Práctica 7. Aprendizaje con GeNIe.

Como ya hemos comentado en clase, en el caso de que se disponga de bases de datos es posible aprender modelos bayesianos a partir de ellos. Vamos a ver cómo se aprenden modelos basados en redes bayesianas y modelos Naive Bayes, cómo se validan, y cómo podemos utilizar GeNIe para generar bases de datos de ejemplos. A lo largo del tutorial se irán planteando preguntas. Para la entrega, escribe las respuestas a dichas preguntas en un fichero (al final de la práctica tendrás que subir dicho fichero, en pdf, como resultado).

TUTORIAL

1. Aprendizaje de redes bayesianas en GeNIe.

Los algoritmos de aprendizaje de redes bayesianas se dividen en dos tipos:

- Algoritmos de aprendizaje estructural: cuyo objetivo es aprender a partir de una base de datos la estructura de la red.
- Algoritmos de aprendizaje de parámetros: cuyo objetivo es aprender, a partir de una base de datos y una estructura dada, los parámetros necesarios. Estos son los algoritmos que hemos visto en clase y vienen descritos en detalle en los apuntes.

Vamos a ver cómo realizar estos dos tipos de aprendizaje en la herramienta GeNIe.

1.1. Aprendizaje estructural

Comenzamos con una base de datos. Vamos a utilizar una de las que viene como ejemplo en GeNIe. Para ello, abre (utilizando la opción "Open data file" del menú File) el fichero de ejemplo retention.txt. En este fichero es donde están los datos. Al abrirlo con GeNIe, el aspecto es este:

	spend	apret	top10	rejr	tstsc	pacc	strat	salar
▶	9855	52.5	15	29.474	65.063	36.887	12	60800
	10527	64.25	36	22.309	71.063	30.97	12.8	63900
	7904	37.75	26	25.853	60.75	41.985	20.3	57800
	6601	57	23	11.296	67.188	40.289	17	51200
	7251	62	17	22.635	56.25	46.78	18.1	48000
	6967	66.75	40	9.718	65.625	53.103	18	57700
	8489	70.333	20	15.444	59.875	50.46	13.5	44000

Este ejemplo procede de un estudio[1] cuyo objetivo era determinar las causas del abandono de las carreras univesitarias en la universidad americana. Las ocho variables fueron seleccionadas de las más de 200 publicadas anualmente por las revistas US News and World Report Magazine. Las variables elegidas fueron:

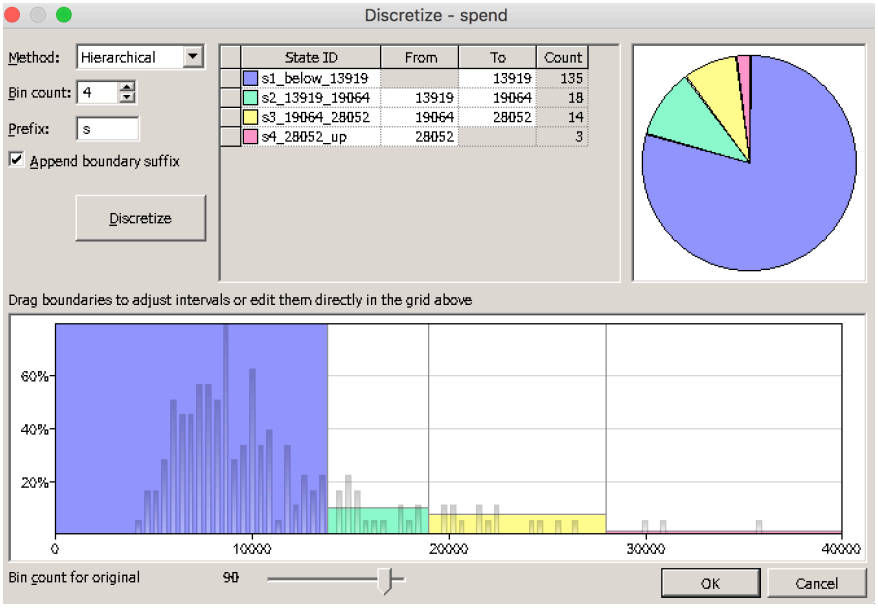
- **Apret** - porcentaje medio de retención, es decir, estudiantes que no abandonan su carrera
- **Reject –** porcentaje de estudiantes que rechazan la oferta de la universidad
- **Tstsc** - nota media obtenida por los estudiantes en los exámenes de acceso
- **Top 10** - porcentaje de estudiantes de nuevo ingreso que estaban entre los diez mejores de su promoción en el instituto)

[1] Marek J. Druzdzel and Clark Glymour. What do college ranking data tell us about student retention: Causal discovery in action. In Proceedings of the Fourth International Workshop on Intelligent Information Systems (WIS-95), pages 138-147, Augustow, Poland, June 5-9, 1995.

