Práctica 2

Mikel de Velasco, Ieltzu Irazu, M Inés Fernandez

18 de octubre de 2015

1 Introducción

Se pide disear un clasificador que implemente el método k-NN básico con la distancia de Minkowski según la expresión (4) donde n representa el número de atributos empleados para caracterizar las muestras.

$$d(a,b) = \left[\sum_{i=1}^{n} |a_i - b_i|^m\right]^{\frac{1}{m}} \tag{1}$$

El clasificador permitirá seleccionar tanto el número de vecinos a explorar (k) como el parámetro m de la expresión (4). Elegir el lenguaje de programación que se considere más apropiado para el diseo. Para inferir el clasificador se dispone de un conjunto de datos (Diabetes.arff). El conjunto de datos dispone de 768 instancias para inferir el modelo. Para describir las instancias se utilizan 8 atributos más la clase. Los atributos son los siguientes:

- Number of times pregnant
- Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- Diastolic blood pressure
- Triceps skin fold thickness
- 2-Hour serum insulin
- Body mass index
- Diabetes pedigree function
- Age
- Class variable

La clase que hay que determinar es si el paciente tiene diabetes o no.

1.1 Parámetro k

El parámetro k es el parámetro que utilizaremos para indicar cuantos vecinos vamos a usar para evaluar el clasificador. Cuantos más usemos mejor, pero, también aumentara el costo del algoritmo.

1.2 Parámetro m

El parámetro m sera el parámetro que utilizaremos para saber que distancia tenemos que utilizar para evaluar nuestro clasificador.

En esta práctica hemos implementado 3 distintas distancias:

1. Distancia de **Manhattan**:

   La distanciade Manhattan (2), entre dos vectores **x**, **y** en un espacio vectorial real n-dimensional con un sistema de Coordenadas cartesianas fijo es la suma de las longitudes de las proyecciones del segmento de lnea entre los puntos sobre el sistema de ejes coordenados.

$$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i| \tag{2}$$

   Donde n es el número de dimensiones.

2. Distancia **Euclídea**:

   Un espacio euclídeo es un espacio vectorial formado sobre los nmeros reales de dimensin finita, en que la norma es la asociada al producto escalar ordinario. Para cada nmero entero no negativo n, el espacio euclídeo n-dimensional se representa por el smbolo $\mathbb{R}^n$ y es el conjunto de todas las tuplas ordenadas $(x_1, x_2, \ldots, x_n)$ en donde cada x_i es un número real, junto con la funcin distancia entre dos puntos $(x1, \ldots, x_n)e(y1, \ldots, y_n)$ definida por la frmula:

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3}$$

   Esta función distancia es una generalización del teorema de Pitágoras y se denomina Distancia euclidiana (3).

3. Distancia de **Minkowski**:

   En fsica matemtica, el espacio de Minkowski (o espacio-tiempo de Minkowski) es una variedad lorentziana de cuatro dimensiones y curvatura nula, usada para describir los fenmenos fsicos en el marco de la teora especial de la relatividad de Einstein.

   En el espacio de Minkowski pueden distinguirse tres dimensiones espaciales ordinarias y una dimensin temporal adicional, de tal manera que todas juntas forman una 4-variedad y as representar al espacio-tiempo.

El espacio-tiempo de Minkowski es una variedad lorentziana de curvatura nula e isomorfa a $\mathrm{M}_0 = (^4, \eta)dondeeltensormtricopuedell$
$-dx^0 \otimes dx^0 + dx^1 \otimes dx^1 + dx^2 \otimes dx^2 + dx^3 \otimes dx^3(4)$

2 Metodología

En este apartado desarrollaremos nuestro propio algoritmo K-Nearest Neighbors y describiremos el diseño y cómo hemos implementado este algoritmo. Además haremos una breve descripción de como hacer funcionar el programa.

El algoritmo que hemos desarrollado es básicamente una adaptación del clasificador IBk ya que hemos cogido su código y le hemos cambiado la forma de estimar que estaban utilizando.

Nuestro programa puede ser ejecutado desde eclipse de una manera muy fácil. Si ejecutas el Probador.java conseguirás una ejecución total del programa y obtendrás los resultados en la carpeta de ficheros.

