

Preparación de datos

Ramon Sangüesa i Solé

PID_00165728

Índice

Introducción	5
Objetivos	6
1. Repaso de conceptos estadísticos	7
1.1. Conceptos mínimos de estadística	8
1.2. La distribución normal	10
1.2.1. Propiedades de la distribución normal	11
1.2.2. Estimación de la probabilidad de un valor muestral	12
1.2.3. Comparación de distribuciones (valores t)	14
1.3. Otras distribuciones	17
2. Terminología de preparación de datos:	
tipos de atributos	18
3. Operaciones de preparación de datos	20
3.1. Transformación de valores	20
3.1.1. Normalización de datos	20
3.2. Discretización	23
3.2.1. Métodos no supervisados	25
3.2.2. Métodos supervisados	32
4. Reducción de dimensionalidad	39
4.1. Reducción del número de atributos	39
4.1.1. Métodos de selección de atributos	40
4.1.2. Fusión y creación de nuevos atributos:	
análisis de componentes principales	46
4.2. Métodos de reducción de casos	47
5. Tratamiento de la falta de datos	49
Resumen	50
Actividades	53
Ejercicios de autoevaluación	53
Bibliografía	55

Introducción


Como ya comentamos al describir el proceso de *data mining*, lo más normal es que los datos necesarios para llevar a cabo un proyecto de *data mining*, una vez seleccionadas, tengan que ser modificados y preparados para, de este modo, poder aplicarles el método de construcción del modelo elegido para la tarea de que se trate.

Podéis ver la preparación de los datos en el subapartado 2.3 del módulo "El proceso de descubrimiento de conocimiento a partir de datos" de esta asignatura.

En este módulo didáctico repasamos las técnicas de preparación de datos más frecuentes:

- a) La transformación de datos; en concreto, la normalización y la discretización.
- b) El tratamiento de los valores no observados (*missing values*).
- c) La reducción de dimensionalidad de los datos; en particular, la selección de atributos.

Haremos un repaso tanto de las técnicas que sólo se dirigen a la transformación de los valores que hay que tratar, como de otras técnicas algo más complejas que requieren tomar decisiones o efectuar hipótesis sobre los valores no presentes o claramente erróneos.

Las técnicas que se presentan a continuación son bastante transversales, independientes, hasta cierto punto, del tipo de modelo que hay que extraer. 

Objetivos

Con el estudio de los materiales didácticos asociados a este módulo, el estudiante alcanzará los objetivos siguientes:

1. Darse cuenta de que, a pesar de la potencia del soporte de *hardware* y *software* que pueden tener los métodos de *data mining*, su complejidad muy a menudo requiere introducir simplificaciones para mantener el coste computacional dentro de unos niveles adecuados sin comprometer la calidad final de los modelos obtenidos.
2. Conocer las principales técnicas de preparación de datos y darse cuenta de su conveniencia para cada tipo diferente de tarea de construcción de modelos.
3. Comprender el concepto de reducción de valores mediante la discretización.
4. Entender el concepto de reducción de dimensionalidad mediante la selección de atributos.
5. Tener una idea de cómo se puede reducir el número de observaciones de un conjunto de datos sin afectar negativamente a la calidad de los modelos que se pueden obtener.

1. Repaso de conceptos estadísticos

Como ya sabemos, en un proyecto de *data mining*, tras seleccionar las fuentes de datos correspondientes –proceso que, por otra parte, puede ser bastante complicado por él mismo– es necesario adecuar tanto los valores como los formatos de los datos.


El **objetivo principal de la preparación** de datos consiste en organizarlos de manera que puedan ser procesados por los programas de construcción de modelos que hayan sido elegidos y, al mismo tiempo, asegurar que los datos se hallan de tal forma que se pueda obtener el mejor modelo posible del conjunto de los datos.

Las operaciones de preparación de datos...

... permiten que éstas puedan ser tratadas por los algoritmos correspondientes y nos den el mejor modelo posible con los datos disponibles.

La preparación de datos...

... ocupa en promedio el 70% del tiempo total de un proyecto *data mining*.

Este último aspecto resulta especialmente difícil, ya que no hay una medida de calidad absoluta ni independiente de la tarea que se quiera realizar, o del tipo de modelo que se quiera obtener. 

Las medidas de calidad de los modelos predictivos, por ejemplo, no tienen por qué ser las mismas que las que se utilizan para obtener buenos modelos de agregación, descripción o explicación. Aun así, se han propuesto métodos generales, no supervisados o “ciegos” al tipo de tarea por realizar que tienen una utilidad lo suficientemente alta.

La calidad...

... del proceso de *data mining* está relacionado con la tarea a la que se desee aplicar el modelo.

Dentro del proceso de preparación de los datos separamos las tareas siguientes:

- **Transformación de valores.** Básicamente, técnicas para cambiar los valores sin perder información y hacer que puedan ser tratados por el método que nos interese.
- **Reducción de dimensionalidad.** Eliminar casos dentro del conjunto de datos original, o bien eliminar atributos, o ambas cosas al mismo tiempo con el objetivo de obtener modelos de la misma calidad con menos esfuerzo computacional.

Con la intención de poder valorar mejor algunas de las técnicas que presentaremos en este módulo y en el resto de la asignatura, será necesario que refresquemos algunos conceptos que proceden de la Estadística. Para profundizar más en algunos detalles, recomendamos la consulta de un texto especializado como la lectura que mencionamos al margen.

Lecturas recomendada

Podéis profundizar en los conceptos de estadística con la lectura de la obra siguiente:
L. Lebart; A. Morineau; J.P. Fenelon (1985). *Tratamiento estadístico de datos*. Barcelona: Marcombo.

1.1. Conceptos mínimos de estadística

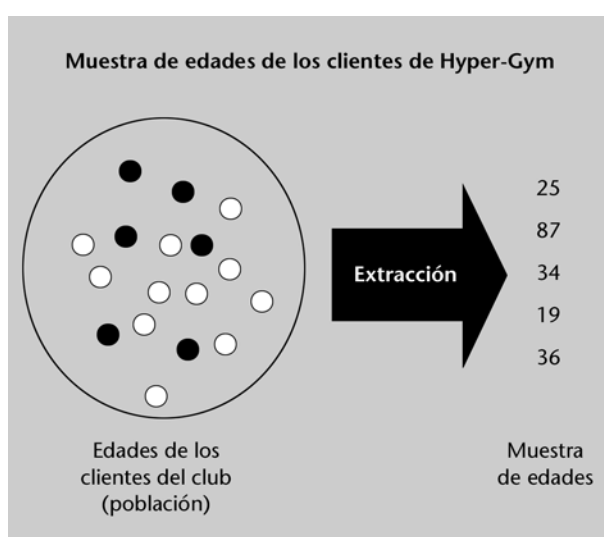
Para hablar con una cierta uniformidad, adoptaremos una buena parte de la terminología utilizada en estadística. En efecto, el proceso de *data mining* incorpora muchos de los intereses y pasos del análisis de datos, y también de la obtención de modelos a partir de conjuntos de datos, que son las tareas propias de la estadística. Es preciso, en cualquier caso, que recordemos una serie de conceptos estadísticos que nos permitirán aclarar las cosas. !

Podemos utilizar toda la serie de conceptos y herramientas aportadas durante años por la inferencia estadística, aquellos conocimientos de estadística que permiten generalizar a partir de los datos observados las propiedades correspondientes a casos que no se pueden observar directamente (en nuestro caso, a futuros casos que tienen que ser clasificados correctamente).

El concepto **población** consiste en todos los objetos que queremos estudiar, y no sólo en aquéllos cuyas características hemos podido observar y recoger directamente. Por ejemplo, todos los clientes de la cadena Hyper-Gym, y no sólo los que hemos utilizado para construir el conjunto de entrenamiento; o todos los clientes (morosos y no morosos) de un banco, y no sólo aquéllos de los cuales hemos utilizado sus datos para construir un modelo de clasificación que prediga quién podría ser moroso y quién, no.

! Podéis ver los casos de ejemplo de los clientes de un banco y de la cadena Hyper-Gym presentados al módulo "El proceso de descubrimiento de conocimiento a partir de datos".

Una **muestra** es un subconjunto elegido aleatoriamente a partir de una población de manera que sea representativo de ésta. Aquí, la dificultad es garantizar que los datos de los casos seleccionados para formar la muestra (los clientes de Hyper-Gym o los del banco) serán efectivamente representativos. Cada conjunto de datos referido a cada caso también se denomina **observación**.



Aquí tenéis una serie de conceptos propios de la estadística descriptiva que nos permitirán hablar con mayor corrección de las propiedades de poblaciones y muestras. En cualquier caso, no los damos de manera formal, sino sólo para

tener un vocabulario de discusión. Para profundizar más en el tema, tenéis que consultar libros de estadística especializados. Los conceptos básicos que nos hacen falta son éstos:

1) **Rango.** Diferencia entre los valores mínimo y máximo de las observaciones de la muestra.

2) **Media muestral.** Suma de todas las observaciones de la muestra dividida por el número de observaciones realizadas. La media muestral (estimada a partir de un conjunto de observaciones) se denota por \bar{X} y se calcula por una muestra de n observaciones según la fórmula:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

en la que x_i es cada uno de los valores observados para la variable X . La media poblacional se representa mediante la letra griega μ .

3) **Varianza muestral.** Medida de la dispersión de los valores de la muestra con respecto a la media. Se denota por s^2 y, para una muestra de n observaciones de una variable x_i , se calcula según la fórmula:

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{X})^2$$

La **varianza poblacional** se representa con la letra griega σ^2 y se calcula según la fórmula:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

4) **Desviación estándar.** Raíz cuadrada de la varianza; podemos volver a aplicar la misma división entre muestral y poblacional. El interés de la desviación estándar se encuentra relacionado con las propiedades de las distribuciones de probabilidad que comentamos a continuación.

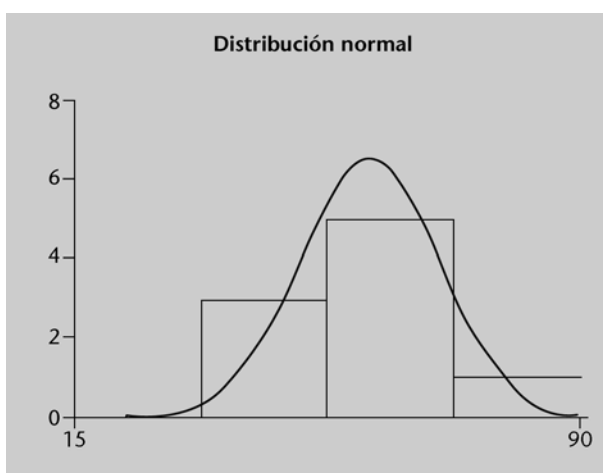
5) **Distribución de probabilidad.** Descripción de cómo se dan los distintos valores correspondientes a una población. Básicamente, la distribución nos indica con qué frecuencia teórica aparecen los valores de la población. La muestra se supone que ha sido extraída de una población que sigue una distribución determinada. Si la discrepancia entre las frecuencias observadas de cada valor en la muestra con respecto a las teóricas de la población es muy fuerte, hay que considerar que (a) los valores no son representativos o (b) la población sigue, en realidad, una distribución diferente.

Lecturas recomendada

Para profundizar en los conceptos de estadística podéis leer la obra siguiente:
L. Lebart; A. Morineau; J.P. Fenelon (1985). *Tratamiento estadístico de datos*. Barcelona: Marcombo.

1.2. La distribución normal

Las distribuciones se formalizan como funciones en las que el eje X representa los valores y el eje Y , la frecuencia teórica de aparición de cada valor (evidentemente, esto admite extensiones a más de una dimensión). Se producen muchos fenómenos en varios ámbitos que siguen distribuciones parecidas. De estos fenómenos se ha obtenido la caracterización funcional de las distintas distribuciones. Las que nos interesan más son la **distribución normal** o **distribución gaussiana** y la **distribución binomial**. Haremos uso de éstas, así como una referencia más detallada cuando la distribución lo requiera. !



La función de densidad de probabilidad de una distribución normal para una variable X presenta el aspecto siguiente (un poco escalofriante a primera vista para quien haga tiempo que no se enfrenta con el análisis matemático):

$$f(X) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

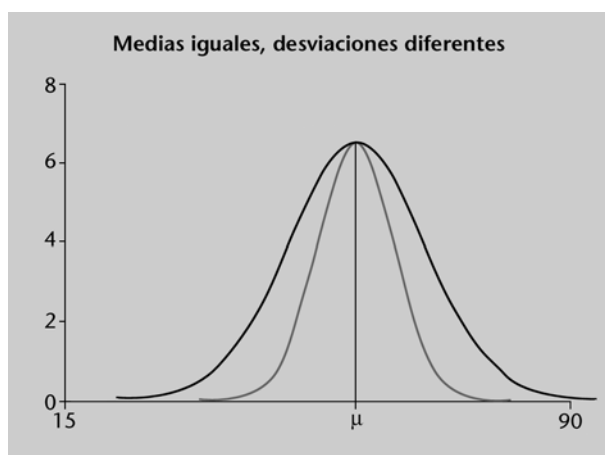
donde π es el número bien conocido por todos; σ es la varianza de la población; μ es su media poblacional y e , el número 2,718... En realidad, esta función no proporciona la probabilidad de cada valor de X . De hecho, la probabilidad es proporcional a la superficie que está comprendida entre esta función y el eje de las abscisas. La suma total de la región bajo la función situada entre los valores mínimo y máximo que puede tomar X es 1, y se puede calcular integrando la función.

No podemos olvidar las características que una función de este tipo da a la distribución. !

