Preparación de los datos

Jordi Casas Roma
Julià Minguilón Alfonso

1 crédito
xxx/xxxxx/xxx

Índice

Introducción

La tercera parte presenta los principales métodos relacionados con la selección y extracción de atributos, que permiten reducir el conjunto de datos original para un ahorro en espacio y tiempo de cálculo de los modelos empleados posteriormente en el proceso de minería de datos, así como para la mejora de sus capacidades predictivas en algunos casos.

Objetivos

En los materiales didácticos de este módulo encontraremos las herramientas indispensables para asimilar los siguientes objetivos:

1. XXX

1. Introducción

La tercera parte presenta los principales métodos relacionados con la selección y extracción de atributos, que permiten reducir el conjunto de datos original para un ahorro en espacio y tiempo de cálculo de los modelos empleados posteriormente en el proceso de minería de datos, así como para la mejora de sus capacidades predictivas en algunos casos.

Recordar notació de dades D

2. Conceptos preliminares

En este capítulo revisaremos algunos conceptos matemáticos que son especialmente útiles en los algoritmos de minería de datos, y que por lo tanto, es interesante conocer antes de iniciar el estudio de los diferentes métodos o algoritmos.

En primer lugar, veremos muy brevemente, los conceptos más básicos del análisis estadístico, en la Sección 2.1. A continuación, en la Sección 2.2 repasaremos algunas de las métricas de distancia o similitud más empleadas por los algoritmos de minería de datos. Finalizaremos este repaso de conceptos preliminares con una breve reseña de la entropía y la ganancia de información, en la Sección 2.3.

2.1. Análisis estadístico

La estadística mira a los datos y trata de analizarlos, por este motivo la minería de datos utiliza tantos conceptos que provienen de esta campo de conocimiento. La recogida de muestras, la exploración de los datos, la inferencia estadística y la búsqueda de patrones forman parte de ambas disciplinas.

2.1.1. La distribución de una muestra

A continuación veremos las bases estadísticas que pueden ayudarnos en la minería de datos, centrándonos en el concepto de distribución con el objetivo de introducirnos en el ámbito de conocimiento de la estadística muestral.

Principales estimadores

La estadística entiende los datos como **observaciones** de una **variable**. Generalmente, se suele trabajar con el alfabeto griego y latino[1] más que con números, de modo que se suele representar una variable con la mayúscula X y sus observaciones con la minúscula $X = \{x_1, x_2, \ldots, x_n\}$.

La **distribución de una variable** son las propiedades de los valores observados, como por ejemplo, los valores mayor y menor o los valores con mayor o menor frecuencia de aparición. Un aspecto interesante en una distribución es el estudio de

[1] Normalmente se utiliza el alfabeto griego para hacer referencia a la población y el latino para la muestra.

su zona central. Es decir, los valores que aproximadamente tienen la mitad de las observaciones por debajo (con menos frecuencia) y la mitad por arriba (con más frecuencia). Aquí es donde encontramos los dos estimadores más básicos: la mediana y la media aritmética.

La **mediana** (*median*) corresponde al valor central de la distribución, que podemos expresar como

$$M_e = x_{\frac{n+1}{2}} \tag{1}$$

donde n es el número total de observaciones.

La **media aritmética** (*arithmetic mean*) de un conjunto de datos numéricos es su valor medio, representado por:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

Tanto la mediana como la media aritmética miden el centro de una distribución, pero lo hacen de forma distinta. Sólo cuando la distribución es simétrica, ambos valores coinciden. Los valores asimétricos y alejados atraen la media, mientras que no afectan a la mediana. Esto hace que la media por sí sola no sea un estimador suficiente para describir la distribución de una variable.

La media aritmética, en ocasiones, puede llegar a ser muy poco representativa del juego de datos. Por este motivo suele ir acompañada de la varianza, donde una varianza pequeña indica que el juego de datos es compacto y en consecuencia la media será muy representativa.