- **Pacc** - porcentaje de estudiantes que aceptan la oferta de la universidad
- **Spend** - gasto total por estudiante
- **Strat** - Ratio estudiantes por docente
- **Salar** - salario medio del profesorado

La mayoría de los algoritmos de aprendizaje estructural trabajan sobre variables discretas. Por este motivo, lo que vamos a hacer es discretizar los datos. Para ello, en el menú *Data* pulsamos la opción *Discretize*.

Nos mostrará la pantalla siguiente, en la que en la casilla "Bin count" podemos seleccionar el número de clases queremos (en este ejemplo hemos puesto 4), así como elegir entre varios métodos para discretizar las variables (jerárquico, anchura uniforme, nº ocurrencias uniforme). El prefijo que introduzcamos se utilizará para ponerle nombre a los intervalos usados para la discretización (GeNIe crea dichos intervalos automáticamente).



Una vez discretizados todos los datos (para cada columna se pueden utilizar parámetros diferentes), podemos utilizar los diferentes algoritmos para aprender la estructura.

Vamos ahora a abrir un fichero que ya tiene los datos discretizados (el fichero retention_discretized.txt) y vamos a aprender la estructura de la red.

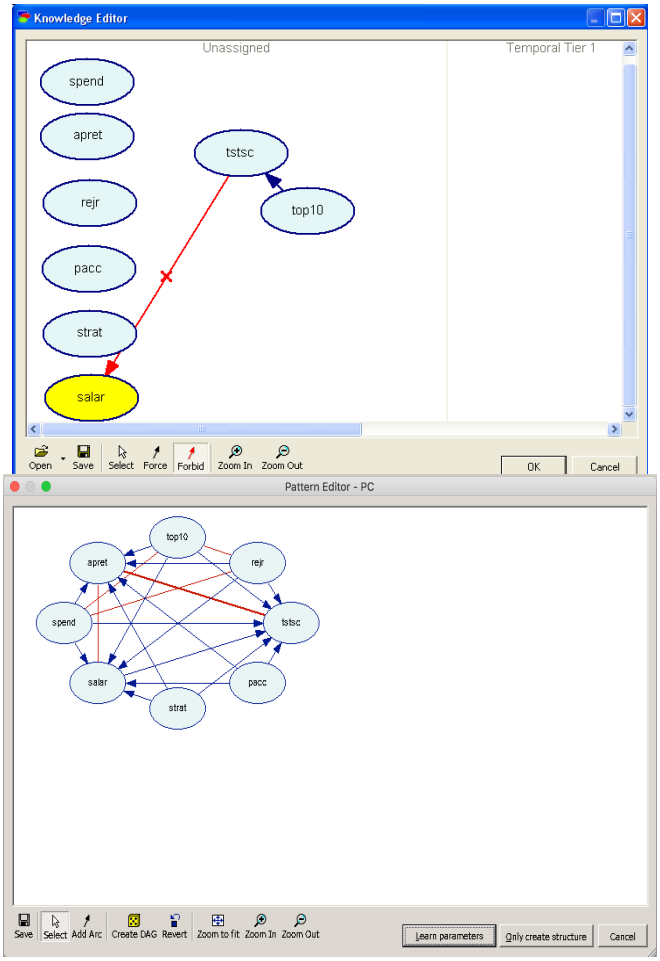
Para ello, nos vamos a "Data" y pulsamos en "Learn new network". Nos aparecerá una pantalla en la que podremos configurar los parámetros.

Podemos marcar las variables que queremos que aprenda (por defecto están marcadas todas) y, si tenemos algún tipo de conocimiento previo sobre el dominio, podemos indicarlo (tanto forzar relaciones que sabemos que existen, como prohibir relaciones que estamos seguros de que no existen). Para ello pulsamos "Background knowledge" y aparece la pantalla que se muestra en la figura inferior a la izquierda.