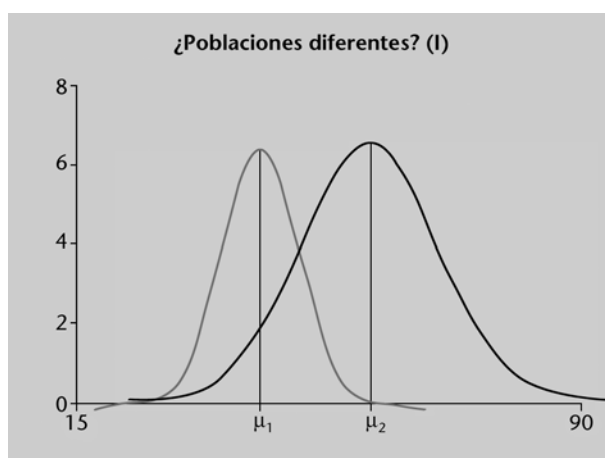
De entrada, tan sólo a partir del aspecto gráfico de la distribución, y si conocemos los valores de la media y la desviación, podemos obtener una primera aproximación a la semejanza o diferencia entre dos poblaciones.

Ejemplo de deducción a partir de dos distribuciones normales

Aquí tenemos dos distribuciones con la misma media, pero con desviaciones diferentes:



A continuación vemos dos distribuciones de las cuales tendríamos la tentación de decir que pertenecen a poblaciones diferentes:



Esta última gráfica corresponde a la distribución de edades entre los hombres y las mujeres que son socios de nuestro club imaginario.

1.2.1. Propiedades de la distribución normal

Lo primero que debemos tener en cuenta es la relación existente entre la media y la varianza poblacional. Si tenemos una población en la que la edad presenta una media de treinta años y una desviación estándar de seis años, una persona de veinticuatro años se sitúa en una desviación estándar de la media poblacional, y una persona de cuarenta y dos años, en dos desviaciones estándar.

Para facilitar las comparaciones, se acostumbra a transformar las distribuciones de manera que la media sea 0, y la desviación típica, 1. Bien, el citado proceso se realiza mediante el cambio de variable siguiente

$$Z = \frac{(x - \mu)}{\sigma}$$

donde x es un valor concreto de la variable X .

Ahora debemos tener en cuenta la simetría de la distribución normal. El punto máximo corresponde justo a la media poblacional (μ). La función decae tanto hacia la izquierda como hacia la derecha. El primer punto de inflexión está a una distancia de una desviación estándar de la media. !

Los valores z tienen la peculiaridad de que permiten comparar medidas parecidas que pueden proceder de poblaciones diferentes.

La importancia de esta forma de representar la distribución normal se ve en calcular la superficie situada bajo la función entre los puntos que se encuentran a una desviación estándar hacia la izquierda y a una desviación estándar hacia la derecha: en esta zona se sitúa el 68% de las observaciones; en dos desviaciones estándar se reúne el 90% de las observaciones. El intervalo normal deja a su derecha el 5% de los casos. Ya veremos qué utilidad tiene todo esto. !

Comparación de medidas de poblaciones diferentes

Nos dicen que una persona procedente de un país donde se come mucha grasa tiene un grado de colesterol al que le corresponde un valor z de 1, y que otra persona de un país donde la población es mayoritariamente vegetariana tiene un grado de colesterol al que le corresponde un valor z de 3. ¿Cuál de ambas está en mejores condiciones?

1.2.2. Estimación de la probabilidad de un valor muestral

Las características de la distribución normal nos ayudarán a contestar preguntas como la de si un determinado valor procede de la población o no.

Intervalo de probabilidad $1 - \alpha$

El intervalo de probabilidad $1 - \alpha$ es el que reúne el $100 - \alpha$ por ciento de observaciones de la población. Este intervalo permite evaluar hasta qué punto las diferencias de valor con respecto a la media poblacional corresponden a variaciones puramente aleatorias. El valor α asociado al intervalo de probabilidad indica la probabilidad de que un valor no pertenezca en el intervalo.


Intervalo de confianza $1 - \alpha$

Si tenemos una muestra que procede de una población desconocida, entonces se puede definir un intervalo simétrico con respecto a una proporción observada de un valor determinado que tenga la probabilidad del $1 - \alpha$ de contener el valor desconocido de esta proporción en la población. También se acostumbra a definir con $\alpha = 0,05$.

Por ejemplo,...

... si hacemos una encuesta entre la gente que accede a la web de Hyper-Gym y alguien nos contesta que tiene cien años y es socio activo del gimnasio, ¿hasta qué punto tenemos que creérmolo? ¿Más que a una persona que afirme que es socia activa y que tiene veintiocho? ¿Menos que a una que afirme que tiene trece?

El **intervalo de probabilidad** presupone que la distribución de la población es conocida. Así pues, indica en qué parte de la distribución se situarían ciertas proporciones de valores. El **intervalo de confianza**, sin embargo, parte del conocimiento de una proporción de apariciones de un valor determinado dentro de una muestra de tamaño n , y nos da un intervalo que tiene una probabilidad elevada de contener la proporción verdadera. El primero permite resolver problemas de predicción, y el segundo, de estimación.

Veremos que algunas de estas ideas se aplicarán a la hora de estimar la calidad de los modelos obtenidos. 


Podéis ver la estimación de la calidad de los modelos obtenidos en el módulo "Evaluación de modelos" de esta asignatura.

A partir del error estándar es posible calcular el intervalo de confianza. En efecto, si en una muestra de n observaciones se ha observado una proporción p para un valor determinado, entonces el error estándar se define como:

$$es = \sqrt{\frac{p(1-p)}{n}}$$

El intervalo de confianza del 95% para un valor p_o se calcula multiplicando su error estándar por dos (lo cual es una aproximación, ya que el valor z que corresponde a $\alpha = 0,05$ es 1,96).

Pruebas de hipótesis

A partir del concepto **intervalo de probabilidad** podemos empezar a contestar preguntas interesantes. Podemos hacer hipótesis y ver si quedan validadas. En otras palabras, tenemos un primer mecanismo para justificar información y generar conocimiento. Ya veremos cómo todo esto también tendrá importancia en la evaluación y validación de modelos. 

Podéis consultar el módulo "Evaluación de modelos".

Pruebas de conformidad

Con relación a las pruebas de conformidad, se trata de saber si los resultados que se han observado en una muestra se avienen con la distribución de población que suponemos.

Ejemplo de pruebas de conformidad

Hemos observado que entre diez socios de uno de los centros de Hyper-Gym, la proporción de personas de treinta años es del 33%, y queremos ver si ese dato se corresponde con las características globales de todos los centros. ¿Qué diríamos si la proporción general fuera del 60%?

Podemos resolverlo comparando la proporción observada con la teórica. Los datos de las edades de los socios son los siguientes:

Edades	30	34	45	30	45	30	60	32	45	43
--------	----	----	----	----	----	----	----	----	----	----

Calculamos el valor z correspondiente:

$$z = \frac{|p - p_o|}{es}$$

En este caso tenemos:

$$es = \sqrt{\frac{0,33 \cdot 0,67}{10}} = \sqrt{0,022} = 0,148$$

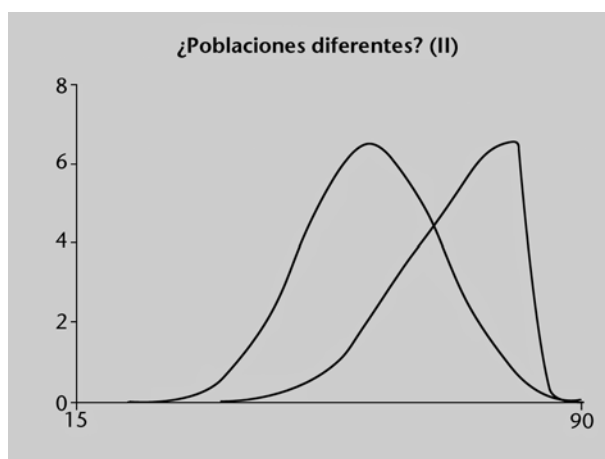
y

$$z = \frac{|0,60 - 0,33|}{0,148} = 1,824$$

Podéis consultar los valores z que se encuentran en el material asociado para saber qué indica este resultado. Fijaos en que el valor de z para un nivel de confianza del 5% es 0,46.

1.2.3. Comparación de distribuciones (valores t)

Vale la pena tener en cuenta los parámetros de una distribución para poder hacer comparaciones. Hay distribuciones con una dispersión muy alta y otras con una distribución muy baja; las primeras son más “planas” que las segundas. Damos, pues, una primera aproximación para saber si dos distribuciones representan la misma población o poblaciones diferentes.



Introduciremos una serie de conceptos y el procedimiento existente para responder a este tipo de preguntas. Por ejemplo, si dos valores proceden de distribuciones diferentes, o si el rendimiento de un método de *data mining* es significativamente mejor que el de otro. En otro módulo se vuelve a hacer referencia a este último punto (punto de gran importancia, por cierto). !

! Podéis ver al respecto el módulo “Validación de modelos” de esta asignatura.

Significación o significancia

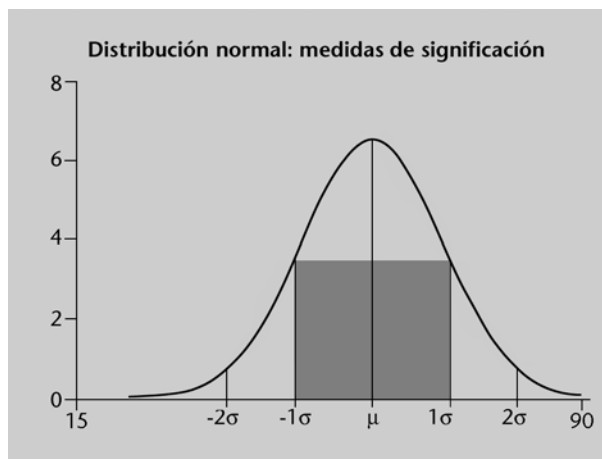
Supongamos que hemos extraído dos muestras de la misma población y apreciamos diferencias entre la frecuencia observada en un mismo valor de cada muestra. Por ejemplo, en la muestra de clientes de Hyper-Gym extraemos dos muestras de cien clientes cada una. En la primera muestra encontramos que la edad es por término medio de treinta y cuatro años, y en la segunda, de treinta y ocho. ¿Es ésta una casualidad o una diferencia significativa? Es decir, ¿las edades de una y otra muestra son lo suficientemente diferentes entre sí?

En general, si la diferencia observada es superior a dos desviaciones típicas y media, consideraremos que, en efecto, la diferencia es significativa porque a esta distancia de la media, bajo la curva de la distribución normal, no queda más que un 5% de la población.

Es decir, sólo tenemos el 5% de posibilidades de encontrar una diferencia así. Estamos seguros de que las muestras proceden de dos poblaciones diferentes con un nivel de confianza del 95%. Veremos otras medidas de significación cuando sea necesario. !

Nos disponemos ahora a precisar un poco más el aspecto de la significación. Supongamos que tenemos dos muestras que no sabemos si proceden de la misma población o de poblaciones diferentes. Llamamos x_1, \dots, x_n a los valores de la primera, e y_1, \dots, y_m a los de la segunda. Sus medias respectivas son \bar{X} e \bar{Y} . Queremos saber si \bar{X} es significativamente diferente de \bar{Y} .

Si el número de observaciones es lo suficiente alto, las medias de un conjunto de observaciones que suponemos independientes (es decir, que no tienen influencia entre sí) siguen una distribución normal.



Supongamos que la media de población es μ . Si conociéramos la varianza de esta distribución normal, podríamos obtener los límites de confianza de μ . Como no la tenemos, podemos estimarla a partir de la varianza muestral. Podemos calcular la varianza de la media \bar{X} dividiendo la varianza muestral (s^2 , en nuestra notación) por el número de observaciones, n . Finalmente, podemos intentar aproximar el valor z correspondiente:

$$\frac{\bar{X} - \mu}{\sqrt{s^2/n}}$$

Este valor, no obstante, no sigue una distribución normal, sino una distribución conocida como t de Student con $n - 1$ grados de libertad, que también está tabulada y podéis encontrar en el mismo material asociado.

El problema que nos planteamos ahora es el de **probar si las dos medias son iguales**, que es lo mismo que ver si la diferencia entre medias, $\bar{X} - \bar{Y}$, es cero.

En este caso se define una variable nueva, m , que recoge las diferencias entre los valores de las observaciones correspondientes, x_i e y_i . Ahora hay que encontrar el estadístico de Student que le corresponde:

$$t = \frac{\bar{m} - 0}{\sqrt{s^2/n}}$$

Consultamos las tablas de t con los correspondientes grados de libertad y un valor de confianza del $\alpha\%$. Si el valor de t es mayor que el que aparece en la tabla, entonces rechazaremos la hipótesis nula (es decir, que las dos medias son iguales).

Ejemplo de significación a partir del estadístico de Student

Podemos comprobar con estos diecinueve valores para la edad de los hombres y diecinueve valores para la edad de las mujeres, seleccionados entre los socios del mismo centro de Hyper-Gym, que la diferencia no es significativa.

Edad (D)	Edad (H)
30	24
32	35
32	36
36	41
30	21
55	56
32	23
40	33
60	32
41	20
36	67
59	24
33	22
34	61
34	30
35	30
32	28
30	23
18	29

Encontraréis que el valor $P(T \leq t)$ es 0,8 y el valor correspondiente a la confianza del 5%, 2,09.

Como este valor está por debajo, tenemos que aceptar que las medias de las dos edades no son significativamente diferentes.

Independencia entre dos variables

Una propiedad interesante en lo que respecta a la relación que hay entre dos variables es la que determina si los valores que adopta una influyen en los que adopta la otra. Si, efectivamente, tienen relación, las variables serán en ese caso dependientes.

