Aprendizaje Automático

GRADO EN INGENIERÍA INFORMÁTICA

Tutorial 2

Curso 2020/2021

Jorge Rodríguez Fraile, 100405951, Grupo 83, 100405951@alumnos.uc3m.es
Carlos Rubio Olivares, 100405834, Grupo 83, 100405834@alumnos.uc3m.es

Índice

Ejercicio 1

Pregunta 1

Dos atributos, compuestos por el nombre de la persona, que es un nominal y su clase que puede tomar los valores + o -, que es nominal.

Pregunta 2

No, ya que solo disponemos del nombre para poder saber si será de tipo positivo o negativo, y este atributo es bastante abstracto, cada persona tiene un nombre propio y sería imposible relacionarlo con la clase.

Ejercicio 2

Pregunta 1

Dentro de las posibilidades, la ejecución es satisfactoria ya que se ha podido otorgar una clase al 100% de las instancias, aunque hay que decir que la probabilidad de que fallara es nula ya que cada nombre es único.

Ejercicio 3

Pregunta 1

Los atributos que podríamos extraer de name son:

- Número de vocales: Veces aparece que aparecen vocales en name para cada nombre.
- Número de consonantes: Veces que aparecen consonantes en cada nombre.
- Inicial del nombre: Carácter con el que empieza el nombre del asistente.
- Longitud del nombre: Número de vocales de caracteres que contiene el nombre.
- Longitud del apellido: Número de caracteres que contiene el apellido.
- Caracteres especiales: Vale ‘0’ si el nombre completo tiene algún punto, guion…En caso contrario se almacenará un ‘1’.

Hemos elegido estos atributos ya que pensamos que son la mejor forma de diferenciar a cada persona teniendo en cuenta solo su nombre, ya que siendo un dominio tan cerrado debemos ser muy específicos con lo que deseamos usar para resolver el problema.

Pregunta 2

Hay 9 atributos que se listan a continuación con su tipo:

- Name: de tipo nominal, nos dice el nombre de la instancia.
- Length: de tipo numeric, nos dice el número de caracteres del nombre.
- Even_odd: de tipo nominal, nos dice si la longitud del nombre es par o impar.
- Frist_char_lowel: de tipo nominal, nos dice si el nombre empieza por vocal (1) o no (0).
- Consonants: de tipo numeric, nos dice el nº de consonantes que se encuentran en el nombre.
- Spaces: de tipo numeric, nos dice el número de espacios que hay.
- Dot: de tipo numeric, nos dice el número de puntos que hay en el nombre.

- Words: de tipo numeric, nos dice el número de palabras que se encuentran en una instancia.
- Class: de tipo nominal, toma los valores + o -.

Pregunta 3

Para los atributos de tipo nominal, aparece para cada valor el número de veces que aparece cada uno y qué peso tiene cada uno, en este caso cada repetición tiene valor de 1.

En los de tipo numeric nos indica cual es el valor mínimo(el valor más pequeño), máximo(el más grande), media y desviación típica entre todas las instancias para ese atributo.

En visualize all, se muestra la distribución de los valores gráficamente en una recta con respecto a una clase que podemos elegir antes de pulsar el botón, en la que la altura indica el número de repeticiones de ese valor para esa clase.

Pregunta 4

Lo que ocurre es que no nos deja generar el clasificador con ID3, dado que ID3 sólo puede generarse si se utilizan atributos de tipo nominal, y en este caso, hay diferentes atributos de tipo numeric por lo que no podemos utilizar este algoritmo.

Ejercicio 4

Pregunta 1

Este filtro divide los datos en diferentes rangos, en este caso, en cinco, que ha venido dado por el atributo ‘bins’ del filtro.

Pregunta 2

Este clasificador tiene un total de 236 instancias acertadas de 294, lo que deja un porcentaje de 80,2721%.

Pregunta 3

En la matriz de confusión, cada columna muestra el número de predicciones de cada clase, mientras que cada fila representa a las instancias con su clase verdadera. Muestra los positivos verdaderos, falso positivos, falso negativo y negativo verdadero.

Pregunta 4

Se han clasificado mal 51 instancias que deberían haber ido a - y se han metido en +, y 7 instancias que deberían haber ido a + han ido a -. Estos datos los extraemos de la matriz de confusión.

Pregunta 5

Con esta opción nos aparece una tabla adicional con la siguiente composición:

inst#	actual	predicted	error	prediction
1	1:-	1:-		1
2	1:-	1:-		1
3	2:+	2:+		0.667
4	1:-	1:-		1
5	1:-	1:-		1
6	2:+	2:+		1
7	1:-	2:+	+	0.571

En esta tabla se muestra la instancia predicha, el valor real que tiene y lo que se ha predicho, además del error de predicción que se ha podido generar. La primera instancia que no se resuelve de manera correcta es la 7 ('Andrey Burago',13,0,1,7,1,0,2,-), ya que la clase que tiene es un - y se ha predicho un +.

Pregunta 6

Donald Trump (borrando el atributo name) sería clasificado de la siguiente manera:

- Length: 12
- Even_odd: 0
- Frist_char_lowel: 0
- Consonants: 8
- Spaces: 1
- Dots: 0
- Words: 2
- Class: +

Con la transformación del filtro, los atributos quedan:

- Length: (10.6-14.2]
- Even_odd: 0
- Frist_char_lowel: 0
- Consonants: (5.6-8.2]
- Spaces: (-inf-1.4]
- Dots: (-inf-0.4]
- Words: (-inf-2.4]
- Class: +

Al introducirlo en el código del archivo, se clasifica de manera correcta, dando el valor + a la clase con nombre ‘Donald Trump’.

