica del método *k*-nn, un nuevo caso es asignado a una clase dada si ésta es la clase más frecuente entre las *k* instancias de entrenamiento *más cercanas*.

Generalmente se usa la *distancia euclidea* para medir la "*cercanía*" entre instancias, la cual viene dada por la siguiente expresión:

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{p}(x_{ir} - x_{jr})^2}$$

donde x_i y x_j corresponden a los casos *i* y *j* del dataset.

También se pueden utilizar otras métricas como la distancia de Manhattan, distancia de Chebychev, distancia del coseno, distancia de Mahalanobis, …

Se pide implementar en Java el método básico *k*-nn, donde *k* es un número natural que introduce el usuario, usando como métrica la distancia euclídea. Los datos del conjunto de entrenamiento se leerán desde un fichero en formato CSV (el nombre del fichero lo proporciona el usuario), donde el último valor almacenado en cada fila corresponde al valor de clase. Los atributos numéricos del dataset deberán ser normalizados entre 0 y 1 (preprocesado de datos), porque en caso contrario siempre prevalecen sobre los demás los de mayor valor absoluto a la hora de calcular las distancias. Para esta normalización se utilizará la expresión:

$$\hat{x}_i = \frac{x_i - min_i}{max_i - min_i}$$

Los casos a clasificar se leerán, o bien desde un fichero, o se introducirán manualmente. El programa mostrará, para cada caso introducido, la clase asignada.

Los ficheros de datos correspondientes a los conjuntos de entrenamiento que se utilizarán en esta práctica son:

iris.csv - El dataset iris, también conocido como dataset Iris de Fisher, es un conjunto de datos multivariante introducido por Sir Ronald Fisher en 1936, para realizar análisis discriminante sobre el mismo. El dataset consiste de 150 instancias: 50 instancias por cada una de las 3 especies de la flor Iris (Iris setosa, Iris virginica e Iris versicolor). Para cada instancia se midieron cuatro características: la longitud y ancho de los sépalos y los pétalos, en centímetros. A cada caso se le asignó la especie a la que pertenecía.

glass.csv – El dataset glass consiste de 214 instancias correspondientes a diferentes tipos de cristal. Para cada caso se almacenan 9 características distintas: RI - Refractive Index, Na – Sodio, Mg – Magnesio, Al – Aluminio, Si – Silicio, K – Potasio, Ca – Calcio, Ba – Bario y Fe – Hierro, así como el tipo de cristal que le corresponde. Hay 6 tipos (clases) distintos: build wind float, build wind not-float, vehic wind float, containers, tableware y headlamps.

Referencias:

Wikipedia: http://es.wikipedia.org/wiki/Knn