Una manera de medir la dependencia o independencia entre las variables es efectuando la **prueba de chi cuadrado** (χ^2). Más adelante, al hablar de discre-

tización, hacemos una descripción más detallada de esta prueba. También encontraréis las tablas necesarias en el material asociado correspondiente.

Podéis ver la prueba de chi cuadrado en el subapartado 3.2 de este módulo didáctico.

Otra forma de medir su dependencia consiste en calcular el **grado de correlación**. Las variables que tienen una correlación alta pueden interpretarse como más dependientes.

1.3. Otras distribuciones

No es cuestión ahora de entrar a discutir y explicar otras distribuciones y tampoco sus propiedades; en todo caso, y dependiendo de si lo necesitamos a lo largo del texto, ya hablaremos de ello en el momento oportuno. Sólo anotaremos aquí que las propiedades que hemos mencionado hasta ahora hacen referencia a distribuciones definidas sobre una única variable (distribuciones univariantes o monovariantes), mientras que en *data mining* los problemas siempre tienen más de una variable. !

En efecto, un conjunto de tuplas extraídas de una base de datos que tiene como atributos X_1, \dots, X_n puede considerarse una muestra extraída de una población definida sobre el mismo conjunto de variables X_1, \dots, X_n . La probabilidad de observar una combinación de valores concretos que corresponden a una de las observaciones de la base de datos $P(X_1=X_{11}, X_2=X_{23}, X_3=X_{33}, \dots, X_n=X_{n3})$ sigue una distribución multivariante, en la que cada una de las variables puede responder a un mismo tipo de distribución (por ejemplo, normal), aunque no necesariamente con idénticos parámetros.

Por ejemplo, ...

... en el caso de nuestro gimnasio ficticio:

$P(\text{Sexo} = \text{'Mujer'}, \text{Edad} = 43, \text{Horario} = \text{'Mañana'}, \text{Años en el club} = 4, \text{Act1} = \text{'Yoga'}, \text{Act2} = \text{'Steps'})$

Y sin tener en cuenta todos los atributos.

Por ejemplo, los parámetros que caracterizan una distribución normal, media y varianza de población no tienen por qué ser los mismos para X_1 que para X_3 . Asimismo, ni siquiera el tipo de distribución que sigue X_1 tiene que ser el mismo que el seguido por X_3 . Así pues, X_1 puede seguir una distribución uniforme, y X_3 , una distribución normal. Llegado el momento oportuno ya veremos hasta qué punto puede afectarnos todo esto en la formalización y caracterización de los métodos de *data mining*. Por ahora basta con que tengamos presentes los conceptos rudimentarios sobre estadística que se han presentado hasta el momento. !

2. Terminología de preparación de datos: tipos de atributos

Recordemos brevemente a continuación los diversos tipos de atributos con que se puede describir un dominio:

- a) **Númericos.** Adoptan valores en los reales, enteros o naturales (en general, en un conjunto numérico que puede tomar infinitos valores).
- b) **Lógicos.** Adoptan los valores 'Verdadero' o 'Falso', normalmente representados por 0 ('Falso') y 1 ('Verdadero').
- c) **Categoricos o discretos.** Aquellos que toman valores en un conjunto finito; por ejemplo {1, 2, 3} o {'Bajo', 'Medio', 'Alto'}.
- d) **Ordenados.** Mantienen una relación de orden entre sí. Por ejemplo, en principio los colores no tienen ningún orden establecido; por lo tanto, si un atributo toma los valores {'Rojo', 'Verde', 'Azul'}, no tenemos por qué considerar que es un atributo con valores ordenados.

También podemos utilizar otros criterios: las dimensiones del rango de valores y la escala de medida utilizada.

Ateniéndonos al rango de valores de la variable, podemos distinguir:

- Las **variables continuas**. Tienen un conjunto de valores infinito no numerable. Caso típico: los números reales.
- Las **variables discretas**. Adoptan valores en un conjunto finito o, como máximo, infinito numerable. Ejemplo: los enteros.
- Las **variables binarias**. Variables discretas que sólo pueden tomar dos valores. Caso típico: la representación de los valores lógicos 'Verdadero' y 'Falso' mediante los números 1 y 0.

Teniendo en cuenta la escala en que se miden las variables, podemos distinguir los tipos que mencionamos a continuación:


- Las **escalas nominales** sólo distinguen dos clases; es decir, para dos valores X e Y , sólo podemos decir si son iguales ($X = Y$) o diferentes ($X \neq Y$). Por ejemplo, si tenemos un atributo que recoge el valor del color de una pieza del almacén, en el que el atributo *Color* se ha definido sobre el conjunto {'Rojo', 'Verde', 'Azul'}.


Lectura complementaria


Encontraréis otros criterios para describir dominios en la obra siguiente:

M. Anderberg (1973).
"Cluster Analysis for Applications". *Academic Press*.


- Las **escalas ordinales** corresponden a los valores sobre los cuales puede definirse una relación de orden. Para dos valores X e Y , además de decir si son iguales ($X = Y$) o diferentes ($X \neq Y$), podemos distinguir $X > Y$ de $X < Y$. Nos permiten, pues, ordenar valores cuando ello sea necesario.
- Una **escala por intervalo** asigna significados a la diferencia de valores. No sólo podemos decir que $X > Y$, sino también que X es $X - Z$ unidades diferente de Y . Ejemplo: las letras del alfabeto en la codificación ASCII.
- Una **escala de proporción** (o **escala de medida**) es un intervalo en el que se ha fijado un valor como origen (o cero). Además de poder decir que $X > Y$, $X < Y$ o $X = Y$, podemos decir que X es X/Y veces mayor que Y . Ejemplo: la temperatura en grados Celsius.

Combinando todas estas escalas, podemos definir las características de prácticamente cualquier tipo de atributo. 

La razón por la cual los valores continuos son tratados por separado es que podemos considerar que pueden llegar a tener tantos valores diferentes que cada uno de esos valores concretos aparece con una frecuencia muy baja. Este hecho puede dar lugar a problemas de precisión en algunos métodos que tienen que hacer comparaciones entre valores, ya que dos valores de este tipo difícilmente serán del todo iguales entre sí. De ahí que se apliquen técnicas de suavización de valores o discretización encaminadas a reducir el número de valores por comparar. 

 Podéis consultar las técnicas de discretización en el subapartado 3.2 de este módulo didáctico.

Observaciones

Los atributos nominales (o atributos simbólicos) son atributos discretos cuyos valores no siguen necesariamente un orden lineal. Como ya hemos señalado, un caso típico es el de los colores. Una variable que represente el color podría tomar los valores 'Rojo', 'Verde', 'Azul', 'Amarillo', 'Negro' y 'Blanco'. A efectos de representación, quizá podríamos codificarlos con los números enteros del 1 al 6 {'Rojo',1), ('Verde',2), ('Azul',3), ('Amarillo',4), ('Negro',5), ('Blanco',6)}. Pero atención: aunque hay un orden entre los enteros que asegura que $1 < 6$, o que 6 es a una distancia de 5 de 1, no tiene ningún sentido utilizar este tipo de conceptos en este caso, porque carece de sentido decir que el negro es mayor que el azul (si no es que, implícitamente, estamos hablando de sus propiedades en términos de longitud de onda). Así pues, habrá que tener mucho cuidado con este tipo de codificaciones, ya que a la hora de hacer comparaciones o reducir valores, podemos introducir una interpretación absolutamente absurda y distorsionar muchísimo tanto el proceso de preparación de datos como los resultados posteriores que se puedan obtener. 

3. Operaciones de preparación de datos

En este apartado describiremos las principales operaciones utilizadas en la preparación de datos, y las dividiremos en dos grandes grupos: la transformación de valores y la reducción del número de atributos que hay que considerar.

3.1. Transformación de valores

Por **transformación de valores** entendemos modificaciones dentro del tipo de valores que pueden adoptar todos o algunos de los atributos.

Las operaciones más habituales son la normalización y la discretización de datos. Comentaremos brevemente la problemática que surge cuando nos faltan los valores de uno o más atributos de algunas observaciones.

3.1.1. Normalización de datos

La **normalización** consiste en situar los datos sobre una escala de valores equivalentes que permita la comparación de atributos que toman valores en dominios o rangos diferentes.

Ejemplo de normalización de datos

Se puede normalizar de manera que los valores numéricos queden dentro del intervalo real $(0, 1)$ o $(-1, 1)$.

Por ejemplo, si nos encontramos con que el atributo *Edad* de nuestro ejemplo del gimnasio oscila entre dieciocho y ochenta y siete años, podemos hacer que vaya a parar a los valores $(0, 1)$. Si el atributo *Años en el club* (que oscila entre 0 y 14) también se transforma de manera que dé valores entre 0 y 1, entonces podemos saber que un valor 0,8 en *Edad* y un valor 0,8 en *Años en el club*, aunque no corresponden al mismo valor, sí que representan valores en la escala alta de los atributos respectivos.

La normalización es útil, o necesaria, para varios métodos de construcción de modelos, como, por ejemplo, las redes neuronales o algunos métodos basados en distancias, como el de los vecinos más próximos. En efecto, si no hay normalización previa, los mencionados métodos tienden a quedar sesgados por la influencia de los atributos con valores más altos, hecho que distorsiona el resultado.

Comentamos a continuación algunas de las técnicas de normalización habituales.

Normalización por el máximo

La normalización por el máximo consiste en encontrar el valor máximo, x_{max} , del atributo por normalizar X y dividir el resto de los valores por x_{max} .

Con este tipo de normalización nos aseguramos que el máximo recibe el valor 1.

Ejemplo de normalización por el máximo

Si tenemos este conjunto de valores para el atributo *Renta*:

53.000.000	15.400.890	7.978.999	3.500.000	12.780.000
------------	------------	-----------	-----------	------------

Entonces x_{max} es 53.000.000 y los valores, una vez transformados, quedan así:

1	0,29	0,15	0,07	0,24
---	------	------	------	------

Normalización por la diferencia

La normalización por la diferencia trata de compensar el efecto de la distancia del valor que tratamos con respecto al máximo de los valores observados.

La normalización por la diferencia consiste en realizar la transformación siguiente:

$$Z_i = \frac{|x_i - x_{min}|}{|x_{max} - x_{min}|}$$

Ejemplo de normalización por la diferencia

Si tenemos al conjunto de valores siguientes para el atributo *Renta*:

53.000.000	15.400.890	7.978.999	3.500.000	12.780.000
------------	------------	-----------	-----------	------------

Entonces x_{max} es 53.000.000 y x_{min} es 3.500.000. Los valores una vez transformados quedan como se sigue:

1	0,24	0,15	0	0,18
---	------	------	---	------

De esta manera, siempre se generan los valores 0 y 1.

Escalado decimal

El escalado decimal permite reducir en un cierto número de potencias de diez el valor de un atributo.

Esta transformación resulta especialmente útil al tratar con valores elevados (por ejemplo, rentas o volúmenes de negocio).

Supongamos un atributo X que muestra un valor X_i . Aquí tenemos la expresión para llevar a cabo su escalado decimal:

$$Z_i = \frac{x_i}{10^j}$$

donde j tiene que ser tal que mantenga el máximo valor que puede adoptar x_i por debajo de uno.

Ejemplo de escalado decimal

Siguiendo con el mismo conjunto de valores para la variable *Renta*:

53.000.000	15.400.890	7.978.999	3.500.000	12.780.000
------------	------------	-----------	-----------	------------

queremos que los valores queden normalizados entre 0 y 1. Entonces tenemos que el valor máximo es 53.000.000. Tenemos que encontrar el valor de j que haga que, una vez transformado, 53.000.000 sea inferior y lo más cercano posible a 1. Este valor de j es 8. El resultado de hacer la transformación:

$$Z_i = \frac{x_i}{10^8}$$

es el que vemos a continuación:

0,53	0,15	0,08	0,04	0,13
------	------	------	------	------

Normalización basada en la desviación estándar: estandarización de valores

Los métodos anteriores no tienen en cuenta la distribución de los valores existentes.

El **método de estandarización de valores** asegura que se obtienen valores dentro del rango elegido que tienen como propiedad que su media es el cero y su desviación estándar vale uno.

La estandarización consiste en hacer la transformación siguiente sobre los valores de los atributos:

$$z_i = \frac{x_i - \mu}{\sigma}$$

Esta normalización resulta adecuada para trabajar después con métodos que utilizan distancias. !

Ejemplo de estandarización de valores

Tenemos este conjunto de valores para el atributo *Edad* de la columna de la izquierda, donde, como podemos comprobar, la media es 37,3 y la desviación estándar, 19,11. Entonces, los resultados que obtenemos de la estandarización son los que vemos en la columna de la derecha:

Edad	Estandarización
22	-0,8006279
43	0,2982732
31	-0,3296703
22	-0,8006279
34	-0,1726845
22	-0,8006279
45	0,4029304
43	0,2982732
23	-0,7482993
88	2,6530615

3.2. Discretización

Recordemos que una buena parte de los métodos de *data mining* trabajan con la suposición de que los atributos con que se describen las instancias u observaciones existentes en la base de datos de la cual se parte son categóricos y no numéricos. Por este motivo, esta es una de las técnicas más utilizadas en la preparación de datos. !

La **discretización** consiste básicamente en establecer un criterio por medio del cual se puedan dividir los valores de un atributo en dos o más conjuntos disjuntos.