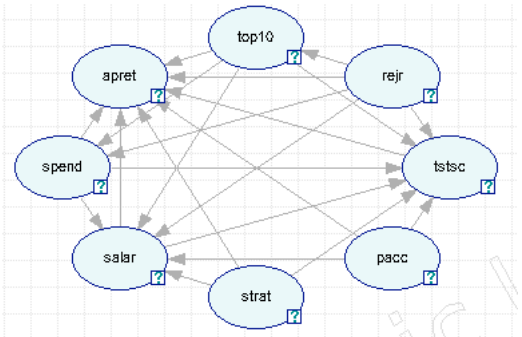
Podemos por ejemplo forzar que se cree un enlace entre topten y tstsc (es decir, indicar que pensamos que la variable "número de estudiantes que estaban en el 10% mejor de su clase" tiene influencia causal en la variable "nota media de acceso" , y prohibir que se cree un enlace entre tstsc y salr (es decir, indicar que pensamos que la variable "nota media de acceso" no tiene relación de influencia causal con la variable "salario".

Una vez introducido el conocimiento previo, pulsamos OK, elegimos el algoritmo PC, y volvemos a pulsar OK. Aparecerá una nueva ventana con los enlaces que haya creado GeNIe.

Vemos que, en algunos casos, ha creado el arco, pero no la dirección (los datos no han dado la evidencia suficiente para ello). Dichos enlaces podemos editarlos y elegir la dirección que pensamos que refleja mejor la relación de causalidad. De este modo, el proceso de creación del grafo es interactivo. Si no queremos decidir nosotros, podemos pulsar el botón DAG (o el botón randomize, en la versión anterior de GeNIe) para que cree un grafo acíclico dirigido (la dirección de los enlaces la determinará de modo que no se creen ciclos, pero si hay más de una posibilidad para ello, la elegirá aleatoriamente).



Ahora podemos elegir entre crear la estructura y los parámetros, o sólo la estructura. Vamos a crear sólo la estructura (pulsando la opción Only Create Structure), y obtendremos:

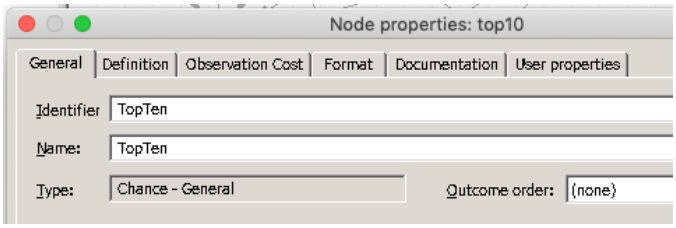


Los enlaces salen de color gris, pues las probabilidades están distribuidas de forma uniforme. Por tanto, esta red aún no vale para razonar, hasta que aprendamos los parámetros. Guardamos el fichero con el nombre PCretention (lo usaremos a continuación para realizar el aprendizaje de parámetros).