Ejercicio 5

Pregunta 1

Predice que el valor de las instancias será siempre el positivo, puesto que es el más común en nuestro conjunto de ejemplos de entrenamiento.

Pregunta 2

Este modelo acierta el 51.0294% de las instancias, que son 150 de las 294 instancias. El porcentaje es mayor que 50% porque coge la clase más común de los ejemplos.

Pregunta 3

Lo clasifica como positivo, puesto que siempre asigna ese valor a los ejemplos, independientemente del valor que tomase una concreta, domina la más común.

Ejercicio 6

Pregunta 1

Genera 19 nodos terminales en el árbol, que se especifica en la salida.

Pregunta 2

Se clasifican correctamente 287 instancias de un total de 294.

Pregunta 3

Se clasifica correctamente un 97.619% de las instancias.

Pregunta 4

De un total de 7 instancias mal clasificadas, a 4 se les ha asignado - y a 3 + incorrectamente.

Pregunta 5

Se clasificaría como +, se puede comprobar tras introducir los datos en el documento de ejemplos, le asigna el + independientemente de si lo introducimos con + o -.

Pregunta 6

Elegiría este modelo, puesto que proporciona una tasa de aciertos más alta, el $\simeq$98% contra $\simeq$80%. Es capaz de clasificar con mayor precisión los datos.

Pregunta 7

No, puesto que no es etiquetar con una precisión del 100% o muy cercana, el 98% es buen resultado, pero no lo suficiente para ser la función exacta.

Ejercicio 7

Pregunta 1

El número de vocales se calcularía como: length-spaces-dots-consonants que es a2-a6-a7-a5.

Pregunta 3

En estadísticas se puede ver que el rango más común es (3.455, 4.182)

Pregunta 5

El porcentaje de instancias clasificadas correctamente nos indica que es del 100%, y la matriz de confusión solo contiene Positivos verdaderos y Negativos verdaderos (144 0 0 150).

Pregunta 6

En las hojas pone +(150.0) y - (144.0) que son el número de instancias que se han etiquetado con cada uno de los valores de la clase. En la rama vemos que los que los seleccione es si tienen más de 4 (-) o menor o igual a 4 (+)

Pregunta 9

Al aumentar Jitter los valores se dispersan alrededor de los valores enteros de número de vocales, dejando una nube de puntos, en vez de estar concentrados en un punto.

Pregunta 10

La característica fundamental que debe poseer los atributos es que sean capaces de dividir el número de instancias de manera uniforme con respecto a los valores que tiene la clase, así se minimiza la complejidad o número de decisiones que se deben tomar para etiquetarlos.

Ejercicio 8

Pregunta 2

Este fichero tiene un total de 15 atributos y 32561 instancias.

Pregunta 3

Al usar cross validation se reparten las instancias en n grupos (en este caso, han sido 10), uno de estos grupos se elige aleatoriamente para que sea el set de validación, mientras que los n-1 restantes serán los sets de entrenamiento, esto se repite para todos los grupos y finalmente se obtiene el modelo.

Pregunta 4

Nos deja un porcentaje de 86.2105% de instancias correctamente clasificadas, lo que deja a 4490 instancias mal clasificadas.

En este caso, el acierto de predicción es algo mayor, en torno a un 1%, aunque el número de instancias usadas es el mismo. El tamaño del árbol y de los nodos terminales es el mismo. También cabe recalcar que al ser el acierto de predicción algo diferente, esto también afecta a la matriz de confusión.

Pregunta 5

En el atributo de salida, que en este caso es el salario hay dos clases, >50K y <=50K, y están distribuidos de tal forma que la primera clase tiene 7841 instancias y la segunda 24720.

La distribución de las clases está bastante desequilibrada, y, aunque no sería totalmente eficiente, ya que no tiene la misma cantidad de ejemplos de una clase que de otra, se podría usar un algoritmo de aprendizaje automático.

Pregunta 6

Ahora el atributo de salida está perfectamente equilibrado, con 16280 instancias para cada clase. El número de instancias ha bajado, pero solo en una unidad, ahora su valor es 32560

Pregunta 7

Los resultados han mejorado de manera significativa, en el caso del cross validation tenemos un 87.1898 % de acierto de predicción, mientras que con el método de *supplied test set* tenemos un 92.7242 %. De nuevo, se puede observar que los resultados del segundo método son algo mejores.

Pregunta 8

Este filtro hace que los datos de tipo numeric se representen en una escala de 0 a 1, siendo el 0 el valor más bajo y el 1 el más alto.

Pregunta 9

Podría ayudar ya que ahora el rango de los datos está bastante más reducido y se podrían clasificar los datos de manera más sencilla.

Pregunta 10

La aplicación del filtro normalize empeora los resultados anteriores,87.1898% vs. 87.1867% y 92.7242% vs. 90.8722% , por lo que la mejor opción sigue siendo utilizar el filtro de Resample y el método *supplied test set.*

Conclusiones y dificultades encontradas

Este tutorial nos ha ayudado a entender la herramienta de Weka y cómo funcionan algunos algoritmos como ID3 y J48. Por otro lado, también hemos podido refrescar nuestra memoria y volver a repasar algunos temas estadísticos como la normalización. En definitiva, este tutorial nos ha parecido más relajado que el anterior y hemos tenido menos problemas realizándolo.