Aunque no es éste el único motivo existente para discretizar datos. Citamos éstos otros:

a) Coste computacional. Teniendo en cuenta que el conjunto de valores sobre el cual se trabajará después de discretizar implica una reducción de los va-

Lectura complementaria

Con relación a la velocidad en el proceso de aprendizaje podéis consultar la obra siguiente:

J. Catlett (1991). "On Changing Continuous Attributes into Ordered Discrete Attributes". En: Y. Kodratoff (ed.). *Proceedings of the European Working Session on Learning: Machine Learning* (pág. 164-178). Springer Verlag.

lores por tratar, el número de comparaciones y cálculos que tendrá que realizar el correspondiente método de *data mining* es menor.

b) Velocidad en el proceso de aprendizaje. Se ha demostrado empíricamente que el tiempo necesario para llevar a cabo un proceso de entrenamiento de un método de *data mining* es más corto si se hace uso de datos discretizados.

c) Almacenamiento. En general, los valores discretos necesitan menos memoria para ser almacenados.

d) Tamaño del modelo resultante. Cuando se trabaja con datos continuos, los modelos clasificatorios que se obtienen son comparativamente mayores. Por ejemplo, los árboles de decisión que se obtienen acostumbran a tener un factor de ramificación más alto cuando se trabaja con datos continuos que cuando se trabaja con datos discretos.

e) Comprensión. La comprensión de algunos modelos mejora en gran medida al describir el elemento utilizando menos términos.

Evidentemente, todo método de discretización está obligado a mantener o mejorar las características del modelo a cuya construcción sirve. Por ejemplo, si introducimos un proceso de clasificación y obtenemos tasas de error más altas o de predicción más bajas que sin discretizar, poco hemos ganado. Por lo tanto, lo que buscan la mayoría de los métodos de discretización es mantener la información asociada al atributo que se discretiza. Un inconveniente que normalmente se cita al hablar de métodos de discretización es precisamente la pérdida de información sobre los valores continuos. Este tipo de efectos puede tener una influencia notable en la precisión de los métodos de aprendizaje.

Un ejemplo típico de pérdida de información...

... en el proceso de discretización es que los dos números situados en los extremos del intervalo de discretización se consideran el mismo número. Este efecto influye en la precisión de los modelos de aprendizaje.

Los métodos de discretización se pueden dividir en varias categorías:

- **Supervisados o no supervisados.** Es decir, métodos que tienen en cuenta los valores del atributo clase o no (los primeros dirigidos a clasificación; los segundos, a otras tareas).
- **Locales o globales.** Métodos limitados a un atributo cada vez, o a un subconjunto de los datos originales o conjunto de datos, o bien actuando sobre todos los atributos y todos los datos.
- **Parametrizados y no parametrizados.** Los primeros empiezan conociendo el número máximo de intervalos que hay que generar para un atributo específico, mientras que los demás tienen que encontrar este número automáticamente.

El número de conjuntos al cual se quiere llegar depende de la aplicación y el dominio concretos en que se trabaje. Según el tipo de resultado que deseamos,

Lectura complementaria

Encontraréis más información con relación a las diversas categorías de los métodos de discretización en la obra siguiente:

J. Dougherty; R. Kohavi; M. Sahami (1995). "Supervised and Unsupervised Discretizations of Continuous Features". *Proceedings of the 12th International Conference on Machine Learning* (pág. 194-202). Morgan Kaufmann Publishers.

estas dos discretizaciones pueden ser igualmente útiles o no. En ambos casos utilizamos el atributo *Años en el club* que toma valores entre 0 y 16:


<i>Años en el club</i>	<i>Conjunto</i>
0-2	Inicial
2-4	Estable
4-6	Fiel
6-11	Senior
11-16	Veterano

O bien este otro:

<i>Años en el club</i>	<i>Conjunto</i>
0-4	Estable
< 5	Fiel

Está claro que los métodos existentes para discretizar dependen en gran medida de la tarea que se quiere realizar. Por tanto, la calidad del método estará en función de la mejora de la calidad del modelo que resulte de aplicar una técnica de *data mining* sobre el conjunto de datos discretizados.

En clasificación podemos utilizar la información con respecto a cuál es la variable de clase para mejorar el resultado de la discretización; en agregación el criterio es diferente. Los métodos para discretizar variables en un caso u otro tendrían que ser diferentes, ya que las tareas son diferentes. Aun así, se han desarrollado varias formas de discretizar que no tienen en cuenta el tipo de aplicación final que tendrán y que, por tanto, han de recurrir a otros criterios para tener una base en la que asegurar que el resultado no será peor que sin discretizar.

El objetivo de todos estos métodos, repetimos, es obtener una división en intervalos a partir de un atributo numérico continuo, de manera que a cada intervalo se le pueda asociar una etiqueta (categoría). 

Haremos una descripción progresiva de los métodos desde los que parecen más naturales y sencillos hasta otros más refinados.

3.2.1. Métodos no supervisados

En los métodos no supervisados no tenemos ninguna información acerca de la tarea que tiene que cumplir el modelo que se construirá a partir de los datos discretizados. En consecuencia, sólo utilizamos la información procedente de las observaciones.

Una buena referencia para este subapartado es la lectura recomendada al margen (Martín, 1999), que resume e implementa una buena parte de los métodos siguientes con una descripción de los resultados experimentales obtenidos con varios conjuntos de datos de referencia. Hemos adaptado algunos de los algoritmos que se proponen en esa obra, pero recoge todavía más bastante interesantes (Valdés, 1997).

Métodos de partición en intervalos de la misma amplitud

El método de partición en intervalos de igual amplitud es el método más sencillo que consiste en el algoritmo siguiente.

Dado un atributo X numérico con valor mínimo x_{min} y máximo x_{max} , se siguen los pasos siguientes:

- 1) Fijar un número k de intervalos que hay que alcanzar.
- 2) Dividir el rango de valores (x_{min}, x_{max}) en k intervalos $\{(x_{1min}, x_{1max}), \dots, (x_{kmin}, x_{kmax})\}$ tales que $x_{1min} = x_{min}$ y $x_{kmax} = x_{max}$.

Se trata de hacer que la distancia entre el máximo y el mínimo de cada intervalo sea la misma: $(x_{min} - x_{max})/k$.

Ejemplo de partición en intervalos de igual amplitud

Aquí tenéis los intervalos que se obtienen con $k = 4$ y el atributo *Años en el club* de la base de datos de Hyper-Gym.

Tenemos *Años en el club*_{min} = 0 y *Años en el club*_{max} = 16. Entonces:

Intervalos
0-4
5-8
9-12
13-16

Como todo método demasiado fácil, el método de partición en intervalos de igual amplitud tiene sus inconvenientes:

- a) Da la misma importancia a todos los valores, independientemente de su frecuencia de aparición.

Por ejemplo, si resulta que sólo tenemos un socio que hace dieciséis años que está en el club, tendríamos que pensar que dieciséis no se puede considerar un valor representativo y que, si lo mantenemos, haremos que tanto la capacidad predictiva como descriptiva de este atributo queden alteradas por la importancia exagerada que damos a un valor poco normal.

Lectura recomendada

Podéis consultar los métodos no supervisados en las obras siguientes:

J.R. Martín (1999). *Discretització de variables contínues i la seva aplicació*. Proyecto de Final de Carrera, Ingeniería Informática. Departamento de Lenguajes y Sistemas Informáticos. Barcelona: Facultad de Informática de Barcelona, Universidad Politécnica de Cataluña.

J.J. Valdés (1997). "Fuzzy Clustering and Multidimensional Signal Processing in Environmental Studies". *Proceedings of the 5th European Congress on Intelligent Techniques and Soft Computing*. Aache (Alemania).

b) En caso de que queramos utilizar el método como punto de partida de clasificación, podemos encontrarnos con que se mezclen dentro de un mismo intervalo valores que corresponden a clases diferentes, intervalos poco homogéneos.

Los métodos que, como los árboles de decisión, se basan en la propiedad que acabamos de mencionar para decidir cuándo un atributo es útil para separar el conjunto de datos en clases pueden quedar afectados por el uso de este método de discretización.

Podéis ver los árboles de decisión en el módulo "Clasificación: árbol de decisión" de esta asignatura.



c) Con este método no hay manera de encontrar un valor de k que sea lo suficientemente bueno.

Obtención de intervalos de discretización de igual frecuencia

Uno de los problemas que hemos citado al comentar el método anterior es que puede generar intervalos en los que las distintas clases o valores se distribuyan con frecuencias diferentes. Por este motivo se puede introducir información sobre la frecuencia requerida de la manera siguiente:

- Indicando el número de intervalos que hay que obtener.
- Indicando la frecuencia que se desea obtener para los intervalos:

$$\text{Frecuencia} = \frac{\text{Número de valores}}{\text{Número de intervalos}}$$

El procedimiento consiste en el algoritmo que tenemos a continuación. Dado un vector A que guarda los n valores del atributo que se quiere considerar, X , es necesario seguir los pasos siguientes:

- 1) Ordenar los valores del atributo X que se quiere considerar de menor a mayor: x_1, \dots, x_n .
- 2) Fijar un número de intervalos k .
- 3) Calcular el valor de la frecuencia por intervalo $frec = n/k$.
- 4) Ejecutar el bucle siguiente:

```
Total := 0
Para Intervaloi = 1 hasta k hacer
  Para xi = 1 hasta frec hacer
    Para punto = 1 hasta frec hacer
      Asignar A (Total + punto) a Intervaloi
    fpor
  Total := Total + frec
fpor
```

El método es sencillo de aplicar, pero depende críticamente de la suposición de que se da una distribución uniforme de los valores del atributo por discretizar. Como el criterio es mantener la misma frecuencia dentro del mismo intervalo, valores muy dispares pueden ir a parar al mismo intervalo.

Ejemplo de obtención de intervalos de discretización de igual frecuencia

Aquí tenemos un conjunto de valores para el atributo *Edad*:

55	22	27	40	28	22	28	31	27	31	31	55
----	----	----	----	----	----	----	----	----	----	----	----

Los ordenamos de menor a mayor:

22	22	27	27	28	28	31	31	31	40	55	55
----	----	----	----	----	----	----	----	----	----	----	----

Y guardamos su información de frecuencia:

Valor	22	27	28	31	40	55
Frec.	2	2	2	3	1	2

Haremos una discretización en tres intervalos ($k = 3$); por lo tanto, la frecuencia por intervalo tendría que ser n/k (donde n indica los valores existentes, en este caso 12). Así pues, la frecuencia deseable por intervalo es 4.

22 22 27 27	28 28 31 31 31	40 55 55
-------------	----------------	----------

Los puntos de corte se encuentran en < 27 y < 31 . Como podéis ver, los intervalos no son demasiado uniformes.

Suposición de normalidad (uso de estimador)

Cuando se está en condiciones de suponer que la distribución que siguen los valores del atributo es normal, se puede aprovechar este punto de partida para obtener discretizaciones que mejoren algunos de los problemas que presentan los métodos anteriores.

Recordemos que la función de densidad de probabilidad propia de la distribución normal es:

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

Si el atributo que queremos discretizar proviene de una distribución normal, entonces para cada intervalo que podamos introducir se tiene que cumplir que en el intervalo i -ésimo (x_{imin} , x_{imax}) la proporción de observaciones de la clase i -ésima es igual al área que hay bajo la función de densidad entre los puntos límite de cada intervalo.

La proporción de la población que se sitúa entre ambos extremos es $P(x_{imax} < x \leq x_{imin}) = F(x_{imax}) - F(x_{imin})$, donde $F(x)$ da el valor de densidad de probabili-

Lectura recomendada

Encontraréis la suposición de normalidad explicada con más detalle en la obra siguiente:

M. Anderberg (1973). "Cluster Analysis for Applications". *Academic Press*.

dad correspondiente al punto X una vez hemos efectuado la transformación $((x - \mu)/\sigma)$, que nos permite establecer la comparación con la distribución normal por término medio 0 y la desviación estándar 1.


Podéis ver la estandarización de valores en el subapartado 3.1.1 de este módulo didáctico.



Comprobaremos si la proporción de valores que aparecen en cada intervalo se aviene con la hipótesis de normalidad. El procedimiento es el que explicamos a continuación. Dado un vector A con los valores, el algoritmo efectúa los pasos siguientes:

- 1) Ordenar los valores del atributo x presente en los datos x_1, \dots, x_n .
- 2) Fijar el número de intervalos k .
- 3) Fijar los valores z mínimos y máximos de cada intervalo de aceptación: z_{imin} y z_{imax} .
- 4) Calcular el rango: $|z_{imin} - z_{imax}|$.
- 5) Calcular el rango de cada intervalo: $|z_{imin} - z_{imax}|/k$.
- 6) Para cada valor extremo calculamos la cantidad P_i (proporción de datos) que le corresponde.
- 7) Mientras que la proporción de valores que se han leído hasta ahora sea menor que P_i , ponemos el valor leído $A(i)$ dentro del Intervalo $_i$.

El primer intervalo tendrá $n \cdot F(x_{1x})$ observaciones si el número total de observaciones en la base de datos es n . El segundo intervalo tendrá $n \cdot F(x_{2x})$, y así sucesivamente.

En principio, el método ofrece resultados más que aceptables cuando la distribución que siguen los datos puede considerarse normal. A veces es útil utilizar el valor máximo observado en la muestra como posible punto de corte del último intervalo. Sin embargo, también puede ocurrir que sea un valor extremo o bien que sea un valor inferior al que le correspondería en la muestra distribucional. 

El método puede hacerse más sencillo si se evita pedir al usuario los valores z y, en cambio, se calculan a partir de los datos.

Método *k-means*

Volveremos a encontrar la idea del método *k-means* en otro contexto, cuando hablemos de métodos de agregación. De hecho, puede considerarse que una manera de afrontar el problema de la discretización o reducción de valores para una variable consiste en repartir los valores entre los distintos intervalos

de manera que, dados k intervalos, se minimice la distancia entre cada valor y el valor medio de su intervalo correspondiente. En el fondo, lo que se hace es generar varios grupos e ir fusionando los grupos (conglomerados o *clusters*) más próximos. !

Trataremos de las funciones de distancia al hablar de los métodos de agregación. Hay varias; aquí sólo adelantaremos una de las más utilizadas en este contexto: la distancia euclídea.

