Actividad 2.1: Ejercicio de simulación

El estudiante generará un conjunto de datos artificial compuesto por 100 instancias caracterizadas por una variable relevante en sentido fuerte, tres variables relevantes en sentido débil y una variable totalmente irrelevante. Esta última se puede generar mediante números aleatorios extraídos de una distribución de probabilidad uniforme o normal (gaussiana). Como indicación sugerimos extender el ejemplo XOR a tres dimensiones (véase el artículo Kohavi, R. & John, G.H., Wrappers for Feature Subset Selection (1997) para obtener indicaciones sobre las definiciones de relevancia y sobre el problema XOR). A continuación, aplicará diferentes técnicas de selección de variables disponibles en weka (un mínimo de tres de filtrado, el análisis de componentes principales y la técnica de envoltura, WrapperSubsetEval, con BayesNet como clasificador y empleando todos los valores por defecto, salvo el número máximo de padres que se debe modificar a 3).

1. Introducción

Para la realización de esta práctica me han sido de mucha ayuda las aportaciones del resto de compañeros en los foros. Retomo los estudios universitarios después de bastantes años sin contacto con la Universidad y supongo que necesito un poco de rodaje y soltura, que espero que pronto llegarán.

2. Descripción del experimento

Tal como se ha sugerido, y siguiendo el artículo recomendado como lectura para la tarea, lo que he hecho es ampliar el ejemplo XOR propuesto. De este modo, he generado un conjunto de datos “aleatorios” formado por 5 variables (X1, X2, X3, X4, X5) y cuya clasificación recae en una sexta variable Y. Todas ellas son de tipo boolean, si bien, en principio entiendo que la variable X5 podría haber sido de cualquier tipo al ser irrelevante como paso a explicar.

En los datos de ejemplo que he construido tenemos las siguientes relaciones definidas a

priori:

- X1, X2, X3 y X5 son **variables** a las que les asigno valores boolean de forma **aleatoria**
- X4: la construyo a partir de X2 y X3 con un xor. **X4 = X2 xor X3**
- Y: es el valor de la clase y lo calculo como **Y = X1 xor X2 xor X3.**

Por lo tanto, a priori, y de cara a los experimentos que expondré mas tarde relativos a la utilización de Weka para selección de variables, podemos observar que:

1. X1 es determinante en la predicción del valor de Y y la información que nos aporta no puede ser sustituida por ninguna otra variable o conjunto de ellas. Pendiente de demostración, parece que podemos decir que **X1 es la variable relevante en sentido fuerte.**
2. Las variables X2, X3 y X4 están relacionadas, es decir, son dependientes pues el valor de cualquiera de ellas puede obtenerse conociendo el valor de las otras dos. Ahora bien, teniendo x1, para saber el valor de y nos basta con conocer uno de los pares siguientes: (x2, x3), (x2, x4) o (x3, x4). Es decir, la información contenida por estas tres variables en relación a Y la podemos encontrar en un par cualquiera de ellas. Pendiente de demostración, podemos ir diciendo tambień que las variables **X2, X3 y X4 serán relevantes en sentido débil.** Es decir, una a una son prescindibles y dependiendo del subconjunto de ellas que seleccionemos, no lo son.
3. La variable **X5** no aporta nada en el conocimiento del valor de Y. Es decir, es claramente **irrelevante** en todos los sentidos.

Al ser las anteriores varibles booleanas y existir las dependencias descritas entre ellas, resulta que las únicas combinaciones posibles son las siguientes (nótese que ignoro por completo la variable X5 en estas apreciaciones):

X1	X2	X3	X4	X5	Y
0	0	0	0	1	0
0	0	1	1	1	1
0	1	0	1	0	1
0	1	1	0	0	0
1	0	0	0	0	1
1	0	1	1	0	0
1	1	0	1	0	0
1	1	1	0	1	1

Como lo que se pide en la práctica son 100 instancias, lo que he hecho para mi conjunto de prueba es repetir la tabla anterior tantas veces como ha sido necesario hasta completar 100 filas. Seguramente sea de agradecer que no adjunte aquí el dataset completo.

3. Demostración probabilística de la relevancia/irrelevancia