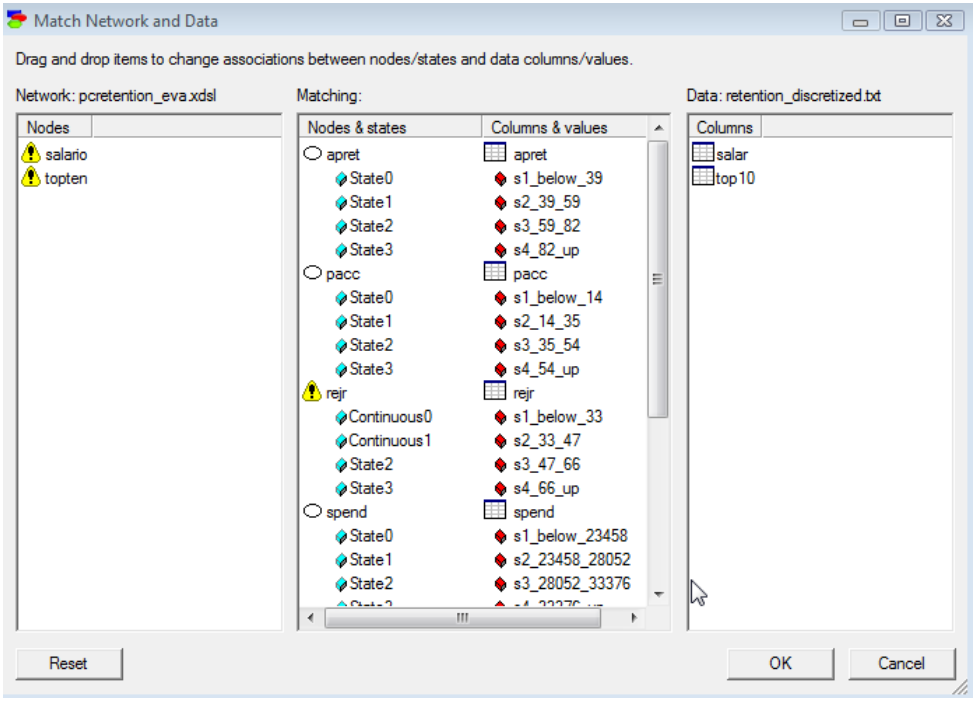
1.2. Aprendizaje paramétrico

En este caso se trabaja con un fichero de datos y una estructura, y el objetivo es aprender los parámetros necesarios. Supongamos que hemos aprendido ya o determinado con ayuda de expertos una estructura para la red, por ejemplo, la que aprendimos con el algoritmo PC (abrimos el fichero PCretention.xsdl).

Para ilustrar mejor el procedimiento, vamos a cambiarle el nombre e identificador de un par de nodos, por ejemplo, al nodo *Salar* le vamos a llamar *Salario* y al nodo *Top10* lo vamos a llamar *TopTen* (lo hacemos en las propiedades del nodo, por ejemplo para el nodo Top10:



Recordemos que, en este fichero, todas las probabilidades estaban distribuidas uniformemente (enlaces en gris). Lo que queremos ahora es aprender los parámetros, es decir, aprender valores para dichas probabilidades, que reflejen aduecuadamente los datos. Para ello, tenemos que tener también abierto el fichero de datos (en nuestro caso, el fichero retention_discretized.txt, que se encuentra en la carpeta *Examples).* Una vez tenemos abierta la estructura y los datos, estamos preparados para aprender los parámetros. En el menú *Data* seleccionaremos "*Learn parameters*". Nos saldrá una ventana como esta:



Como vemos, si una columna del fichero de datos tiene el mismo nombre que el identificador de un nodo de la red, automáticamente se establecen las asignaciones (columna central de la figura). Para los nodos o columnas cuyos nombres no coinciden, es necesario realizar las asignaciones de modo manual, para lo cual se utiliza la acción "coger y soltar" del ratón (se selecciona el nodo y se suelta encima de la columna a la que queramos asignarlo, o bien se selecciona la columna y se suelta encima del nodo).

Una vez puestos en correspondencia los elementos de la red, se ejecuta el algoritmo EM (el que hemos visto en clase), que aprende los parámetros necesarios. Vemos que los enlaces ya no están en gris, y para encontrar dichos datos solo tenemos que mirar las tablas de probabilidad de los nodos del fichero PCretention).

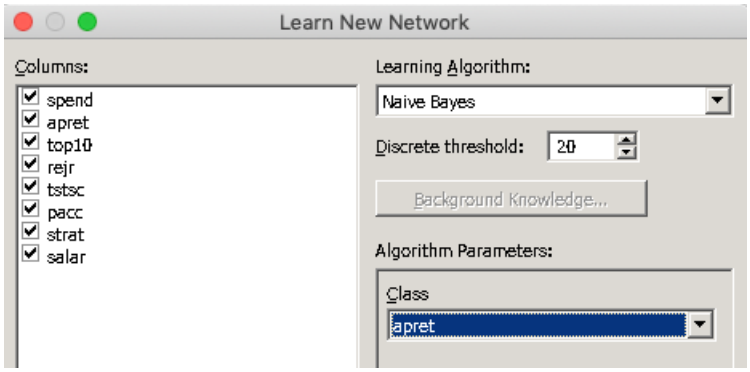
También se nos muestra el valor de lo que en los apuntes hemos llamado p (en GeNIe lo llama Log(p) y este valor se corresponde con la suma de la los logaritmos neperianos de la función de verosimilitud). Como hemos visto en clase, este valor nos sirve para determinar la calidad del modelo aprendido y poder compararlo con otros modelos.

Ahora podemos guardar el modelo generado, con el nombre PCretention.xsdl. Este modelo servirá para hacer predicciones de nuevos casos.