La **distancia euclídea** entre dos valores x_1 y x_2 se calcula mediante la fórmula siguiente:

$$Dist(x_1, x_2) = \sqrt{\sum (x_1 - x_2)^2}$$

El **método de *k-means*** consiste en comparar cada valor x_j con el valor medio de los dos intervalos adyacentes: $Intervalo_i$, que es el que ocupa en un momento dado, y el siguiente, $Intervalo_{i+1}$:

- Si la distancia al valor medio de su intervalo \bar{x}_i es menor que con respecto al siguiente \bar{x}_{i+1} , el valor permanece en su intervalo.
- En caso contrario, se pasa al intervalo siguiente.

En el caso de la distancia euclídea, la fórmula de cálculo de la distancia es:

$$Dist(x_j, x_i) = \sqrt{\sum (x_j - \bar{x}_i)^2}$$

A partir de esta distancia y de la partición en intervalos que hay en un momento dado, podemos ver que las particiones totales existentes son homogéneas. Una manera de calcular esto es ver cuál es el error medio que se comete con la asignación actual de valores a intervalos. Se calcula, a tal efecto, la distancia media de los valores de cada intervalo a sus “centros” (o medianas) correspondientes, y se extrae la media aritmética para la actual configuración de intervalos. Para un conjunto de valores X y un conjunto de particiones k con medias de intervalo $\bar{x}_1, \dots, \bar{x}_k$, se tiene la expresión del error siguiente:

$$Error_{actual}(configuración) = \frac{1}{k} \sum_{i=1}^k \sqrt{\sum_{j=1}^{i_{max}} (x_j - \bar{x}_i)^2}$$

donde i_{max} es el número de valores que tiene el intervalo i en la configuración actual.

Si en la iteración siguiente el error de la configuración aumenta, eso será una indicación de que introducimos intervalos todavía más desordenados que los que teníamos en el paso anterior. En tal caso, el algoritmo se detiene.

Lectura recomendada

Encontraréis el método *k-means* explicado en la obra siguiente:

J.I. Hartigan; M.R. Wong (1979). “A *k-means* Clustering Algorithm”, ALGORITHM AS 136. *Applied Statistics* (vol. 28, núm. 1).

El algoritmo de discretización para k -means funciona de la manera que explicamos a continuación. Si tenemos un vector A con los valores x_1, \dots, x_n , se efectúan los pasos siguientes:

- 1) Ordenar los valores del atributo X presente en los datos x_1, \dots, x_n .
- 2) Fijar el número de intervalos k y el número de valores diferentes n .
- 3) Ejecutar el bucle siguiente:

Para cada $Intervalo_i$ **hacer**
 Asignar a $Intervalo_i$ **los** n/k **valores siguientes**
 Fpor {esto genera la $Configuración_1$ }
 Calcular $Error_{actual}(Configuración_1)$
 Repetir

- 4) Guardar la configuración actual en $Configuraciones_i$:

$$Error_{anterior}(Configuraciones_i) = Error_{actual}(Configuraciones_i)$$

- 5) Asignar los valores en cada intervalo:

Para cada valor x **de** $Intervalo_i$ **hacer**
 Si $Dist(x_j, \bar{x}_i) > Dist(x_j, \bar{x}_{i-1})$ **entonces asignar** valor x_j **a** $Intervalo_{i-1}$ **fsi**
 Si $Dist(x_j, \bar{x}_i) > Dist(x_j, \bar{x}_{i+1})$ **entonces asignar** valor x_j **a** $Intervalo_{i+1}$ **fsi**
 Fpor {Aquí se ha generado una nueva configuración, $Configuraciones_{i+1}$ }
 Calcular $Error_{actual}(Configuraciones_{i+1})$
 Hasta que $Error_{actual}(Configuraciones_i) \geq Error_{anterior}(Configuraciones_{i+1})$

Ejemplo de método de discretización por k -means

Podemos ver un ejemplo de un conjunto de edades:

22	22	22	27	27	28	31	31	31	40	55	55
----	----	----	----	----	----	----	----	----	----	----	----

El número de valores diferentes es $n = 6$. Supongamos que queremos obtener una partición con $k = 3$ intervalos. Entonces los intervalos iniciales están formados por n/k elementos ($6/3 = 2$) y son:

Intervalo 1	Intervalo 2	Intervalo 3	Intervalo 4	Intervalo 5	Intervalo 6
22 22	22 27	27 28	31 31	31 40	55 55

Sus medianas y errores son éstas:

Intervalo	Media	Distancia media (error del intervalo)
1	22	0
2	24,5	1,666
3	27,5	0,333
4	31	0
5	35,5	3
6	55	0

El error medio de esta configuración es 1,666. La nueva configuración es:

Intervalo 1	Intervalo 2	Intervalo 3	Intervalo 4
22 22 22	27 27 28	31 31 31 40	55 55

Y las medias y distancias correspondientes son las siguientes:

Intervalo	Media	Distancia media (error del intervalo)
1	22	0
2	27,333	0,444
3	33,25	4,5
4	55	0

Podemos ver fácilmente que la configuración final obtenida es:

Intervalo 1	Intervalo 2	Intervalo 3
22 22 22 27 27 28	31 31 31 40	55 55

Otras variantes del método uniformizarían todavía más el resultado asignando a cada intervalo el valor mayoritario.

3.2.2. Métodos supervisados

Los métodos que hemos discutido hasta ahora no se dirigían a ninguna tarea concreta, podríamos utilizarlos tanto para clasificación como para agregación. Ahora bien, en el caso de la clasificación nos interesa asegurar que dentro de cada intervalo generado los valores de las distintas clases que pueden existir se distribuyen lo más uniformemente posible. Ésta es una problemática que volvemos encontrar en otro nivel al hablar de árboles de decisión. Algunas de las soluciones que se han aportado son comunes al problema de la discretización y a algunos pasos de la construcción de árboles de decisión. !

Podéis ver los árboles de decisión en el módulo "Clasificación: árbol de decisión" de la presente asignatura.

Método chi merge

Una buena discretización de datos tendría que mostrar la propiedad consistente en que dentro de cada intervalo apareciesen representadas todas las clases de forma parecida, y muy diferente respecto de otros intervalos. En términos de frecuencias, eso querría decir que dentro de un intervalo habría que esperar a que la frecuencia de aparición de valores de cada clase fuera parecida, y muy diferente de la de otros intervalos. Si la primera condición no se cumple, quiere decir que tenemos un intervalo demasiado heterogéneo y que sería preciso subdividirlo. Si no se cumple la segunda, entonces quiere decir que el intervalo es muy parecido a algún otro intervalo adyacente, de manera que, en tal caso, sería necesario unirlos. Una manera de medir estas propiedades es recurriendo al estadístico χ^2 para averiguar si las frecuencias correspondientes a dos intervalos son significativamente diferentes (y, por lo tanto, ya es correcto que los intervalos estén separados), o bien que no lo son y hay que unirlos.

Recordemos que el estadístico se calcula mediante la siguiente expresión:

$$\chi^2 = \sum_{i=1}^n \sum_{j=1}^k \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$$

donde:

- n es el número de intervalos que hay que comparar entre sí; para simplificar, haremos $n = 2$.
- k es el número de clases que hay.
- N_{ij} es el número de valores de la clase j presente en el intervalo i .

En este método resulta crítico situar el nivel de significación de χ^2 bastante adecuado para evitar situaciones no deseadas. Por ejemplo, si el valor es demasiado alto, obligaremos a efectuar muchas uniones de intervalos consecutivos; por el contrario, un valor creará bajo un número elevado de intervalos de poca amplitud.

La idea es fusionar intervalos hasta que cada intervalo muestre un nivel de χ^2 adecuado. De manera alternativa, se puede prefijar un número mínimo o máximo de intervalos por construir.

El algoritmo chi merge consta de los pasos siguientes:

- 1) Ordenar los valores del atributo x_1, \dots, x_m .
- 2) Fijar el valor de significación de χ^2 , s .


Lectura recomendada

Podéis consultar el método chi merge en la obra siguiente:

R. Kerber (1992). "Chimerge: Discretization of Numeric Attributes". *Proceedings of the 10th National Conference on Artificial Intelligence* (pág. 123-127).

- 3) Fijar los n grados de libertad para la distribución de χ^2 ; $n = k - 1$, donde k es el número de clases.
- 4) Obtener una primera discretización en la que cada valor es su propio intervalo: $Intervalo_1, \dots, Intervalo_m$.
- 5) Repetir los pasos siguientes hasta que para todo par $Intervalo_i, Intervalo_j$ adyacentes $significación(Intervalo_i) > s \wedge significación(j) > s$:
 - a) Calcular χ^2 para cada par de intervalos adyacentes $Intervalo_i, Intervalo_j$.
 - b) Encontrar el par de intervalos con χ^2 menor $Intervalo_{imin}, Intervalo_{jmin}$.
 - c) Fusionar los intervalos.

Este método posee la ventaja de utilizar un estadístico que no depende del número de clases o ejemplos. En general, se comporta correctamente sin introducir puntos de corte que no separen intervalos buenos, en este caso sin proponer puntos de corte intermedios.

El problema es que exige conocer el atributo de clase. También es un algoritmo demasiado local: sólo considera los intervalos adyacentes dos a dos. 

Métodos basados en medidas de entropía

Este método parte de la misma idea que hay tras los algoritmos de construcción de árboles de decisión basados en medidas de información. En efecto, en el caso de los árboles se trataba de encontrar particiones del conjunto original de datos que fueran internamente tan homogéneas como se pudiera, y con respecto al resto de las particiones, tan diferentes como fuera posible. Éstas son las mismas características de los intervalos de una buena discretización. ¿Cómo podemos aprovecharlo?

En el caso de los árboles de decisión que trabajan sobre variables continuas para crear un nodo de decisión sobre un atributo determinado, el objetivo es encontrar aquel valor del atributo que permite alcanzar una mejor separación de los datos. De esta manera se establece un límite o umbral. Las observaciones que para el atributo considerado presentan un valor menor en el umbral se asignan al subárbol izquierdo, y las demás, al derecho.

La idea general del método es la siguiente:

- 1) Ordenar los valores del atributo continuo X_i que se esté considerando: x_1, \dots, x_m .
- 2) Crear el conjunto de puntos de corte candidatos:
 - a) Considerar como umbral posible el punto medio, x_{medio} , entre dos valores consecutivos de X_i : $x_{medio} = (x_i + x_{i+1})/2$:

$$candidatos = \{x_{medio}, \dots, x_{(1+m)/2}\}.$$


Lectura recomendada

Podéis consultar los métodos basados en medidas de entropía en la obra siguiente:
U.M. Fayyad; K.B. Irani
 (1993). "Multi-interval Discretization of Continuous Valued Attributes for Classification Learning".
Proceedings of the 13th International Joint Conference on Artificial Intelligence
 (pág. 1022-1027).

3) Para cada $x_i \in \text{candidatos}$, ejecutar los pasos siguientes:

- a) Establecer $C_{i1} = \{\text{observaciones } C_j \text{ de la base de datos tales que el valor de } X_j < x_i\}$.
- b) Establecer $C_{i2} = \{\text{observaciones } C_j \text{ de la base de datos tales que el valor de } X_j \geq x_i\}$.
- c) Calcular $E_1 = \text{Entropía}(C_{i1})$.
- d) Calcular $E_2 = \text{Entropía}(C_{i2})$.

4) Elegir x_i , tal que $\text{Entropía}(C_{i1} \cup C_{i2})$ sea mínima.

Podemos utilizar aquí los conceptos de entropía y entropía asociada en una partición que también mencionamos al hablar de árboles de decisión de forma similar a como hemos definido la entropía de clase al discutir los árboles de decisión. 

Podéis ver los conceptos de entropía y entropía asociada en el subapartado 2.1 del módulo "Clasificación: árboles de decisión de esta asignatura."

En efecto, supongamos que hemos elegido un valor de corte x_t . Si tenemos un conjunto inicial de valores del atributo X , el punto de corte divide este conjunto de valores en dos: los que son inferiores a x_t (que denotamos por X_1) y los que son superiores a x_t (que denotamos por X_2). Suponiendo que hay k clases, C_1, \dots, C_k , y que la proporción de valores de X que corresponden a la clase C_i sea $P(X, C_i)$, entonces la entropía de la clase C_i para un conjunto de valores X_i es la siguiente:

$$I(X_i) = - \sum_{i=1}^k P(C_i, X_i) \log_2 P(C_i, X_i)$$

Si definimos X como el conjunto de valores del atributo correspondiente, X_1 , como el conjunto de valores inferiores al punto de corte y X_2 , como el conjunto de valores superiores, la entropía de clase debida a la elección de un punto de corte x_t es:

$$I(X_i, x_t) = \frac{|X_1|}{|X|} I(X_1) + \frac{|X_2|}{|X|} I(X_2)$$

Hay dos posibilidades. Veámoslas:

- Dividir los valores del atributo X en un solo punto de corte que haga que la entropía de las particiones sea mínima.
- Dividir el conjunto de valores por el punto de corte que tenga la entropía mínima (a menos que todos los valores del conjunto sean iguales o correspondan a la misma clase).

Es necesario que aclaremos el criterio que se utiliza para definir el final de este proceso de partición repetida sobre el punto de entropía mínima. Tenemos que detenemos en el momento en que tengamos intervalos bastante homogéneos. ¿Cómo caracterizamos ese momento?

Un criterio muy útil y que da buenos resultados para saber en qué momento hay que detener el proceso de partición es el de minimización de la longitud de la descripción. Este concepto muy general puede resumirse diciendo que el mejor modelo (M) es, dado un conjunto de datos (D) aquel que minimiza el tamaño del modelo (es decir, el que nos da el modelo más compacto) y, simultáneamente, también requiere el mínimo de información (datos) para construirlo. Esta intuición se formaliza teniendo en cuenta que ambas cosas, el modelo y los datos, pueden codificarse mediante bits (unidades de información). Por lo tanto, ante un modelo y unos datos, se trata de minimizar la longitud de la codificación.

