Práctica 1 Búsqueda Local APC
Metaheurísticas
Algoritmos KNN, Relief y Búsqueda Local

Ignacio Aguilera Martos
DNI: 77448262V e-mail: nacheteam@correo.ugr.es
Grupo de prácticas 1 Lunes 17:30-19:30

Curso 2017-2018

Índice

1. Introducción del problema

Para el problema de clasificación partimos de un conjunto de datos dado por una serie de tuplas que contienen los valores de atributos para cada instancia. Esto es una n-tupla de valores reales en nuestro caso.
El objetivo del problema es obtener un vector de pesos que asocia un valor en el intervalo $[0, 1]$ indicativo de la relevancia de ese atributo. Esta relevancia va referida a lo importante que es en nuestro algoritmo clasificador ese atributo a la hora de computar la distancia entre elementos.
Resumiendo lo que tenemos es un algoritmo clasificador que utiliza el vector de pesos calculado para predecir la clase a la que pertenece una instancia dada. Este algoritmo clasificador es el KNN con k=1. Lo que hace es calcular según la distancia euclídea (o cualquier otra) la tupla más cercana a la que queremos clasificar ponderando cada atributo con el correspondiente peso del vector, es decir, la distancia entre dos elementos sería:

$$d(e, f) = \sqrt{\sum_{i=0}^{n} w_i * (e_i - f_i)}$$

Donde e y f son instancias del conjunto de datos, w el vector de pesos y n la longitud de e y f que es la misma.
La calificación que se le asigna al vector w depende de dos cosas: la tasa de aciertos y la simplicidad.
La tasa de aciertos se mide contando el número de aciertos al emplear el clasificador descrito y la simplicidad se mide como el número de elementos del vector de pesos que son menores que 0.2, ya que estos pesos no son empleados por el clasificador, o lo que es lo mismo, son sustituidos por cero. Por lo tanto las calificaciones siguen las fórmulas:

$$Tasa\_acierto = 100 \cdot \frac{n^o aciertos}{n^o datos} \ , \ Tasa\_simplicidad = 100 \cdot \frac{n^o valoresdew < 0{,}2}{n^o deatributos}$$

$$Tasa\_agregada = \frac{1}{2} \cdot Tasa\_acierto + \frac{1}{2} \cdot Tasa\_simplicidad$$

Cabe destacar que todas las tasas están expresadas en porcentajes, por lo tanto cuanto más cercano sea el valor a 100 mejor es la calificación.
De esta forma a través del algoritmo que obtiene el vector de pesos para el conjunto de datos dado y el clasificador obtenemos un programa que clasifica de forma automática las nuevas instancias de datos que se introduzcan.

2. Introducción de la práctica

En esta práctica he analizado el comportamiento de los algoritmos KNN con k=1, el algoritmo greedy Relief y una implementación del algoritmo de búsqueda local para el problema de obtención de un vector de pesos para clasificar un conjunto de datos. Así mismo he realizado la implementación del algoritmo KNN con k variable para poder estudiar si varían los resultados al aumentar el valor de K o por contra obtenemos demasiado ajuste.
Para empezar al leer los ficheros de datos dados para el problema me he dado cuenta de que tenemos tuplas repetidas, cosa que he tenido en cuenta para no usarlas en la clasificación, ya que siempre obtendríamos distancia 0 para dicha tupla. Para ello en vez de comprobar el índice dentro del vector he comprobado si las tuplas son iguales para no usarlas.
Así mismo he implementado diferentes distancias a parte de la euclídea para comprobar si los resultados son mejores o peores en función de la distancia para cada conjunto de datos.
Para terminar, antes de analizar los datos, se debe considerar que los datos han sido redondeados a 4 decimales para no obtener tablas excesivamente largas. Si se desea obtener los datos completos se puede ejecutar el programa como se describe en la sección 7.