Como parámetros de entrada tendremos el siguiente: ruta del archivo que queramos analizar, en nuestro caso, ficheros/diabetes.arff

Dentro de la ejecución total tenemos varias partes de ejecuciones. La primera nos vale para quitar atributos y instancias que no nos van a servir para la evaluación de nuestro método. Después haremos la inferencia y la evaluación del clasificador Ibk. Una vez que tengamos los datos del primer clasificador, haremos la evaluación de nuestroModelo.

Los resultados quedaran dentro de la carpeta ficheros. Si se quiere comparar los clasificadores solamente tendremos que fijarnos en los datos que hemos sacado en las evaluaciones.

Además tenemos otros métodos que ya hemos mencionado pero no les hemos dado importancia. Tenemos varias clases que utilizamos para la lectura y escritura de todo tipo de archivos, aunque en este caso solamente necesitemos lectura y escritura de archivos de texto.

3 Resultados

En esta sección mostraremos los resultados que hemos obtenido de las ejecuciones anteriores.

Resultados del clasificador Ibk:

```
*****************************************************

****Hold Out 70 30****

*****************************************************
F-Measure Batazbestekoa: 0.7560923349427191
K Maximoa: 25
Distance Weighting: 1
Revision: 10141
Training Times: 212
Nearest Neighbour Searh Algorithm: weka.core.neighboursearch.BallTree@4d591d15
K Maximoa metodo ez exhaustiboarekin: 15
Precision Batazbestekoa: 0.7621770668504911
Recall Batazbestekoa: 0.7652582159624414
ROC Area Batazbestekoa: 0.8079297245963913
F-Measure V1: 0.8275862068965517
F-Measure V2: 0.6323529411764706
```

```
Recall V1: 0.8888888888888888
Recall V2: 0.5512820512820513
Precision V1: 0.7741935483870968
Precision V2: 0.7413793103448276
Correctly Classified Instances: 76.52582159624413
=== Confusion Matrix ===

   a   b   <-- classified as
 120  15 |   a = tested_negative
  35  43 |   b = tested_positive

*******************************************************


Resultados del clasificador creado por nosotros:


*******************************************************

****Estimacion Nuestro modelo****

*******************************************************
K Maximoa: 38
Distance Weighting: NoDistance
Distance Type: Minkowski
Nearest Neighbour Searh Algorithm: 1
---------------------------------
    Precision = 38.61408684976535
    Recall = 50.64933378353808
    Accuracy = 0.5283013528046641
    F-Measure = 43.82036358733511
---------------------------------
------------Matriz De Confusin---------------------
---------------------------------
---|---------|-------------------
---|-TP=39---|-FP=62--------
---|---------|-------------------
---|-TN=73---|-FN=38--------
---|---------|-------------------
```

4 Conclusiones

Realizar un clasificador desde cero ha sido una tarea complicada a que nos hemos basado en el modelo del IBk, y le hemos ajustado el pseudocódigo que teníamos. Ha sido una tarea entretenida y muy dinámica pero muy costosa. Como conclusión tenemos que decir que no merece la pena crear un modelo propio habiendo ya el IBk.

5 Valoración Subjetiva

Una vez finalizado todo el proyecto, cada uno hemos hecho una valoración de todo lo que ha supuesto el proyecto individualmente.

Ieltzu: Personalmente ha sido una tarea que no me ha gustado mucho. He visto inecesaria la practica y para lo único que me ha servido es para darme cuenta de lo difícil que es implementar un clasificador propio.

Mikel: Bajo mi prespectíva, esta práctica, ha ayudado a comprender como funciona el algoritmo k-NN con los diferentes párametros pasados. Sabiendo que alcanzar la eficiencia del algoritmo ya implementado por weka es difícil (ya que este ha sido testeado y comprobado por numerosos investigadores) es un buen mtodo para aprender a utilizar herramientas que están bien testeadas ademas de comprender por dentro como funcionan dichas herramientas intentando implementarlas nosotros.

Maria:

References

https://es.wikipedia.org/wiki/Espacio-tiempo$_{d}e_{M}inkowski$