En otras palabras, dados unos datos D , un modelo M y un esquema de codificación C , entonces la **longitud de codificación** es:

$$L_c(D) + L_c(M|D)$$

es decir, la longitud de codificación es la longitud de la codificación de los datos observados según el esquema de codificación elegido, más la longitud de la codificación del modelo M teniendo en cuenta que procede del conjunto de datos D .

¿Cuál es el modelo M y los datos en nuestro problema de discretización?

La codificación de algún tipo de información se realiza mediante bits. Se puede demostrar que el número de bits necesarios para codificar un valor x_i es el logaritmo en base 2 de su probabilidad (lo cual es, de hecho, una medida de la información de este valor).

En nuestro caso, el modelo es el punto de corte, los datos, los valores y las clases asociadas. Queremos comparar el tamaño de codificación del modelo y la cantidad de datos que necesitamos para especificarlo. Si no decidimos introducir una partición en este punto, si mantenemos el intervalo tal como está, entonces tenemos que describir los datos resultantes utilizando las etiquetas de clase de cada valor. Si introducimos un punto de corte, entonces tenemos que codificar:


- El valor del punto de corte. Necesitamos $\log_2(N - 1)$ para codificarlo (donde N es el número de valores que hay).
- Las clases de los valores que hay por debajo del punto de corte.
- Las clases de los valores que hay por encima del punto de corte.

Lecturas complementarias

Encontraréis más información sobre el criterio de la longitud de la descripción en las obras siguientes:

J.R. Quinlan (1989). "Inferring Decision Trees Using the Minimum Description Length Principle". *Information and Computation* (núm. 80, vol. 3. pág. 227-248).

M. Rissanen (1985). "The Minimum Description Length Principle". En: S. Kotz; N.L. Johnson (ed.). *Encyclopedia of Statistical Sciences* (vol. 5). Nueva York: John Wiley & Sons.

Dependiendo de la homogeneidad de los intervalos resultantes, valdrá la pena o no introducir un punto de corte. Si la codificación obtenida introduciendo el punto de corte es más corta que sin hacerlo, la introduciremos; en caso contrario, no la introduciremos. 

Si hay exactamente el mismo número de valores que pertenecen a la clase 1 que los que pertenecen a la clase 2, entonces codificar cada valor cuesta aproximadamente un bit. Ahora bien, si introducimos el punto de corte de manera que por debajo suyo todos los puntos pertenecen a una clase, y por encima, a la otra, entonces codificar cada valor tiene un coste muy próximo a cero.

La partición que genera un punto de corte determinado es válida si aporta una ganancia de información que supera un umbral determinado. No entraremos a precisar cómo se deriva esta fórmula, pero tenemos que si el punto introduce una partición en dos intervalos y definimos los elementos siguientes:

- N es el número de valores total.
- El atributo X cuyos valores estamos considerando.
- k es el número de clases por considerar.
- El número de clases diferentes asociadas a los valores del primer intervalo X_1 (lo que quedaría por debajo del punto de corte) es k_1 .
- El número de clases diferentes asociadas a los valores del segundo intervalo X_2 (lo que quedaría por encima del punto de corte) es k_2 .
- La entropía del conjunto de valores inicial, sin partir, es I .
- La entropía del conjunto que queda por debajo del punto de corte es I_1 .
- La entropía del conjunto que queda por encima del punto de corte es I_2 .

Entonces, la expresión de la **ganancia de información** sobre un conjunto de N valores del atributo X introducida por un punto de corte x_t que parte el conjunto original en dos subconjuntos X_1 y X_2 es:

$$Ganancia(X, x_t, X_1, X_2) = \frac{\log_2(N-1)}{N} + \frac{\log_2(3^k - 2) - kI + k_1I_1 + k_2I_2}{N}$$


Observad que la primera parte de esta fórmula indica la cantidad de información necesaria para codificar el valor del punto de corte; la segunda parte indica la cantidad de información necesaria para codificar a qué clases pertenecen los valores del intervalo inferior y las del superior.

Lectura complementaria

Encontraréis la derivación de la expresión de la ganancia de información en la obra siguiente:

U.M. Fayyad; K.B. Irani (1993). "Multi-interval Discretization of Continuous Valued Attributes for Classification Learning". *Proceedings of the 13th International Joint Conference on Artificial Intelligence* (pág. 1022-1027).

Podemos minimizar la longitud de descripción total si en cada paso de discretización elegimos aquel punto que maximiza la ganancia de información. Se puede demostrar que todo punto que minimice la entropía de la partición tiene que ser un punto de corte, y no sólo eso, sino que se asegura que los puntos de corte así escogidos siempre se encuentran entre valores que pertenecen a clases diferentes.

En conjunto, éste es un comportamiento muy adecuado y la combinación de la discretización basada en la entropía, junto con el criterio de no introducir particiones o detener el proceso de partición cuando no aumenta la ganancia de información, da como resultado muy buenas discretizaciones. 

Ejemplo de método basado en medidas de entropía

Veamos un ejemplo sencillo de aplicación de un método basado en medidas de entropía.

Aquí tenemos información sobre el valor del atributo *Edad* para un conjunto de socios de Hyper-Gym que pertenecen a la clase 1 o a la clase 2 (a efectos del ejemplo es irrelevante, aunque remarcamos que la clase 1 corresponde a los socios que solicitan entrenador personal, y la clase 2, a aquellos que no lo solicitan):

Valores	22	24	28	35	38	42	43	47	49	55
Clase	1	2	2	12 21	2 2	1 1	2	1	1	2

Como podemos ver, hay valores que son menos homogéneos que otros. Por ejemplo, el 35 acumula dos observaciones de la clase 2 y otras dos de la clase 1.

¿Cuántos lugares posibles puede ocupar el punto de corte? Recordemos que inicialmente se elige como punto de corte el punto medio entre dos valores consecutivos $(x_i + x_{i+1})/2$. Éstos son los siguientes:

23	26	31,5	36,5	40	42,5	45	48	52
----	----	------	------	----	------	----	----	----

Tenemos que ver la calidad de cada uno de ellos. Tenemos que calcular el contenido de información de los intervalos que generaría cada punto de corte. Por ejemplo, el 23 deja a su izquierda un único valor de clase 1 y a su derecha, ocho valores de la clase 2 y seis valores de clase 1. ¿Qué cantidad de información posee esta partición?

$$I(\text{Edad}, 23) = \frac{1}{10}I(\text{Edad}, \text{Edad}_1) + \frac{9}{10}I(\text{Edad}, \text{Edad}_2)$$

Donde Edad_1 , corresponde al conjunto de puntos inferiores a 23, y Edad_2 , al conjunto de puntos con valores superiores a 23.

Calculamos los valores correspondientes:

$$\begin{aligned} I(\text{Edad}, \text{Edad}_1) &= -P(C_1, \text{Edad}_1) \log_2 P(C_1, \text{Edad}_1) - P(C_2, \text{Edad}_1) \log_2 P(C_2, \text{Edad}_1) = \\ &= -(1/1) \cdot \log_2 1 - (0/1) \cdot \log_2 0 = 0 \end{aligned}$$

$$\begin{aligned} I(\text{Edad}, \text{Edad}_2) &= -P(C_1, \text{Edad}_2) \log_2 P(C_1, \text{Edad}_2) - P(C_2, \text{Edad}_2) \log_2 P(C_2, \text{Edad}_2) = \\ &= -(6/14) \log_2 (6/14) - (8/14) \log_2 (8/14) = \\ &= 0,42 \times 1,22 + 0,57 \times 0,80 = 0,98 \end{aligned}$$

Por lo tanto,

$$I(\text{Edad}, 23) = \frac{1}{10}I(\text{Edad}, \text{Edad}_1) + \frac{9}{10}I(\text{Edad}, \text{Edad}_2) = 0,88$$

Si repitiéramos el proceso, encontraríamos el conjunto de puntos de corte final.

4. Reducción de dimensionalidad


Una vez tenemos los datos en el formato adecuado para el tipo de modelo que se quiere obtener y el método para construirlo, todavía es posible aplicar una serie nueva de operaciones con el fin de asegurar dos objetivos: la reducción del número de atributos por considerar y la reducción del número de casos que hay que tratar, asegurando, asimismo, que se mantendrá la calidad del modelo resultante.

El motivo para efectuar la reducción acostumbra ser doble:

- 1) El programa de construcción del modelo elegido no puede tratar la cantidad de datos de que disponemos.
- 2) El programa puede tratarlos, pero el tiempo requerido para construir el modelo es inaceptablemente largo (días de UCP, por ejemplo).

Hay algún otro motivo un poco más sutil que vendría a decirnos que no siempre más datos quiere decir que vayamos a tener un mejor modelo. En efecto, cuando se consideran muchos datos, hay que ajustar el modelo a más casos, con lo que nos perdemos detalles que pueden ser importantes.

Con las **operaciones de reducción de dimensionalidad** que comentaremos más adelante queremos mantener las características de los datos iniciales, ya que eliminamos datos que no son necesarios.

Otros métodos alteran los datos, bien por fusión de atributos, bien por cambio de valores, de manera que cuesta recuperar los datos originales de cara a futuras interpretaciones. 

4.1. Reducción del número de atributos

La reducción del número de atributos consiste en encontrar un subconjunto de los atributos originales que permita obtener modelos de la misma calidad que los que se obtendrían utilizando todos los atributos. Este problema se denomina **problema de la selección óptima de atributos**.

La selección de atributos...

... es un proceso que intenta seleccionar aquel subconjunto de los atributos originales que mantendrá la calidad de los modelos que se puedan extraer a partir de los datos.

El sentido común nos impone un procedimiento que parece el más evidente, pero que presenta claras desventajas. Se trata de ir generando todos los sub-


conjuntos posibles de $n - 1$ atributos, elegir el mejor, generar los de $n - 2$, etc., y así hasta que no haya ninguna mejora o se empiece a degradar la calidad de los modelos.

Con el solo hecho de considerar cuántos subconjuntos se generan por medio de este proceso ya nos daremos cuenta de que no es práctico. Así pues, se han desarrollado métodos indirectos que, a partir de medidas sobre las propiedades de los atributos o de las relaciones existentes entre éstos, nos permiten detectar cuáles son irrelevantes, redundantes o equivalentes.

4.1.1. Métodos de selección de atributos

Los métodos de selección de atributos consisten en saber qué subconjunto de atributos puede generar un modelo con la misma calidad que el que se podría extraer con todo el conjunto de atributos iniciales. Se trata de ver qué medidas sobre los atributos pueden utilizarse para asegurar que se puede introducir una reducción sin pérdida de información.

En parte, todo se reduce a saber hasta qué punto determinados valores pertenecen a la distribución que se corresponde con los datos. Si, por ejemplo, un atributo concreto no tiene correspondencia con la distribución, podríamos descartarlo como erróneo o irrelevante. El problema es que no conocemos la distribución de la población y no podemos hacer este tipo de tratamiento. Sólo podemos aproximarnos mediante la estimación de los parámetros correspondientes.

Nos centraremos en los métodos de selección de atributos desarrollados para clasificación. 

Los **métodos de selección de atributos desarrollados para clasificación** pretenden obtener un modelo que, con menos atributos, tenga el mismo poder de clasificación que el modelo que se obtendría a partir del conjunto original de atributos.


Conocido el valor C_i que puede tomar el atributo clase dentro de un conjunto finito de valores de clase $\{C_1, \dots, C_n\}$, se trata de ver, por ejemplo, si dos atributos diferentes, X_1 y X_2 , muestran valores de la media muy iguales o no para un atributo clase determinado. Si, en efecto, son muy iguales, entonces podremos considerar ambos atributos igualmente predictivos para aquella clase, y uno podrá ser sustituido por el otro. En cambio, en el caso de que los valores sean muy diferentes, nos hallamos ante un atributo interesante, ya que permite discriminar mejor si una observación tiene que ir a una clase o a otra. Evidentemente, es necesario extender este tipo de análisis a todos los atributos

por considerar, así como ponderar adecuadamente la importancia de cada uno de éstos.

Aquí mostramos una aproximación sencilla a este problema basada en el análisis de los parámetros estadísticos más corrientes.

Prueba de significación

La prueba de significación compara si un atributo X_i es realmente relevante o no a partir de sus valores.

Una manera de hacerlo, bastante clásica y procedente de la estadística, consiste en realizar un test de significación para sus valores medios en relación con las distintas clases que hay. Si el atributo tiene el mismo valor para cada clase, entonces podríamos considerarlo poco relevante para la clasificación, y viceversa. Simplificaremos la presentación suponiendo un problema sencillo de clasificación en el que sólo puede haber dos clases, la 1 y la 2. El análisis puede extenderse fácilmente a problemas de multclasificación. 


La prueba de significación consiste en hacer una prueba de hipótesis para averiguar si la media del atributo X_i en los casos observados que tienen como etiqueta la clase 1 es igual a la media de los casos correspondientes a la clase 2 o es diferente. En definitiva, se trata de dar una prueba de hipótesis sobre la igualdad o no de la media del atributo para cada clase:

$$est = \sqrt{\frac{\text{var}(X_{i1})}{n_1} + \frac{\text{var}(X_{i2})}{n_2}}$$

donde X_{i1} es el valor del atributo X_i en las observaciones que corresponden a la clase 1; X_{i2} es el equivalente para las observaciones que corresponden a la clase 2; n_1 es el número de casos correspondientes a la clase 1, y n_2 , el de los correspondientes a la clase 2. var es la varianza de X_i .