La **varianza**, entendida como la suma de los cuadrados de las distancias a la media, permite medir la dispersión alrededor de la media aritmética. Se calcula de la siguiente forma:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \tag{3}$$

La **desviación estándar** es simplemente la raíz cuadrada de la varianza, de modo que nos da una medida de la dispersión entorno a la media, pero expresada en la misma escala que la variable. Su fórmula de cálculo es:

$$s = \sqrt{s^2} \tag{4}$$

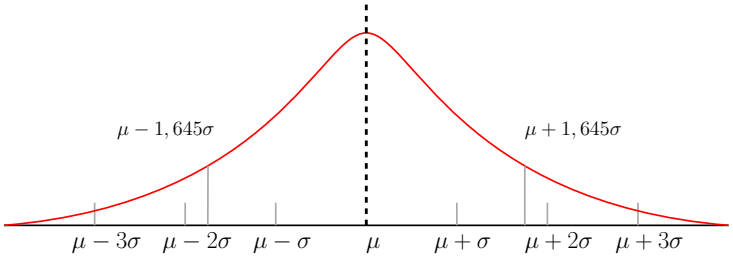
La importancia de la desviación estándar, como medida de dispersión de la distribución de una variable, radica en el hecho de que "la mayoría"[2] de las desviaciones respecto de la media caen dentro de un intervalo igual a 1 o 2 veces la desviación estándar, es decir, $(\overline{x} - s, \overline{x} + s)$ o $(\overline{x} - 2s, \overline{x} + 2s)$, como se puede ver en la Figura 1.

La distribución normal

La **curva de densidad** es la representación gráfica de la distribución de una variable suponiendo que fuéramos capaces de obtener muchísimas observaciones. Una **distribución normal**[3] cumple la propiedad de que su curva de densidad es aproximadamente simétrica, como se puede ver en la Figura 1, y por lo tanto la mediana y la media aritmética coinciden en su centro. La fórmula matemática de la distribución normal viene dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{5}$$

Figura 1. Curva de densidad de una distribución normal



Debido a las propiedades especiales que posee la distribución normal, reservamos letras específicas para referirnos a su media y a su desviación estándar:

- La letra μ se utiliza para referirse a la media aritmética de una distribución normal.

[2] El concepto de "la mayoría" lo definiremos con el estudio de la distribución normal.

[3] La distribución normal también es conocida como distribución de Gauss o distribución gaussiana.

- La letra σ se utiliza para referirse a la desviación estándar de una distribución normal.

Si nos fijamos en estos intervalos, sucede que:

- En toda distribución normal, el área debajo de la curva para el intervalo de una desviación estándar es de 0.68. Es decir, en el intervalo $(\mu-\sigma, \mu+\sigma)$ encontraremos el 68 % de la población.
- En el intervalo $(\mu - 1{,}645\sigma, \mu + 1{,}645\sigma)$ encontramos el 90 % de la población.
- En el intervalo $(\mu - 2\sigma, \mu + 2\sigma)$ encontraremos el 95 % de la población.
- Y finalmente, en el intervalo $(\mu - 3\sigma, \mu + 3\sigma)$ encontraremos el 99,7 % de la población.

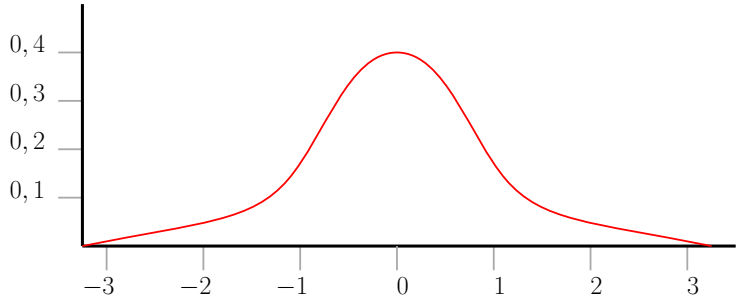
La distribución normal estándar

La **distribución normal estándar** es una distribución normal que tiene una media aritmética con valor 0 y una desviación estándar con valor igual a 1.