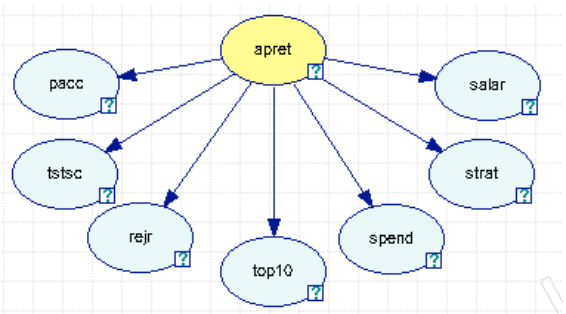
Contesta ahora a las preguntas 1 y 2 (ver sección "Entrega")

2. Aprendizaje de clasificadores *Naive Bayes* en GeNIe

Para aprender un clasificador *Naive Bayes*, sólo tenemos que tener abierta la base de datos (en este caso, retention_discretized.txt) e indicar cuál de las variables es la clase Y. Esto es porque, como explicamos en clase, el modelo *Naive Bayes* va a suponer siempre la misma estructura: la variable clase Y es la raíz del árbol, y el resto de las variables dependen directamente de ella.
En el menú Data pulsamos "Learn new network", seleccionamos el modelo *Naive Bayes*, y tenemos que indicar también cual es la variable clase (en este caso, es Apret, que es lo que queremos determinar (recuerda que la variable Apret representa el porcentaje medio de retención).



Aprenderá la estructura del Naive Bayes:



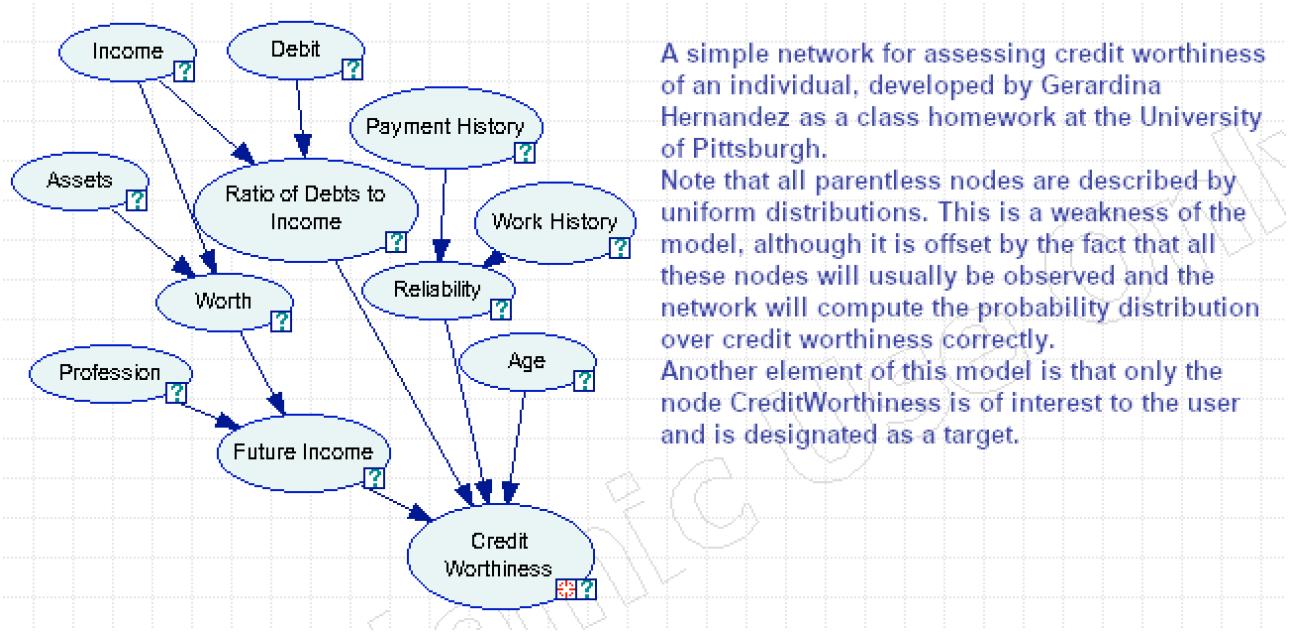
Y los parámetros más adecuados según los datos disponibles. Esta red la podremos utilizar ahora para clasificar nuevos ejemplos.