Comparamos ambos atributos y vemos si su nivel de significación indica realmente que las diferencias corresponden a algo más que a una sencilla variación aleatoria. Si no hay diferencia, se toma la decisión de eliminar uno de los dos atributos.

En caso de tener que comparar más de dos atributos, entonces se puede reiterar esta comparación de atributos dos a dos.

Hay que tener en cuenta que la suposición inicial de este tipo de procedimiento es que los atributos son independientes entre sí. 

Ejemplo de prueba de significación


Aquí podemos ver un ejemplo sencillo de análisis de prueba de significación. Supongamos el atributo *Edad* que presenta los valores siguientes para las clases 1 y 2:

Clase 1	Clase 2
22	23
34	34
23	23
21	56
34	67
23	87
21	56
34	23
65	59
12	77

La media de la edad a la clase 1 es de 28,9, y en la clase 2, de 50,5. Parece que tenemos un posible atributo interesante para discriminar entre los elementos de las clases 1 y 2.

Obtenemos la diferencia entre medias, 21,6. El valor de *est* es 8,3, lo cual nos da un valor de 2,6. Consultando las tablas de significación, podemos comprobar que la diferencia es significativa y, en principio, parece que el atributo *Edad* puede ser relevante para discriminar entre una clase y la otra.


Análisis de grupos de atributos

El método de selección de atributos anterior, como hemos remarcado, sólo sería realmente aplicable sobre atributos independientes entre sí. En la mayoría de los casos nos da una idea aproximada de la relevancia final de cada atributo y da pie a tomar decisiones bastante acertadas. De todos modos, en la realidad la suposición de independencia puede ser un poco excesiva, por lo que habrá que recurrir a otros métodos que no la utilizan. 

Puede suceder que determinados atributos sean útiles cuando los consideramos de forma aislada, pero dejen de serlo en conjunción con otros. En consecuencia, nos interesa detectar qué agrupaciones de atributos X_1, \dots, X_k son realmente relevantes y si lo siguen siendo cuando dejamos de considerar uno o más atributos X_j , $1 < j < k$.

Muchas veces un grupo de atributos puede ser sustituido por otro atributo (perteneciente al conjunto o no) sin alterar en absoluto la calidad del modelo resultante. Para hacerlo, tenemos que estudiar conjuntamente los grupos de atributos.

Supongamos que trabajamos sobre un conjunto de m atributos, X_1, \dots, X_m . Cada atributo puede considerarse una variable aleatoria. Tomados conjuntamente, los m atributos siguen una distribución de probabilidad conjunta

$P(X_1, \dots, X_m)$. Podemos suponer, en tal caso, que esa distribución de probabilidad sigue una distribución normal multivariante. 

Una distribución de este tipo puede caracterizarse mediante un vector que recoge las medias de todas las variables X y por la matriz de covarianzas de las medias C , que es una matriz $m \times m$.

Cada uno de los elementos c_{ij} de la matriz C refleja la relación con los atributos X_i y X_j según la fórmula siguiente:

$$c_{ij} = \frac{1}{n} \sum_{k=1}^n [(x_{ki} - \bar{x}_i) \cdot (x_{kj} - \bar{x}_j)], \text{ donde:}$$

- n es el número total de observaciones que hay en el conjunto de datos;
- x_{ki} es el valor del atributo x_i en la observación k -ésima;
- x_{kj} es el valor del atributo x_j en la observación k -ésima;
- \bar{x}_i es la media del atributo x_i sobre todo el conjunto de datos;
- \bar{x}_j es la media del atributo x_j sobre todo el conjunto de datos.

Los elementos de la diagonal de la matriz C , c_{ii} , corresponden a las varianzas de cada atributo i . Los términos externos a la diagonal de la matriz, c_{ij} con $i \neq j$, indican la correlación entre los atributos correspondientes x_i y x_j . A partir de la información de las correlaciones y varianzas se puede detectar la existencia de atributos redundantes.

En efecto, supongamos que volvemos a encontrarnos en una situación de clasificación que admite dos clases, 1 y 2. Nos interesa volver a encontrar si hay diferencias significativas entre las características de los atributos con respecto a una clase y la otra. La diferencia ahora es que, en lugar de hacer esta prueba atributo por atributo o de dos en dos, como no suponemos independencia entre los atributos, establecemos una prueba simultáneamente sobre todas las medias del conjunto de los atributos.

Dada la clase 1, podemos indicar por \bar{X}_{c1} el vector que contiene las medias de todos los atributos calculadas a partir de las observaciones correspondientes a la clase 1. Indicamos por \bar{X}_{c2} el vector correspondiente para la clase 2.

Nuestra intuición nos dice que, si estos dos vectores son muy parecidos entre sí, entonces el grupo de n atributos que consideramos no es demasiado relevante para discriminar entre una clase y la otra.

Para establecer la similitud entre los dos vectores podemos recurrir de nuevo al concepto de **distancia**, aunque en esta ocasión su expresión ha de tener en cuenta las características del espacio donde calculamos estas distancias. Así, extendemos la prueba de significación entre las diversas medias de todos los atributos simultáneamente. Definimos la distancia siguiente:

$$Dist = (\bar{X}_{c1} - \bar{X}_{c2})(C_1 + C_2)^{-1}(\bar{X}_{c1} + \bar{X}_{c2})^T$$

donde:

$$(\bar{X}_{c1} + \bar{X}_{c2})^T$$

el vector invertido del resultado de sumar los vectores de medias para la clase 1 y 2, y:

$$(C_1 + C_2)^{-1}$$

es la inversa de la suma de la matriz de covarianzas.

Si los atributos son independientes entre sí, entonces todos los elementos que no se encuentren en la diagonal de la matriz inversa de covarianzas son cero y los valores de la diagonal son:

$$\frac{1}{var_{xi}}$$


para cada atributo X_i . Tenemos que encontrar aquellos atributos para los cuales se maximice la distancia, es decir, aquéllos para los que cada uno de los componentes del vector muestre los valores máximos. Cada componente del vector de distancias es el valor dado por la expresión:

$$\frac{(\bar{X}_{iC1} - \bar{X}_{iC2})^2}{var_{xiC1} + var_{xiC2}}$$

donde:

\bar{X}_{ic2} es la media de X_i para la clase C_1 (análogamente para C_2).

var_{xiCi} es la varianza de X_i para la clase C_1 (análogamente para C_2).

Por lo tanto, el problema consiste en encontrar un conjunto de k atributos con k inferior al número de atributos total presente en el dominio m tales que el valor correspondiente a esta fórmula quede maximizado. Bien, para encontrarlo hay que hacer lo siguiente: 

1) Encontrar los subconjuntos de k atributos dentro del conjunto de m atributos inicial.

2) Evaluarlos.

3) Quedarse con el subconjunto que muestre un valor de *Dist* mayor.

Evidentemente, el número de combinaciones es muy alto, pero hay algoritmos de optimización que permiten afrontarlo.

Una alternativa interesante a lo anterior consiste en empezar con un único atributo e ir añadiendo atributos midiendo el valor de *Dist*. La idea se basa en ir incrementando el número de atributos del conjunto inicial de manera que ese valor mejore.

En cada paso de la iteración también hay que ver si es necesario eliminar alguno de los atributos considerados hasta ese momento porque no aporta la suficiente ganancia. La idea consiste en introducir en cada paso los atributos que maximicen la calidad y eliminar los que la minimicen. La condición de final de la iteración es cuando ya no se añade ni elimina ningún atributo, o bien cuando se ha superado la calidad del conjunto total de n atributos en una proporción lo suficientemente significativa (por ejemplo, entre el 75% y el 95%).

A continuación presentamos el esquema del algoritmo asociado a este método de análisis de grupos de atributos. Si tenemos X_1, \dots, X_n atributos que describen un dominio y un umbral de calidad λ , el algoritmo es el siguiente:

```

Seleccionar un atributo  $X_i$ .
Conjunto_Actual =  $\{X_i\}$ .
No_final = Cert
Mientras No_final hacer
    Resta =  $\{X_1, \dots, X_n\} - \text{Conjunto\_Actual}$ 
    Máximo = 0; Atributo_máximo = {}
    Para todo  $X_j \in \text{Resta}$  hacer
        Si Calidad ( $\text{Conjunto\_Actual} \cup \{X_j\}$ ) > Máximo
            entonces Atributo_máximo =  $\{X_j\}$ 
            Máximo = Calidad( $\text{Conjunto\_Actual} \cup \text{Atributo\_máximo}$ )
    fsi
    fpor
        Conjunto_Actual =  $\text{Conjunto\_Actual} \cup \{\text{Atributo\_máximo}\}$ 
        Mínimo = Calidad( $\text{Conjunto\_Actual}$ );
        Para todo  $X_j \in \text{Conjunto\_Actual}$  hacer
            Si Calidad( $\text{Conjunto\_Actual} - \{X_j\}$ ) < Mínimo
                entonces Atributo_mínimo =  $\{X_j\}$ 
                Mínimo = Calidad( $\text{Conjunto\_Actual} - \text{Atributo\_mínimo}$ )
        fsi
        fpor
            Conjunto_Actual =  $\text{Conjunto\_Actual} - \{\text{Atributo\_mínimo}\}$ 
    fMientras
  
```

4.1.2. Fusión y creación de nuevos atributos: análisis de componentes principales

El método de selección de atributos que hemos comentado hasta este punto es el equivalente a eliminar en la forma tabular de los datos toda una columna; el resto de las columnas queda completamente inalterado.

Una alternativa a este método consiste en fusionar atributos, introduciendo, si se desea, atributos “híbridos”, de manera que se creen atributos nuevos que correspondan a valores también nuevos. También podemos influir en el resto de los atributos modificando sus valores.

Dados n atributos, X_1, \dots, X_n , se puede convertir en un atributo nuevo X_{n+1} mediante la aplicación de una combinación lineal de pesos a los atributos X_1, \dots, X_n .

$$X_{n+1} = \sum_{i=1}^n X_i \cdot w_i$$

donde w_i , $1 < i < n$ es un peso, un factor numérico, que indica en qué grado el atributo X_i contribuye al nuevo atributo.

Evidentemente, la combinación lineal de todos los atributos nos conduciría a una reducción extrema de atributos en la cual n atributos quedarían reemplazados por uno solo, el nuevo atributo X_{n+1} . Esto, en general, no nos garantizará comportamientos correctos.

El **método de los componentes principales** efectúa hasta n transformaciones. En cada etapa se realiza una única transformación. En cada una de estas etapas o transformaciones se aplica un vector \mathbf{W} de pesos diferente al que actúa como componente principal. Cuando ya hemos obtenido n transformaciones, entonces se eligen aquellas k de mejor calidad.

Suponiendo un conjunto de datos original definido como un vector con tantos elementos como atributos tiene el dominio \mathbf{D} , podemos establecer la relación entre este conjunto original y el que resulta tras haber aplicado i transformaciones, \mathbf{D}_i , mediante una matriz \mathbf{P} que recoge los distintos componentes principales. En efecto, podemos ver que se cumpliría la siguiente relación:

$$\mathbf{D} = \mathbf{D}\mathbf{P}$$

donde cada columna de \mathbf{P} es un componente principal formado por m pesos. El resultado de multiplicar en la transformación i -ésima la observación k -ésima

Ejemplo de selección de atributos

Si decidimos que, en el caso de Hyper-Gym, para una determinada operación de construcción de modelos el atributo *Renta* resulta irrelevante, lo eliminaremos, pero no introduciremos ningún nuevo atributo ni modificaremos los valores de los atributos que permanezcan en la base de datos.

ma por la columna j -ésima de la matriz \mathbf{P} es el valor del nuevo atributo j para la observación k -ésima.


Para encontrar los pesos de los componentes principales, primero se normalizan los datos por estandarización. El primer componente principal es la línea que se ajusta mejor a los datos, algo que podemos medir utilizando la distancia euclídea mínima:

$$D_{eu} = \sum_{i,j} (D_{ij} - w_j D_{ij})$$

Se puede demostrar que el atributo que se ha generado en esta transformación es el que presenta varianza máxima. En principio, parece que los atributos que tienen esta propiedad deberían ser buenos candidatos para separar clases.

¿Qué criterios hay para seleccionar estas transformaciones?

4.2. Métodos de reducción de casos

Otra manera de reducir la dimensionalidad consiste en eliminar el número de casos por considerar. Este problema ya ha sido afrontado tradicionalmente desde la estadística y el análisis de datos. 

El **método de reducción de casos** consiste en encontrar una muestra, un subconjunto del conjunto original de casos, que muestre un comportamiento parecido.

El **muestreo aleatorio** intenta extraer el subconjunto de casos adecuado. Consideraremos dos métodos de muestreo:

- a) Obtención incremental de la muestra. Se trata de elegir un conjunto inicial de muestra e ir añadiendo casos al mismo hasta que se obtiene un nivel de calidad aceptable prefijado.
- b) Consenso de muestras. Consiste en obtener varios modelos para N conjuntos de muestras de dimensiones relativamente pequeñas e intentar fusionar los modelos obtenidos.


Es importante tener en cuenta que ambos métodos requieren la aplicación del programa de construcción de modelos de que se trate.

Obtención incremental de muestras

La obtención incremental de muestras consiste en empezar por un conjunto pequeño de casos seleccionados aleatoriamente; por ejemplo, empezar con un subconjunto inicial correspondiente al 10% de los casos. Entonces, hay que evaluar el proceso e ir repitiéndolo con porcentajes de datos progresivamente mayores.

Consenso de muestras

El consenso de muestras consiste en dividir los N casos en un número determinado de muestras. A cada una se le aplica el método que buscamos y, finalmente, un algoritmo de consenso.

Curiosamente, con este método se obtienen mejores resultados que aplicando el método a todos los casos. 

5. Tratamiento de la falta de datos

Uno de los problemas más habituales en el tratamiento previo de los datos es la ausencia de valores para un atributo determinado.