3. Descripción común a todos los algoritmos

Los algoritmos empleados han sido el KNN, el algoritmo greedy Relief y la metaheurística de búsqueda local.
Estos algoritmos comparten ciertos métodos y operadores que pasaré a explicar en esta sección.
Para empezar se debe destacar que la representación escogida para las soluciones es un vector de números reales, es decir, si n es el número de características:

$$w \in \mathbb{R}^n \ t.q. \ \forall i \ con \ 0 \leq i < n \ se \ tiene \ w_i \in [0,1]$$

O lo que es lo mismo, un vector de tamaño n con todas las posiciones rellenas con números del intervalo [0,1].
A estos números me referiré como pesos asociados a las características, ya que lo que nos indican es el grado de importancia de dicha característica a la hora de clasificar los datos, siendo 1 el máximo de relevancia y 0 el mínimo.
Así mismo cabe destacar que nuestra intención en este problema es obtener una buena calificación de dicho vector de pesos. Esto lo medimos mediante las tasas de acierto y simplicidad que se definen como:

$$Tasa\_acierto = 100 \cdot \frac{n^o aciertos}{n^o datos} \ , \ Tasa\_simplicidad = 100 \cdot \frac{n^o valoresdew < 0{,}2}{n^o deatributos}$$

$$Tasa\_agregada = \frac{1}{2} \cdot Tasa\_acierto + \frac{1}{2} \cdot Tasa\_simplicidad$$

La tasa de aciertos lo que nos mide es en un porcentaje cuántas instancias hemos clasificado correctamente mediante el algoritmo KNN usando el vector de pesos w.
La tasa de simplicidad nos mide cuántos de los valores que tiene el vector de pesos son menores que 0.2. Esto se hace ya que como imposición del problema tenemos que si alguna de los pesos es menor que 0.2 no debemos usarlo, o lo que es lo mismo, debemos sustituirlo por un 0 en la función de la distancia que luego describiré. Midiendo esto obtenemos un dato de cuanto sobreajuste ha tenido nuestro algoritmo a la hora de obtener el vector de pesos. Cuantas menos características necesitemos para discernir la clase a la que pertenece una instancia de los datos, más simple será clasificar dicha instancia. Se expresa en porcentaje indicando 0 como ninguna simplicidad y 100 como la máxima simplicidad.

De esta forma combinando ambas tasas obtenemos la tasa agregada que nos hace la media entre ambas tasas, de forma que le asignamos la misma importancia a acertar en la clasificación de las instancias y a la simplicidad en la solución. Cabe destacar que es imposible obtener una tasa de un 100 % a no ser que los datos se compongan únicamente de un punto ya que ello implicaría que la simplicidad ha de ser un 100 % (todos las posiciones del vector menores que 0.2) y por tanto la distancia sería 0 en todos los casos. De esta forma aspiraremos a una calificación lo mas alta posible pero teniendo en cuenta las restricciones de la función objetivo construida.
Las funciones y operadores de uso común los he agrupado en un fichero llamado auxiliar.py. Este fichero contiene las funciones de lectura de datos, distancias, una función que devuelve el elemento más común de una lista, la norma euclídea y una función para dividir los datos en el número de particiones que queramos manteniendo el porcentaje de elementos de cada clase que había en el conjunto de datos original.

3.1. Función de lectura de datos

La función de lectura de datos recibe la ruta del fichero arff y lee el contenido del mismo dando como resultado una lista de listas en la que cada una de ellas es una tupla o instancia de los datos.
El pseudocódigo de la función es:

Algorithm 1 lecturaDatos(nombre_fich)

```
data ← [ ]
for linea de nombre_fich do
  if se ha leído @data then
    data ← [data,linea]
  end if
end for
return data
```

Para esta implementación en concreto nos apoyamos en que Python tiene polimorfismo para todos los tipos de datos sin necesidad de declarar las variables, de forma que no nos importa que los datos sean numéricos o de tipo string.