Contesta ahora a la pregunta 3 (sección "Entrega").

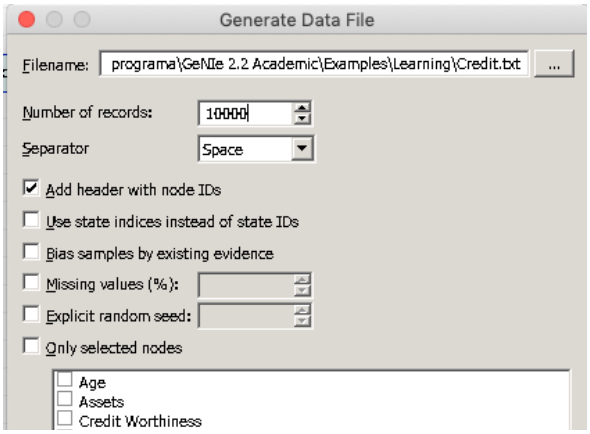
3. Generación de bases de datos a partir de redes bayesianas en GeNIe.

Vamos a aprender también como generar una base de datos "sintética". Una *base de datos sintética* es una base que no se corresponde con datos del mundo real, sino que ha sido generada por ordenador. Las bases de datos sintéticas se pueden generar de diversos modos y con diferentes algoritmos, y se utilizan para hacer pruebas sobre ellas.

En GeNIe es posible a partir de una red bayesiana generar un fichero de datos. Abramos por ejemplo la red "Credit.xsdl" que está dentro de la carpeta "Learning" en la carpeta "Examples".



Seleccionamos ahora la opción "Generate Data File" del menú "Learning", donde podremos configurar diversos parámetros: seleccionar los nodos cuyos datos queremos generar, el número de casos, ect.



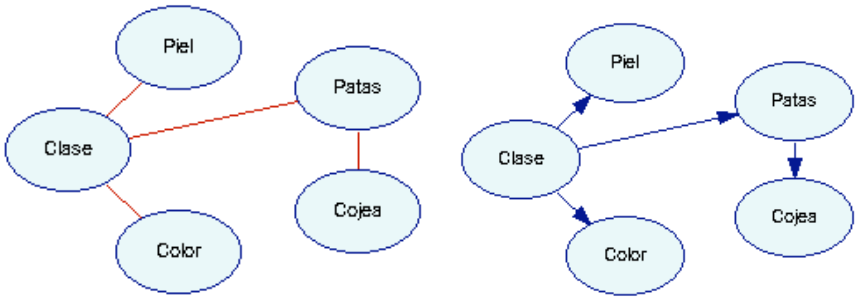
Vamos a generar un fichero con 10000 casos. Utilizamos la opción "Open in GeNIe" y tendremos un conjunto de datos parecido a este:

	PaymentHistory	WorkHistory	Reliability	Debit	Income	RatioDebInc	Assets	Worth	Profession	FutureIncome	Age	CreditWorthiness
	NoAceptable	Unstable	Unreliable	a25901_more	s30001_70000	Unfavorable	average	High	Low_income_profession	Not_promissing	a16_21	Negative
	Excellent	Stable	Reliable	a0_11100	s70001_more	Favorable	poor	Medium	High_income_profession	Not_promissing	a66_up	Positive
	NoAceptable	Stable	Unreliable	a0_11100	s30001_70000	Favorable	wealthy	High	Low_income_profession	Promissing	a16_21	Positive
	Excellent	Justified_no_work	Reliable	a11101_25900	s70001_more	Favorable	poor	High	Low_income_profession	Not_promissing	a22_65	Positive
	Without_Reference	Unjustified_no_work	Reliable	a11101_25900	s0_30000	Unfavorable	poor	Low	Medium_income_profession	Not_promissing	a66_up	Negative
	NoAceptable	Unstable	Unreliable	a11101_25900	s70001_more	Favorable	average	Medium	High_income_profession	Promissing	a16_21	Positive
	Without_Reference	Unstable	Reliable	a11101_25900	s0_30000	Unfavorable	average	Low	Low_income_profession	Not_promissing	a16_21	Negative
	NoAceptable	Unjustified_no_work	Unreliable	a11101_25900	s0_30000	Unfavorable	poor	Low	Medium_income_profession	Promissing	a22_65	Negative
	Aceptable	Stable	Unreliable	a0_11100	s0_30000	Unfavorable	wealthy	High	Medium_income_profession	Promissing	a16_21	Negative
	Without_Reference	Justified_no_work	Unreliable	a25901_more	s30001_70000	Unfavorable	poor	Medium	High_income_profession	Not_promissing	a66_up	Negative

Guardamos este fichero con el nombre CreditTraining, y lo cerramos. Ahora generamos otro fichero con 2000 casos, y lo guardaremos con el nombre CreditTest. Utilizaremos dichos ficheros para validar los modelos generados.