Es interesante, dada una variable, poder **estandarizarla**, es decir, realizar los cálculos de probabilidades (áreas debajo de la curva de densidad) necesarios y regresar de nuevo a la variable original con la traducción de las probabilidades calculadas. Para realizar este proceso, partiremos de una variable X normalmente distribuida, con media aritmética μ y desviación estándar σ, y la estandarizaremos restando a cada observación la media aritmética y dividiendo el resultado por la desviación estándar. El resultado será la variable **variable normal estandarizada**, que se calcula de la siguiente forma:

$$Z = \frac{X - \mu}{\sigma} \tag{6}$$

Figura 2. Curva de densidad de una distribución normal estándar



Si llamamos P a la probabilidad o área debajo de la curva de densidad, viendo la Figura 2 y consultando las tablas de densidad de la distribución normal estándar sabemos que, al ser una distribución simétrica centrada en el cero se cumple que:

$$P(Z < -1) = P(Z > 1) = 0{,}1587 \tag{7}$$

y por deducción, como todo el área bajo la curva de dicha distribución es 1, podemos calcular la probabilidad de que un valor de Z se encuentre en la zona central, que corresponde a:

$$\begin{aligned} P(-1 < Z < 1) &= 1 - P(Z < -1) - P(Z > 1) \\ &= 1 - 2 * 0{,}1587 = 0{,}6826 \end{aligned} \tag{8}$$

Para realizar el camino inverso, es decir, de la distribución normal estándar a la distribución normal original, usaremos la siguiente expresión:

$$X = \mu + Z\sigma \tag{9}$$

2.2. Métricas de distancias o similitud

Antes de adentrarnos en la sección de algoritmos, nos centraremos en los conceptos de distancia y similitud (o disimilitud) puesto que éstos constituyen la base sobre la que se construyen muchos los algoritmos de segmentación y clasificación.

La elección de una métrica apropiada para la distancia tiene mucho impacto en el resultado final de cualquier algoritmo de segmentación. Es más, la elección de la métrica de distancia adecuada suele ser uno de los aspectos clave en la construcción de este tipo de algoritmos.

Cuando hablamos de distancia, en realidad nos estamos refiriendo a una forma de cuantificar cuán similares (o disimilares) son dos objetos, variables o puntos.

Planteado de esta forma, el concepto de distancia es muy abstracto y etéreo, por este

motivo los científicos quieren ponerle algunos límites o condiciones. Para un conjunto de elementos X se considera distancia a cualquier función $f : X \times X \to \mathbb{R}$ que cumpla las siguientes tres condiciones:

- **No negatividad**: La distancia entre dos puntos siempre debe ser positiva.

$$d(a, b) \geq 0 \quad \forall a, b \in X \tag{10}$$

- **Simetría**: La distancia entre un punto a y un punto b debe ser igual a la distancia entre el punto b y el punto a.

$$d(a, b) = d(b, a) \quad \forall a, b \in X \tag{11}$$

- **Desigualdad triangular**: La distancia debe coincidir con la idea intuitiva que tenemos de ella en cuanto a que es el camino más corto entre dos puntos.

$$d(a, c) \leq d(a, b) + d(b, c) \quad \forall a, b, c \in X \tag{12}$$

Veamos, a continuación, algunas de las métricas de distancia o similitud más empleadas por los algoritmos de minería de datos.

Distancia euclídea

La distancia euclídea o euclidiana és una de las distancias más utilizadas, especialmente en el caso de tratar con atributos numéricos. La distancia euclídea coincide plenamente con lo que nuestra intuición entiende por "distancia". Su expresión matemática para un espacio de m dimensiones es:

$$d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \tag{13}$$

Uno de sus principales inconvenientes se presenta al utilizar esta distancia en un proceso de segmentación, donde este cálculo de la distancia no tiene en cuenta las diferentes unidades de medida en las que pueden estar expresadas la variables x e y. Por ejemplo, si estamos segmentado familias o razas de ratones, donde la variable x representa la edad de los ratones y la variable y representa la longitud de la cola

del ratón (en milímetros), entonces la variable x tomará valores de en el rango $[0, 3]$, mientras que la variable y utilizará el rango $[90, 120]$. Parece claro que cuando calculemos la distancia o similitud entre dos individuos, pesará injustamente mucho más la longitud de la cola que la edad.