3.2. Funciones de distancia

Las funciones de distancia siguen todas el mismo esquema de código, cambiando únicamente la fórmula empleada en cada caso para computar la distancia. Voy a describir las 3 distancias que he implementado teniendo esto en cuenta.
Se debe tener en cuenta que e1 y e2 son ambos dos tuplas del conjunto de datos de las que vamos a obtener la distancia y w es el vector de pesos que toma parte en el cómputo.

4. KNN

5. Relief

6. Búsqueda Local

7. Ejecución del programa y explicación

	Ozone				Parkinsons				Spectf-Heart			
	%_clas	%_red	Agr.	T (seg)	%_clas	%_red	Agr.	T (seg)	%_clas	%_red	Agr.	T (seg)
Partición 1	60.9375	0.0000	30.4688	0.1136	100.0000	0.0000	50.0000	0.0106	69.1176	0.0000	34.5588	0.0613
Partición 2	71.8750	0.0000	35.9375	0.1116	97.3684	0.0000	48.6842	0.0103	67.6470	0.0000	33.8236	0.0592
Partición 3	67.1875	0.0000	33.5938	0.1118	97.3684	0.0000	48.6842	0.0107	75.0000	0.0000	37.5000	0.0604
Partición 4	60.9375	0.0000	30.4688	0.1123	81.5789	0.0000	40.7895	0.0102	66.1765	0.0000	33.0882	0.0609
Partición 5	64.0625	0.0000	32.0313	0.1127	76.7442	0.0000	38.3721	0.0130	61.0390	0.0000	30.5195	0.0767
Media	65.0000	0.0000	32.5000	0.1122	90.6120	0.0000	45.3060	0.0144	67.7960	0.0000	33.8980	0.0836

Cuadro 1: Resultados 1NN

	Ozone				Parkinsons				Spectf-Heart			
	%_clas	%_red	Agr.	T (seg)	%_clas	%_red	Agr.	T (seg)	%_clas	%_red	Agr.	T (seg)
Partición 1	67.1875	0.0000	33.5938	0.1127	100.0000	0.0000	50.0000	0.0113	72.0588	0.0000	36.0294	0.0641
Partición 2	76.5625	0.0000	38.2813	0.1115	92.1053	0.0000	46.0526	0.0108	70.5882	0.0000	35.2941	0.0619
Partición 3	62.5000	0.0000	31.2500	0.1123	94.7368	0.0000	47.3684	0.0119	79.4118	0.0000	39.7059	0.0607
Partición 4	68.7500	0.0000	34.3750	0.1115	76.3158	0.0000	38.1579	0.0107	67.6471	0.0000	33.8235	0.0622
Partición 5	76.5625	0.0000	38.2813	0.1119	76.7442	0.0000	38.3721	0.0135	71.4286	0.0000	35.7243	0.0773
Media	70.3125	0.0000	35.1563	0.1120	87.9804	0.0000	43.9902	0.0116	72.2269	0.0000	36.1134	0.0653

Cuadro 2: Resultados 3NN

	Ozone				Parkinsons				Spectf-Heart			
	%_clas	%_red	Agr.	T (seg)	%_clas	%_red	Agr.	T (seg)	%_clas	%_red	Agr.	T (seg)
Partición 1	70.3125	0.0000	35.1563	0.1205	100.0000	0.0000	50.0000	0.0122	73.5294	0.0000	36.7647	0.0656
Partición 2	78.1250	0.0000	39.0625	0.1207	86.8421	0.0000	43.4211	0.0123	75.0000	0.0000	37.5000	0.0625
Partición 3	59.3750	0.0000	29.6875	0.1194	97.3684	0.0000	48.6842	0.0122	80.8824	0.0000	40.4412	0.0625
Partición 4	68.7500	0.0000	34.3750	0.1192	76.3158	0.0000	38.1579	0.0121	70.5882	0.0000	35.2941	0.0653
Partición 5	79.6875	0.0000	39.8438	0.1182	69.7674	0.0000	34.8837	0.0156	74.0260	0.00000	37.0130	0.0811
Media	71.2500	0.0000	35.6250	0.1196	86.0588	0.0000	43.0294	0.0129	74.8052	0.0000	37.4026	0.0674