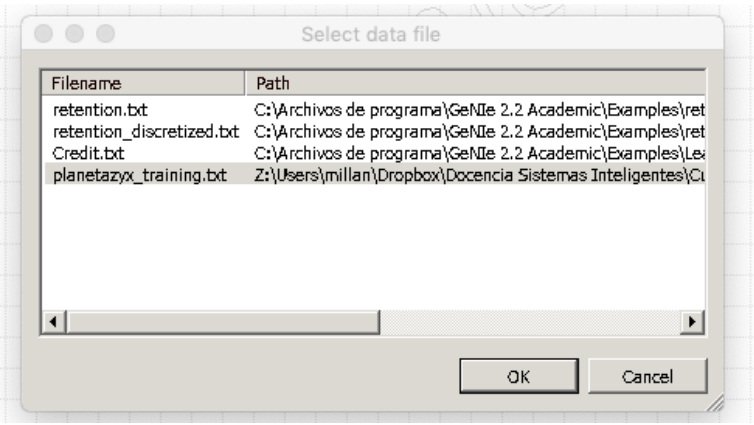
4. Validación de modelos y obtención de medidas de rendimiento con GeNIe

En este apartado vamos a aprender a validar los modelos obtenidos. Para ello, vamos a utilizar dos bases de datos sintéticas, que se han generado a partir de la red del planeta Zyx. Dichas bases de datos se llaman planetazyx_training (datos de entrenamiento) y planetaxyx_test (datos de prueba)
En primer lugar, vamos a utilizar la base d de datos de entrenamiento. Para ello, abrimos el fichero planetazyx_training, y aprendemos la red bayesiana. En primer lugar, aprendemos la estructura, con el algoritmo PC. Veremos que ha sido capaz de determinar las relaciones (arcos), pero no su dirección. Seleccionamos nosotros la dirección adecuada (según el modelo visto en prácticas anteriores) y tendremos que:

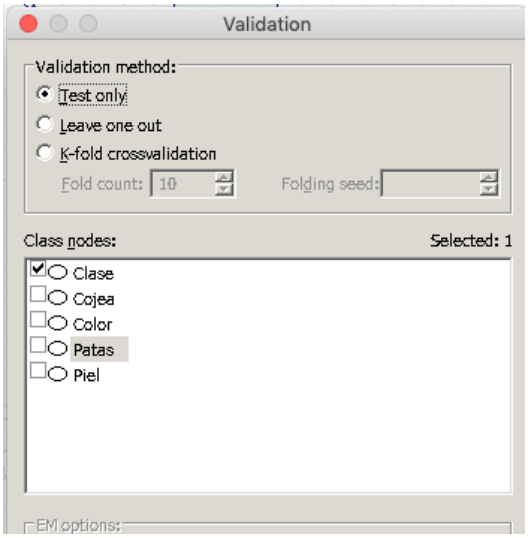


Definida la estructura, pulsamos en la opción Learn Parameters y ya tendremos la red aprendida, con sus parámetros correspondientes. La guardamos con el nombre Planetazyx_aprendido.