Distancia estadística o de Gauss

Para superar la distorsión provocada por las diferentes unidades de medida usadas en las distintas variables, tenemos la distancia estadística, que simplemente "normaliza" la variables para situarlas todas bajo la misma escala.

Su expresión analítica para un espacio de m dimensiones, en la que nuestra distribución de puntos en estas dimensiones tuviera una desviación estándar σ viene dada por la fórmula:

$$d(x, y) = \sqrt{\sum_{i=1}^{m} \left(\frac{x_i - y_i}{\sigma_i} \right)^2} \tag{14}$$

Nuevamente, este concepto de distancia sigue teniendo problemas. No tiene en cuenta la correlación entre las variables, es decir, si nuestras variables de trabajo fueran totalmente independientes, no habría ningún problema, pero si tienen algún tipo de correlación, entonces una influye sobre la otra y esta "influencia" no queda bien reflejada si usamos la distancia estadística.

Por ejemplo, las variables *entrenar y rendimiento* están correlacionadas, de modo que más entrenamiento siempre implica más rendimiento, pero esta regla no se cumple infinitamente, ya que entrenamiento infinito no implica rendimiento infinito (lo vemos continuamente en el deporte). Si no tenemos en cuenta esta correlación y queremos comparar a dos deportistas (medir su distancia) podemos llegar a conclusiones erróneas.

Distancia de Mahalanobis

Esta métrica pretende corregir la distorsión provocada por la correlación de las variables mediante la siguiente expresión para un espacio de m dimensiones es:

$$d(x,y) = \sqrt{(x_1 - y_1, \ldots, x_m - y_m) \begin{pmatrix} Cov(x_1,y_1) & \ldots & Cov(x_1,y_n) \\ \vdots & \ddots & \vdots \\ Cov(x_n,y_1) & \ldots & Cov(x_n,y_n) \end{pmatrix}^{-1} \begin{pmatrix} x_1 - y_1 \\ \vdots \\ x_m - y_m \end{pmatrix}} \tag{15}$$

donde $Cov(x, y)$ representa la covarianza entre las variables x e y.

La distancia de Mahalanobis es una generalización de la distancia de Gauss, la cual, a su vez, es una generalización de la distancia Euclídea:

- Si en la distribución de puntos, las m dimensiones son totalmente independientes, tendremos que su covarianza es 0, de manera que la distancia de Mahalanobis coincide con la distancia estadística.

- Si en la misma distribución con variables independientes tenemos que su distribución está normalizada, es decir, que las dos variables tienen la misma escala, entonces su varianza es 1, de manera que la distancia estadística coincide con la distancia euclídea.

Distancia de Hamming

Hasta ahora hemos revisado algunas de las distancia más importantes para lidiar con valores numéricos. En el caso de tratar con atributos nominales o binarios, una de las métricas más utilizadas es la distancia de Hamming. Esta métrica se define como el número de atributos distintos que existen entre dos vectores de atributos.

$$d(x,y) = \sum_{i=1}^{m} \delta(x_i, y_i) \tag{16}$$

donde $\delta(x_i, y_i)$ toma el valor 0 si $x_i = y_i$ y 1 en caso contrario.

El lector interesado en otras métricas posibles adaptadas a la naturaleza de los datos puede consultar el trabajo de Cha [**?**], por ejemplo.

2.3. Entropía y ganancia de información

En la base conceptual de varios métodos de minería de datos se encuentra el concepto de ganancia de la información, como hilo heurístico que puede guiar el proceso de aprendizaje.

Empezaremos por introducir la idea de **entropía** (*entropy*) como medida de cómo de predecible es un resultado en un juego de datos. También puede ser pensada como el grado de desorden o de incertidumbre presente en un juego de datos. Su expresión matemática es la siguiente:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{17}$$

donde X es una variable aleatoria discreta con un conjunto de posibles valores $X = \{x_1, \ldots, x_n\}$ y $p(x_i)$ es la probabilidad de que la variable X tome el valor x_i.