Dos opciones típicas son las siguientes:

- Sustituir el valor que falta por la media de los valores que presenta el atributo en los datos.
- Sustituirlo por el valor más frecuente.


Ejemplos típicos de tratamiento de falta de datos

En el caso de las edades de los clientes de Hyper-Gym podríamos encontrar esta situación:

22	43	31	22	?	22	45	43	23	88
----	----	----	----	---	----	----	----	----	----

Las dos soluciones típicas serían:


- Sustituir el valor que falta por la media de los valores que tenemos; en este caso sería 37,3.
- Sustituir el valor que falta por el más frecuente; en este caso sería 22.

Cada alternativa puede llegar a dar resultados diferentes, especialmente cuando el número de valores que faltan es elevado. En general, estos métodos demasiado sencillos llevan a introducir sesgo en los datos. 

Acostumbra a suceder que los casos para los cuales no se ha recogido una observación reciben algún valor especial, por ejemplo -1 ó 999. Es importante no tratarlos como un valor numérico más, sino como lo que realmente son: ausencia de información con respecto a un atributo en un caso observado.

Otras soluciones incluyen estos puntos:

- No tratar aquellos atributos que presentan un número elevado de valores no observados.
- No tratar los casos que presentan un atributo no observado.
- No tratar los casos que presentan más de un atributo no observado.

También es importante saber si se trata de falta de observación o bien que aquel atributo no se considera significativo o relevante. En el segundo caso, no hay que tratar el atributo en absoluto. 

Resumen

La **fase de preparación de datos** introduce modificaciones sobre los valores presentes en el conjunto de datos principalmente por dos motivos:

- 1) Exigencias de formato de los métodos de *data mining*, como el caso de las redes neuronales (valores de entrada entre 0 y 1) o de ciertos métodos de aprendizaje que sólo admiten determinados tipos de variables.
- 2) Necesidad de simplificación de los cálculos que tiene que efectuar el método que se aplique, reduciendo el número de valores de los atributos, o el conjunto de atributos iniciales o bien el conjunto de casos por tratar.

Hemos visto las características de los atributos en función de la escala, el orden y el tamaño de su rango. Existen varios **tipos de transformaciones** en función de la división en variables continuas, discretas, ordinales y nominales.

Se han descrito las **normalizaciones**, que llevan los valores sobre un rango estandarizado y comparable, y también las discretizaciones, que reducen el número de valores de los atributos partiéndolos en varios intervalos que definen puntos de corte sobre los que se pueden efectuar comparaciones.

Los **métodos de discretización** se pueden aplicar de manera general, sin conocer qué tipo de métodos de *data mining* se aplicarán sobre los datos discretizados. Este tipo de métodos reciben el nombre de **métodos no supervisados**.

Por el contrario, los **métodos supervisados** parten del conocimiento de qué tipo de tarea hay que realizar con el modelo que resulte de tratar los datos discretizados. Hemos revisado dos métodos supervisados para discretizar datos para clasificación: *k-means* y métodos basados en entropía de clase. Estos últimos nos han permitido introducir tres conceptos importantes. Veámoslos:

a) Por una parte, el concepto de **medida de distancia**, que también utilizamos al hablar de agregación.

Podéis ver la agregación en el módulo "Agregación" de esta asignatura.


b) Por otra parte, la **entropía** como medida de la homogeneidad de una clase (que volvemos a encontrar cuando hablamos de árboles de decisión).

Podéis ver los árboles de decisión en el módulo "Clasificación: árboles de decisión" de la presente asignatura.

c) Finalmente, también hemos presentado el **principio de la mínima longitud de descripción** (que se aplica de manera general para la construcción y evaluación de muchos tipos de modelos para varias tareas).

También hemos presentado métodos para reducir el número de atributos que hay que utilizar. Hemos aprendido a detectar atributos irrelevantes dos a dos y grupos de atributos que pueden ser sustituidos por uno solo.

Y ya por último, hemos comentado la posibilidad de utilizar un conjunto de datos menor que el conjunto original, a fin de evitar el tratamiento de volúmenes de observaciones demasiado grandes. Tan sólo hemos mencionado la posibilidad de conseguir muestras de forma incremental y de consensuar varias muestras.

Todos estos cambios, en definitiva, nos obligan a asegurarnos de que el modelo final mantiene unas características de calidad buenas. 

Actividades

1. Comparad los distintos métodos de discretización que ofrecen los sistemas:

- WEKA
- SIPINA
- Cviz

¿Qué limitaciones imponen?

2. Consultad el artículo original de Fayyad e Irani para ver cómo se desarrolla la prueba de que el criterio MDL es apropiado para dar una condición de final a la discretización por entropía de clase.

3. Con respecto a la primera propuesta de proyecto que habíais efectuado en las actividades del módulo anterior:

- Estudiad qué necesidades de transformación de datos podrían darse.
- Estudiad qué necesidades de discretización podrían darse.
- Aportad al foro algún conjunto de datos que necesitéis para llevar a cabo el proyecto, y comentad qué métodos de normalización y discretización aplicaríais.
- ¿Habéis encontrado algún atributo irrelevante?

4. Comparad vuestras propuestas y dificultades con las de vuestros compañeros y compañeras; discutid las ventajas e inconvenientes que hayáis encontrado.

Lectura recomendada

Encontraréis el artículo original de Fayyad e Irani en la obra siguiente:
U.M. Fayyad; K.B. Irani (1993). "Multi-interval Discretization of Continuous Valued Attributes for Classification Learning". *Proceedings of the 13th International Joint Conference on Artificial Intelligence* (pág. 1022-1027).

Podéis ver la actividad 1 del módulo "El proceso de descubrimiento de conocimiento" de esta asignatura.



Ejercicios de autoevaluación

1. Clasificad los tipos de variables siguientes en continuas, ordinales, discretas o nominales:

- Horario de Hyper-Gym
- Renta de los socios de Hyper-Gym
- Sexo de los socios de Hyper-Gym
- Profesión de los socios de Hyper-Gym
- Día de la semana
- Años de pertenencia al club

2. Normalizad, si podéis, por el máximo, por diferencia y por escala decimal los valores siguientes:

- Valores de X : -999; 985,34; 234,06 y 2.002,999, donde X puede tomar valores reales entre 1.000 y 2.003 bajo el 1.000.
- Valores de X : 1; 224; 30.000 y 154, donde X puede adoptar valores enteros entre 0 y 150.000. Tenéis que conseguir que los valores resultantes estén entre -1 y 1.
- Valores de la variable X que recoge valores enteros correspondientes a las rentas de los miembros del club Hyper-Gym. El valor mínimo que puede tomar es 0, y el valor máximo, 10.000.000. El rango final tiene que encontrarse entre 0 y 1 (valores reales).

3. Discretizad los conjuntos de valores siguientes para obtener tres intervalos:

Variable 1	Variable 2
6.000.000,0	40,0
0,0	30,0
0,0	32,0
0,0	32,0
4.000.000,0	30,0
1.200.000,0	55,0
0,0	32,0
3.500.000,0	40,0
2.500.000,0	60,0
4.000.000,0	41,0

Variable 1	Variable 2
3.500.000,0	36,0
1.800.000,0	59,0
3.200.000,0	33,0
5.000.000,0	34,0
4.000.000,0	34,0

- Por el método de igual amplitud de intervalo.
- Por el método de igual frecuencia.
- Por el método de la distribución normal.
- Por el método *k-means*.
- Comparad el resultado:
 - Sin transformar los datos.
 - Normalizando los valores de cada variable para que queden en el rango (0, 1).
 - Con los resultados que obtengáis con una discretización con cinco intervalos.

3. Discretizad este conjunto de datos, que es una muestra de la base de datos IRIS del repositorio de la UCI:

Encontraréis los datos utilizados en este ejemplo en la dirección siguiente:
<http://www.gmd.de/mlarchive/frames/datasets/datasets-frames.html>

Clase	Longitud del sépalo	Anchura del sépalo	Longitud del pétalo	Anchura del pétalo
<i>Iris-setosa</i>	5,1	3,5	1,4	0,2
<i>Iris-setosa</i>	4,9	3	1,4	0,2
<i>Iris-setosa</i>	4,7	3,2	1,3	0,2
<i>Iris-setosa</i>	4,6	3,1	1,5	0,2
<i>Iris-setosa</i>	5	3,6	1,4	0,2
<i>Iris-setosa</i>	5,4	3,9	1,7	0,4
<i>Iris-setosa</i>	4,6	3,4	1,4	0,3
<i>Iris-virginica</i>	6,5	3	5,8	2,2
<i>Iris-virginica</i>	7,6	3	6,6	2,1
<i>Iris-virginica</i>	4,9	2,5	4,5	1,7
<i>Iris-virginica</i>	7,3	2,9	6,3	1,8
<i>Iris-virginica</i>	6,7	2,5	5,8	1,8
<i>Iris-virginica</i>	7,2	3,6	6,1	2,5
<i>Iris-virginica</i>	6,5	3,2	5,1	2
<i>Iris-setosa</i>	5,8	4	1,2	0,2
<i>Iris-setosa</i>	5,7	4,4	1,5	0,4
<i>Iris-setosa</i>	5,4	3,9	1,3	0,4
<i>Iris-setosa</i>	5,1	3,5	1,4	0,3
<i>Iris-setosa</i>	5,7	3,8	1,7	0,3
<i>Iris-setosa</i>	5,1	3,8	1,5	0,3
<i>Iris-versicolor</i>	6,3	2,3	4,4	1,3
<i>Iris-versicolor</i>	5,6	3	4,1	1,3
<i>Iris-versicolor</i>	5,5	2,5	4	1,3
<i>Iris-versicolor</i>	5,5	2,6	4,4	1,2
<i>Iris-versicolor</i>	6,1	3	4,6	1,4

- Por *chi merge*.
- Por minimización de la entropía de clase.

5. Dados los datos siguientes, determinad los atributos relevantes. Estos datos proceden de una simplificación de la calificación por barrios de la ciudad de Boston, en Estados Unidos, que intenta relacionar varios factores del barrio donde se reside con el nivel de renta alcanzado por una persona.

Nivel de renta (clase)	Índice de criminalidad	Ocupación industrial	Nivel de óxido nítrico en el aire	Número de habitantes por vivienda	Nivel de escolarización
0-10K	20,0849	18,1	0,7	4,368	20,200001
10-20K	0,17004	7,87	0,524	6,004	15,2
20-30K	0,00632	2,31	0,538	6,575	15,3
30-40K	0,02729	7,07	0,469	7,185	17,799999
40-50K	0,08187	2,89	0,445	7,82	18
30-40K	0,03237	2,18	0,458	6,998	18,700001
30-40K	0,06905	2,18	0,458	7,147	18,700001
20-30K	0,02985	2,18	0,458	6,43	18,700001
20-30K	0,08829	7,87	0,524	6,012	15,2
20-30K	0,14455	7,87	0,524	6,172	15,2
10-20K	0,21124	7,87	0,524	5,631	15,2

6. Con los mismos datos del ejercicio anterior, comprobad la relevancia de los atributos dos a dos.

Bibliografía

Anderberg, M. (1973). "Cluster Analysis for Applications". *Academic Press*.

Catlett, J. (1991). "On Changing Continuous Attributes into Ordered Discrete Attributes". En: Kodratoff, Y. (ed.). *Proceedings of the European Working Session on Learning: Machine Learning* (págs. 164-178). Springer Verlag.

Dougherty, J.; Kohavi, R.; Sahami, M. (1995). "Supervised and Unsupervised Discretizations of Continuous Features" (págs. 194-202). *Proceedings of the 12th International Conference on Machine Learning*. Morgan Kaufmann Publishers.

Fayyad, U.M.; Irani, K.B. (1993). "Multi-interval Discretization of Continuous Valued Attributes for Classification Learning". *Proceedings of the 13th International Joint Conference on Artificial Intelligence* (págs. 1022-1027).

Hartigan, J.I.; Wong, M.R. (1979). "A k-means Clustering Algorithm", ALGORITHM AS 136. *Applied Statistics* (vol. 28, núm. 1).

Huan Liu, H.; Motoda, H. (1998). *Feature Selection for Knowledge Discovery and Data Mining*. Boston: Kluwer Academic Publishers.

Kerber, R. (1992). "Chimerge: Discretization of Numeric Attributes". *Proceedings of the 10th National Conference on Artificial Intelligence* (págs. 123-127).

Lebart, L.; Morineau, A.; Fenelon, J.P. (1985). *Tratamiento estadístico de datos*. Barcelona: Marcombo.

Martín, J.R. (1999). *Discretització de variables contínues i la seva aplicació*. Proyecto de Final de Carrera, Ingeniería en Informática. Departamento de Lenguajes y Sistemas Informáticos. Barcelona: Facultad de Informática de Barcelona, Universidad Politécnica de Cataluña.

Pyle, D. (1999). *Data Preparation For Data Mining*. Morgan Kaufmann Publishers.

Quinlan, J.R. (1989). "Inferring Decision Trees Using the Minimum Description Length Principle". *Information and Computation* (núm. 80, vol. 3, págs. 227-248).

Rissanen, M. (1985). "The Minimum Description Length Principle". A: Kotz, S.; Johnson, N.L. (eds.). *Encyclopedia of Statistical Sciences* (vol. 5.). Nueva York: John Wiley & Sons.

Simoudis, E.; Fayyad, U.M. (1997). "Data Mining Tutorial". *First International Conference on the Practical Applications fo Knowledge Discovery in Data Bases* (marzo). Londres.

Valdés, J.J. (1997). "Fuzzy Clustering and Multidimensional Signal Processing in Enviromental Studies". *Proceedings of the 5th European Congress on Intelligent Techniques and Soft Computing*. Aache (Alemania).