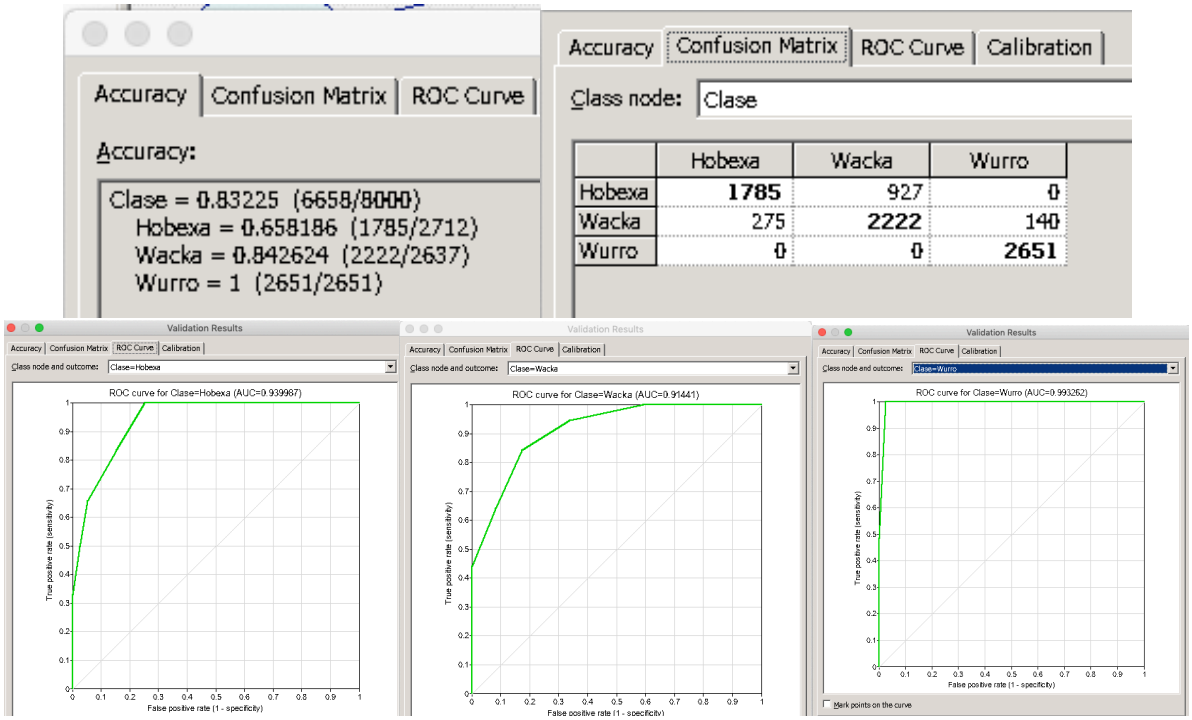
Vamos ahora a probar lo bien que es capaz de predecir esta red los nuevos ejemplos. Para ello vamos a utilizar el conjunto de datos de prueba, planetazyx.test. Para ello, utilizamos la opción Validate del menú Learning.



Nos aparece de nuevo la pantalla que permite emparejar los nodos de la red con las columnas de la base de datos, pero en este caso como tienen los mismos nombres, no hay que hacer nada. Pulsamos Ok. Obtendremos entonces una pantalla donde tendremos que elegir la clase, que en este caso se llama precisamente "Clase". Elegiremos también el método de validación (en este caso, como la validación la hacemos sobre un conjunto de pruebas que no es el mismo que hemos utilizado para entrenar, elegimos "test only"). Pulsamos Ok.



Nos aparecerá una ventana que en sus diferentes pestañas nos muestra los resultados de la ejecución del método: accuracy, matriz de confusión, curva ROC. Observa en tu pantalla que, justo encima de la curva ROC, aparece el valor del área bajo la curva (AUC).



Repite ahora todo este proceso para aprender y validar un modelo Naive Bayes para las mismas bases de datos, y contesta a la pregunta 4.

TAREA Y ENTREGA

Tarea: Contesta a las preguntas que figuran a continuación
Entrega: Documento pdf con la solución (capturas de pantalla y textos descriptivos)

Pregunta 1. Indica el valor de log(p) que te ha dado en el proceso de aprendizaje.

Pregunta 2. Supón un nuevo ejemplo en el que Tstsc toma el valor de más de 78; Top 10 más de 75); Pacc más de 54, Spend más de 33376, Strat menos de 19; Salar más de 74900 y Rejr más de 66. Calcula la probabilidad del nodo Apret para este ejemplo (captura la pantalla que muestra las probabilidades), y di cómo se clasificaría este ejemplo.

Pregunta 3. Repite lo pedido en la pregunta 2, con el modelo *Naive Bayes* obtenido.

Pregunta 4. Escribe un breve informe acerca de la calidad del modelo aprendido, tanto para el caso de redes bayesianas como el modelo Naive Bayes (incluye también los valores obtenidos para el área bajo la curva ROC, en ambos casos). A la vista de los resultados obtenidos, ¿qué modelo es mejor?