Valores de entropía altos indican que el resultado es muy aleatorio y en consecuencia, poco predecible. Veamos, como ejemplo, los siguientes experimentos:

- Tirar un dado con 6 caras al aire puede darnos 6 posibles resultados, $X \in \{1, 2, 3, 4, 5, 6\}$. La entropía de este experimento es:

$$H(X) = -6 \cdot \left(\frac{1}{6} \log_2 \left(\frac{1}{6} \right) \right) \approx 2{,}58$$

- Tirar una moneda al aire puede darnos 2 posibles resultados, $X \in \{cara, cruz\}$. La entropía de este experimento es:

$$H(X) = -2 \cdot \left(\frac{1}{2} \log_2 \left(\frac{1}{2} \right) \right) = 1$$

Observamos como la entropía de lanzar un dado con 6 caras es mucho más alta y, por lo tanto, mucho más aleatoria que la entropía de lanzar una moneda.

La **ganancia de información** (*information gain*) nos da una medida de cómo de relevante es un atributo dentro de un juego de datos, de modo que un atributo con mucha ganancia será muy relevante en el juego de datos, es decir, muy determinante para predecir el atributo objetivo o clase.

La ganancia de información refleja el cambio en la entropía del juego de datos cuando

tomamos parte de la información como dada. Dicho de otro modo, la ganancia de información del atributo a respecto del atributo objetivo x es la diferencia entre la entropía del atributo objetivo y la entropía del atributo a:

$$IG(X, a) = H(X) - H(X, a) \tag{18}$$

$$H(X, a) = \sum_{v \in valores(a)} p(v) H(\{x \in X \mid valor(x, a) = v\}) \tag{19}$$

donde $H(X, a)$ es la entropía del juego de datos cuando lo particionamos en base al atributo a, o visto de otro modo, es la entropía del atributo a en el juego de datos.

En la expresión anterior, el valor $H(\{x \in X \mid valor(x, a) = v\})$ es la entropía del juego de datos cuando fijamos el atributo a al valor v, o visto de otro modo, es la entropía del atributo a cuando toma el valor v en el juego de datos X.

3. Preparación de los datos

Las tareas o técnicas de preparación de datos en un proyecto de minería de datos se orientan a la adecuación del juego de datos para que pueda ser usado, posteriormente, por algoritmos de clasificación, segmentación o regresión.

Estas tareas las podemos agrupar en los siguientes bloques temáticos:

- Tareas de **limpieza de datos**, que permiten corregir o eliminar ruido o datos no válidos.
- Tareas de **normalización de datos**, que facilita la presentación de los datos en el mismo rango.
- Tareas de **discretización**, entendidas como procesos de conversión de variables continuas a categóricas.
- Tareas de **reducción de la dimensionalidad**, que nos ayudará a desarrollar modelos con juegos de datos reducidos.

3.1. Limpieza de datos

En el proceso de limpieza de datos (en inglés, *data cleansing* o *data scrubbing*) se llevan a cabo actividades de detección, eliminación o corrección de instancias corrompidas o inapropiadas en los juegos de datos.

A nivel de valores de atributos se gestionan los valores ausentes, los erróneos y los inconsistentes. Un ejemplo podrían ser los valores fuera de rango *(outliers)*.

El proceso de integración de datos puede ser una de las principales fuentes de incoherencias en los datos, fruto de la fusión de juegos de datos distintos se pueden generar inconsistencias que deben ser detectadas y subsanadas.

3.2. Normalización de datos

La normalización de datos consiste en modificar los datos para lograr que estén en una escala de valores equivalentes que simplifique la comparación entre ellos. La normalización es útil para varios métodos de minería de datos, que tienden a quedar sesgados por la influencia de los atributos con valores más altos, distorsionando de

esta forma el resultado del modelo.