Cuadro 3: Resultados 5NN

	Ozone				Parkinsons				Spectf-Heart			
	%_clas	%_red	Agr.	T (seg)	%_clas	%_red	Agr.	T (seg)	%_clas	%_red	Agr.	T (seg)
Partición 1	67.8431	94.5205	81.1818	0.0612	82.8025	86.9565	84.8795	0.0078	81.6143	48.8889	65.2516	0.0439
Partición 2	68.7500	63.0137	65.8818	0.0642	81.5287	91.3043	86.4165	0.0081	90.4580	80.0000	85.2290	0.0435
Partición 3	68.5039	82.1918	75.3479	0.0613	83.4395	86.9565	85.1980	0.0078	93.5018	51.1111	72.3065	0.0439
Partición 4	61.7188	94.5205	78.1196	0.0657	82.8025	95.6522	89.2274	0.0079	89.0152	62.2222	75.6187	0.0436
Partición 5	57.4219	95.8904	76.6561	0.0638	76.3158	86.9565	81.6362	0.0104	80.1932	42.2222	61.2077	0.0557
Media	64.8475	86.0274	75.4375	0.0607	81.3778	89.5652	85.4715	0.0101	86.9565	56.8889	71.9227	0.0568

Cuadro 4: Resultados Relief con K=1

	Ozone				Parkinsons				Spectf-Heart			
	%_clas	%_red	Agr.	T (seg)	%_clas	%_red	Agr.	T (seg)	%_clas	%_red	Agr.	T (seg)
Partición 1	65.8824	94.5205	80.2015	0.0607	84.7134	86.9565	85.8349	0.0078	69.9552	48.8889	59.4220	0.0457
Partición 2	75.3906	63.0137	69.2022	0.0613	80.2548	91.3043	85.7796	0.0078	73.2824	80.0000	76.6412	0.0437
Partición 3	74.0157	82.1918	78.1038	0.0626	82.8025	86.9565	84.8795	0.0086	76.1733	51.1111	63.6422	0.0453
Partición 4	57.0313	94.5205	75.7759	0.0618	83.4395	95.6522	89.5458	0.0078	81.4394	62.2222	71.8308	0.0436
Partición 5	63.6719	95.8904	79.7811	0.0599	85.5263	86.9565	86.2414	0.0099	73.9130	42.2222	58.0676	0.0571
Media	67.1984	86.0274	76.6129	0.0613	83.3473	89.5652	86.4563	0.0084	74.9527	56.8889	65.9208	0.0471

Cuadro 5: Resultados Relief con K=3

	Ozone				Parkinsons				Spectf-Heart			
	%_clas	%_red	Agr.	T (seg)	%_clas	%_red	Agr.	T (seg)	%_clas	%_red	Agr.	T (seg)
Partición 1	65.0980	94.5205	78.8093	0.0637	80.8917	86.9565	83.9241	0.0086	68.6099	48.8889	58.7494	0.0443
Partición 2	70.3125	63.0137	66.6631	0.0614	79.6178	91.3043	85.4611	0.0086	74.4275	80.0000	77.2137	0.0470
Partición 3	72.4409	82.1918	77.3164	0.0631	82.1656	86.9565	84.5611	0.0081	83.0325	51.1111	67.0718	0.0510
Partición 4	56.6406	94.5205	75.5806	0.0625	83.4395	95.6522	89.5458	0.0088	81.4394	62.2222	71.8308	0.0466
Partición 5	66.0156	95.8904	80.9530	0.0649	86.1842	86.9565	86.5704	0.0108	76.3285	42.2222	59.2754	0.0614
Media	66.1015	86.0274	76.0645	0.0631	82.4598	89.5652	86.0125	0.0090	76.7675	56.8889	66.8282	0.0501

