Práctica 9
Inteligencia Artificial

Autores (Grupo 5):
José Javier Cortés Tejada
Pedro David González Vázquez

Apartado 1

Descripción del dataset

Para esta práctica hemos cogido de internet un dataset sobre el cáncer de mama, el cual pretende determinar si una persona que ha sufrido esta enfermedad vuelva a padecerla. Para ello, utilizaremos 10 campos con información sobre el individuo, los cuales son:

- Edad (*age*): edad del individuo, comprendida de forma discreta en intevalos de 10 años, empezamdo en 10 y acabando en 100.
- Menopausia (*menopause*): indica el estado de la menopausia del paciente en el momento del diagnóstico.
- Tamaño del tumor (*tumor-size*): tamaño del tumor en milímetros (mm).
- Numero de linfomas (*inv-nodes*): cantidad de linfomas encontrados en la examinación.
- Tumor extendido (*node-caps*): indica si se ha roto la membrana del nodo afectado, afectando a otros.
- Grado de malignidad (*deg-malig*): indica el grado en que un tumor es malo para el individuo, de 1 a 3.
- Pecho (*breast*): indica el pecho afectado por la enfermedad, derecho o izquierdo.
- Cuadrante del pecho (*breast-quad*): cuadrante del pecho afectado.
- Quimioterapia (*Irradiat*): indica si el individuo ha sido sometido a sesiones de quimio.
- Clase (*Class*): indica si hay reaparición de los síntomas.

Para este ejemplo tenemos 286 instancias, las cuales vamos a clasificar en función de la recurrencia de los sintomas del cáncer, trabajando sobre los datos conocidos.

J48 con *training set*

La ejecución nos deja los siguientes resultados:

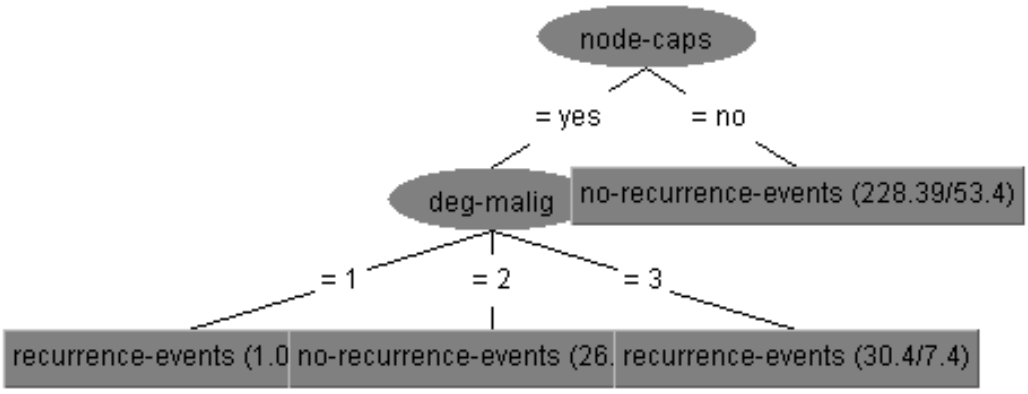
Correctly Classified Instances 217− 75,8741 %
Incorrectly Classified Instances 69− 24,1259 %
Kappa statistic 0,2899
Mean absolute error 0,3658
Root mean squared error 0,4269
Relative absolute error 87,4491 %
Root relative squared error 93,4017 %
Total Number of Instances 286

Tasa TP	Tasa FP	Precisión	Recall	Clase
96, 5 %	72, 9 %	75, 8 %	96, 5 %	Evento no recurrente
27, 1 %	3, 5 %	76, 7 %	27, 1 %	Evento recurrente

Matriz de confusión:

Evento no recurrente	Evento recurrente
194	7
62	23

Árbol de decisión:



De acuerdo a los datos obtenidos en el sumario y a la matriz de confusión, tenemos un total de 69 instancias mal clasificadas (hemos contado tanto *FP* como *FN*).Respecto del árbol, tenemos 4 nodos terminales (hojas del árbol) y el primer nivel del árbol de clasificación se determina con el atributo *node-caps*, dado que es el que mayor entropía tiene respecto del resto de opciones.

Comparación de los métodos de validación

	J48 *traning set*	J48 *cross-validation*	J48 *percentaje-split*
Instancias bien clasificadas	$75,87\,\%$	$75,52\,\%$	$68\,\%$
Precisión	$76\,\%$	$75\,\%$	$66\,\%$
Recall	$76\,\%$	$75\,\%$	$68\,\%$

Fijándonos en la forma de clasificar, el *traning-set* es que que nos proporciona mejores resultados en cuanto a la clasificación (algo mejores que los de *cross-validation*), pues los valores que nos proporcionan son mayores en todos los apartados. En cuanto a *percentaje-split*, los resultados obtenidos son bastante peores en comparación con el resto, debido a que solo se ha trabajado con un tamaño menor de la muestra.

A su vez, observamos que *traning set* y *cross-validation* son bastantes cercanos entre ellos y por tanto son realistas ambos.

De cara a *percentaje-split*, tenemos una tasa del $69,7\,\%$ en FP, mientras que en TP tenemos una tasa del $30,3\,\%$ (son complementarias).

Apartado 2

En este caso, el problema asociado es clasificar a los individuos del dataset en dos clases distintas, presencia de eventos recurrentes y no recurrentes. Para la simulación hemos aplicado el clustering jerárquico propuesto en la práctica, y los resultados obtenidos son los siguientes:

Cluster 0 (no recurrente)	Cluster 1 (recurrente)	
201	0	eventos no recurrentes
84	1	eventos recurrentes

Podemos comprobar que el dataset que hemos empleado no es adecuado para *clustering* dado que hemos obtenido casi un 30 % de instancias mal agrupadas; de dicho porcentaje, el total de ellas son los 84 eventos recurrentes que han sido asignados al cluster 0. Este error no sería tan significativo si los errores estuviesen repartidos entre las dos clases, pero dado que todos están agrupados en la misma clase podemos concluir que aún con un mayor volumen de eventos recurrentes la tasa de error aumentará.

Dado que en el dataset empleado solo tenemos dos clases, es absurdo repetir el clustering con un tercio de las mismas, pues solo tendríamos un cluster y todo iría a él, luego no hemos realizado esta parte.

Referencias

El dataset que hemos usado para el desarrollo de la práctica lo hemos cogido del siguiente sitio http://repository.seasr.org/Datasets/UCI/arff/.