Algunos de los principales métodos de normalización de atributos son:

- **Normalización por el máximo**, que consiste en encontrar el valor máximo del atributo a normalizar y dividir el resto de los valores por este valor máximo, asegurándonos, por tanto, que el máximo recibe el valor 1 y el resto valores en el rango $[0, 1]$.

$$z_i = \frac{x_i}{x_{max}} \tag{20}$$

  donde z_i es el valor normalizado, x_i el valor original del atributo y x_{max} el valor máximo para este atributo de todo el conjunto de datos.

- **Normalización por la diferencia**, que intenta compensar el efecto de la distancia del valor que tratamos con respecto al máximo de los valores observados. Su fórmula de cálculo es:

$$z_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{21}$$

  donde x_{min} es el valor mínimo para este atributo en el conjunto de datos.

- **Normalización basada en la desviación estándar**, también llamada estandarización de valores, que asegura que se obtienen valores dentro del rango elegido que tienen como propiedad que su media es el cero y su desviación estándar es uno.

$$z_i = \frac{x_i - \mu}{\sigma} \tag{22}$$

  donde μ representa la media y σ la desviación típica.

3.3. Discretización de datos

La discretización es el proceso mediante el cual los valores de una variable continua se incluyen en contenedores, intervalos o grupos, con el objetivo de que haya un número limitado de estados posibles. Los contenedores se tratarán entonces como si fueran categorías.

La discretización es una tarea habitual en los procesos de minería de datos puesto que muchos algoritmos de clasificación requieren que sus atributos sean discretizados, bien porque solo acepten valores nominales, bien porque trabajar con atributos nominales en lugar de continuos disminuye el coste computacional y acelera el proceso inductivo. Además, la discretización de atributos puede, en ocasiones, mejorar el rendimiento de un clasificador.

Los métodos de discretización se pueden clasificar según tres grandes categorías:

- **Supervisados o no supervisados**. Es decir, métodos que tienen en cuenta los valores del atributo clase o no.

- **Locales o globales**. Métodos limitados a un atributo, o en general, a un subconjunto horizontal (subconjunto de instancias) o vertical (subconjunto de atributos) del conjunto de datos; o bien actuando sobre todos los atributos y todos los datos a la vez.

- **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.

3.3.1. Método de igual amplitud

El método de partición en intervalos de la misma amplitud es un método no supervisado que busca crear grupos con el mismo número de instancias en su interior.

El método consiste en los siguientes pasos:

1) Se parte de un atributo con n valores, el cual queremos discretizar en k intervalos de igual longitud.

2) La longitud de cada intervalo será:

$$\alpha = \frac{x_{max} - x_{min}}{k} \tag{23}$$

donde x_{min} y x_{max} representan el valor mínimo y máximo, respectivamente, que

toman los atributos.

3) Entonces, los intervalos quedan definidos por la expresión:

$$Intervalo_i = x_{min} + i\alpha,\ \forall i \in \{1, \ldots, k\} \tag{24}$$

Algunos de los principales problemas de este método son:

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

2) Puede generar intervalos poco homogéneos, donde se mezclen dentro de un mismo intervalo valores que corresponden a clases diferentes.

3) La elección del parámetro k no es trivial en la mayoría de casos.

3.3.2. Método de igual frecuencia

Este método es similar al método de igual amplitud, pero en este caso se requiere que cada grupo contenga el mismo número de instancias en su interior.

El método consiste en los pasos siguientes:

1) Ordenar los valores del atributo X de menor a mayor: $x_1 \leq \ldots \leq x_n$.

2) Fijar un número de intervalos k.

3) Calcular el valor de la frecuencia de cada intervalo: $f = \frac{n}{k}$.

4) Asignar a cada intervalo los f valores ordenados, de tal forma que cada cada intervalo o grupo contendrá f elementos ordenados, donde el valor máximo del intervalo i será menor o igual que el valor mínimo del intervalo $i + 1$.

El principal inconveniente de este método es que valores muy dispares pueden ir a parar al mismo intervalo, dado que el criterio es mantener la misma frecuencia en todos los intervalos. Por el contrario, el método funciona correctamente cuando se da una distribución uniforme de los valores del atributo a discretizar.