Cuadro 6: Resultados Relief con K=5

	Ozone				Parkinsons				Spectf-Heart			
	%_clas	%_red	Agr.	T (seg)	%_clas	%_red	Agr.	T (seg)	%_clas	%_red	Agr.	T (seg)
Partición 1	74.1176	53.4247	63.7712	34.5843	82.1656	43.4783	62.8219	0.5610	81.1659	35.5556	58.3607	5.3001
Partición 2	67.1875	35.6164	51.4020	6.4653	82.8025	43.4783	63.1404	0.5545	87.4046	31.1111	59.2579	3.5880
Partición 3	66.5354	53.4247	59.9800	13.2504	92.3567	39.1304	65.7436	0.6979	93.8629	40.0000	66.9314	4.7296
Partición 4	71.0938	47.9452	59.5195	23.0546	89.1720	21.7391	55.4556	0.2889	87.5000	35.5556	61.5278	9.3608
Partición 5	69.1406	43.8356	56.4881	10.8910	92.1053	43.4783	67.7917	0.3560	76.8116	51.1111	63.9613	11.3526
Media	69.6150	46.8493	58.2321	17.1638	87.7204	38.2609	62.9906	0.5722	85.3490	38.6667	62.0078	7.6077

Cuadro 7: Resultados Búsqueda Local con K=1

	Ozone				Parkinsons				Spectf-Heart			
	%_clas	%_red	Agr.	T (seg)	%_clas	%_red	Agr.	T (seg)	%_clas	%_red	Agr.	T (seg)
Partición 1	69.8039	43.8356	56.8198	12.9352	88.5350	39.1304	63.8327	0.3989	75.7848	28.8889	52.3368	4.6527
Partición 2	69.5313	50.6849	60.1081	18.8173	81.5287	26.0869	53.8078	0.2219	78.2443	44.4444	61.3444	4.9285
Partición 3	70.4724	43.8356	57.1540	13.0915	82.1656	43.4783	62.8219	1.0002	71.4801	53.3333	62.4067	7.6825
Partición 4	67.1875	41.0959	54.1417	17.3212	89.1720	34.7826	61.9773	0.5576	74.6212	37.7778	56.1995	7.1972
Partición 5	67.5781	46.5753	57.0767	22.6021	87.5000	47.8261	67.6630	0.7952	77.7778	37.7778	57.7778	9.2464
Media	68.9146	45.2055	57.0601	16.9535	85.7803	38.2609	62.0206	0.5948	75.5816	40.4444	58.0130	6.7415

Cuadro 8: Resultados Búsqueda Local con K=3

	Ozone				Parkinsons				Spectf-Heart			
	%_clas	%_red	Agr.	T (seg)	%_clas	%_red	Agr.	T (seg)	%_clas	%_red	Agr.	T (seg)
Partición 1	66.6667	50.6849	58.6758	16.2974	83.4395	34.7826	59.1110	0.8260	74.8879	40.0000	57.4439	11.5613
Partición 2	64.0625	32.8767	48.4696	9.4039	83.4395	21.7391	52.5893	0.3869	74.4275	40.0000	57.2137	9.4207
Partición 3	74.8031	57.5342	66.1687	24.0036	84.7134	47.8261	66.2697	0.8540	83.7545	44.4444	64.0995	9.1497
Partición 4	76.1719	60.2740	68.2229	34.7944	88.5350	52.1740	70.3545	0.7861	78.0303	57.7778	67.9040	9.0839
Partición 5	68.7500	47.9452	58.3476	20.8456	96.0526	34.7826	65.4176	0.5576	78.2609	37.7778	58.0193	4.626
Media	70.0908	49.8630	59.9769	21.0690	87.2360	38.2609	62.7484	0.6821	77.8722	44.0000	60.9361	8.8356

Cuadro 9: Resultados Búsqueda Local con K=5