3.3.3. Método *chi merge*

Chi merge es un algoritmo simple que utiliza el estadístico chi-cuadrado para discretizar los atributos numéricos. Se trata de un método de discretización de datos supervisado.

Este método intenta que dentro de cada intervalo aparezcan representadas todas las clases de forma parecida, y muy diferente respecto de otros intervalos. En términos de frecuencias, eso quiere decir que dentro de un intervalo hay que esperar que la frecuencia de aparición de valores de cada clase sea 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 chi-cuadrado 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).

3.3.4. Métodos basados en entropía

Los métodos basados en medidas de entropía parten de la idea de encontrar particiones del conjunto original de datos que sean internamente tan homogéneas como sea posible, y tan diferentes como sea posible con respecto al resto de las particiones.

Estos métodos trabajan sobre variables continuas con el objetivo de crear un punto de decisión (llamado límite o umbral) sobre un atributo determinado, tal que permita alcanzar la mejor separación de los datos en dos subconjuntos disjuntos.

Se pueden definir distintos criterios para escoger la partición "óptima" en cada iteración del algoritmo, pero una de las métricas más empleadas para tal fin es la entropía de la información asociada a cada posible partición. Es decir, se toma como criterio de discretización el valor que minimice la entropía de los subconjuntos generados utilizando ese punto de corte.

3.4. Reducción de la dimensionalidad

En algunos procesos de minería de datos podemos encontrarnos con el problema de que el método de construcción del modelo elegido no puede tratar la cantidad de datos de que disponemos, o simplemente que el tiempo requerido para construir el modelo no es adecuado para nuestras necesidades.

En estos casos, es posible aplicar una serie de operaciones con el fin de facilitar dos objetivos, y asegurando, además, que se mantiene la calidad del modelo resultante:

- la **reducción del número de atributos** a considerar

- y la **reducción del número de casos** que hay que tratar

Si denotamos como $D_{n,m}$ a nuestro conjunto de datos, en el primer caso el objetivo será reducir el valor de m (reducción de columnas), mientras que en el segundo caso el objetivo será reducir el valor de n (reducción de filas).

3.4.1. 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, pero con una reducción en la complejidad temporal y/o de cálculo del método de minería de datos.

Existen dos grandes grupos:

- Los **métodos de selección de atributos**, que consisten en saber qué subconjunto de atributos puede generar un modelo con la misma, o similar, calidad que el modelo obtenido con el conjunto de atributos original. En esencia, se trata de encontrar métricas que indiquen los atributos que pueden ignorarse en la creación del modelo sin sufrir una considerable pérdida de información.

  Algunos métodos realizan este tipo de pruebas a cada atributo de forma individual, como por ejemplo la prueba de significación, que compara si un atributo es relevante o no a partir de sus valores. Otros métodos, en cambio, parten de que la suposición de independencia entre los atributos puede ser un poco excesiva, y analizan las dependencias y correlaciones entre atributos para determinar grupos de atributos.

- La **fusión y creación de nuevos atributos**, que consiste en crear nuevos atributos "híbridos" a partir de la integración o fusión de los atributos existentes. En general, podemos crear un atributo nuevo a_{n+1} mediante la aplicación de una combinación lineal de pesos a los atributos $a_1, \ldots, a_n$:

$$a_{n+1} = \sum_{i=1}^{n} a_i w_i \tag{25}$$

  donde w_i es el peso que indica en qué grado contribuye el atributo a_i en la creación del nuevo atributo.

  El método de análisis de componentes principales (*Principal Component Analysis*, PCA) es uno de los métodos más conocidos para la fusión y creación de nuevos atributos.

3.4.2. Métodos de reducción de casos

En minería de datos no siempre es posible disponer de todos los datos de la población que se estudia. Es en esta circunstancia en la que el concepto de "muestra" se convierte en relevante para el analista. A una muestra de una población se le exige que sus propiedades sean extrapolables a las propiedades de la población.

Llamaremos **espacio muestral** al conjunto de muestras que se extrae de una población. La variable que asocia a cada muestra su probabilidad de extracción, sigue lo que se llama **una distribución muestral** cuyo estudio nos permitirá calcular la probabilidad que se tiene, dada una sola muestra, de acercarse a las propiedades de la población.

Respecto de las técnicas de muestreo, podemos clasificarlas en dos grupos: muestreo probabilístico y muestreo no probabilístico o subjetivo.

Técnicas de muestreo probabilístico

Se trata del conjunto de técnicas más aconsejable, puesto que permiten calcular la probabilidad de extracción de cualquiera de las muestras posibles.

A continuación enumeramos algunas de las más importantes:

- **Muestreo estratificado**: Consiste en dividir la población de estudio en grupos homogéneos respecto de alguna de las características que se quiere medir. Una vez determinados los grupos a éstos se les aplicaría, por ejemplo, una técnica de muestreo sistemática. Casos de grupos pueden ser por sexo, por sector laboral, etc.

- **Muestreo sistemático**: Si el número total de de miembros de nuestra población es N y el número de miembros de la muestra que queremos tomar es n, entonces tendremos que el intervalo de selección será $\frac{N}{n}$. Escogemos al azar un miembro de la población q, y a partir de éste, se escogen los siguientes miembros siguiendo el orden del intervalo de selección. Esto es $\{q + i\frac{N}{n} \mid i = 0, \ldots, n-1\}$.

- **Muestreo por estadios múltiples**: Se divide la población en niveles estratificados, de modo que vamos tomando muestras de los distintos niveles considerados. Por ejemplo, si queremos realizar un estudio de intención de voto, empezaríamos por dividir la población total en regiones, localidades y barrios.

- **Muestreo por conglomerado**: Esta técnica está indicada para poblaciones que ya de forma natural se encuentran divididas en grupos que contienen toda la variabilidad de la población total. Es decir, que el estudio de uno de los grupos es suficiente para extrapolar los resultados al resto de población.

Técnicas de muestreo no probabilístico o subjetivo

Se dan cuando no es posible determinar la probabilidad de extracción de una determinada muestra o incluso cuando ciertos elementos de la población no tienen posibilidades de ser seleccionados (individuos fuera de cobertura).

Como la selección de muestras no es aleatoria tampoco aplica el concepto de estimación de errores de la muestra. Esta circunstancia provoca lo que se conoce como **sesgo de selección**, que es la distorsión provocada por el hecho de que la muestra tiene muchas limitaciones a la hora de representar toda la variabilidad de la población total.

A continuación enumeramos algunas de las más importantes:

- **Muestreo por cuotas**: Podemos pensarlo como el muestreo por estratos, pero ponderado por cuotas. Es decir, se divide la población en estratos definidos por alguna variable de distribución conocida (por ejemplo el sexo). Posteriormente calculamos la proporción de cada estrato respecto de la población total. Y finalmente se multiplica cada proporción por el tamaño de cada muestra. Esta técnica es habitual en sondeos de opinión y estudios de mercado.

- **Muestreo de bola de nieve**: En esta técnica se selecciona inicialmente un pequeño grupo de individuos que nos ayudarán a encontrar más miembros representativos de la población. Está indicada para los casos en que la población se encuentra muy dispersa o incluso escondida.

4. Extracción y selección de atributos

En el proceso de descubrimiento de conocimiento en grandes volúmenes datos es vital escoger las variables y características más adecuadas para presentar al algoritmo de minería de datos. Este problema puede tener diferentes enfoques, entre otros: escoger los mejores atributos de los datos a partir de su análisis preliminar, eliminar los atributos redundantes o que aportan poca información al problema que se desea resolver, o reducir la dimensionalidad de los datos generando nuevos atributos a partir de atributos existentes. En cualquiera de estos casos, el objetivo es reducir el coste espacial y computacional de creación del modelo, permitiendo además la creación de un modelos de similar o mejor calidad.

La **selección de atributos** consiste en escoger únicamente aquellos atributos que son realmente relevantes para el problema a resolver, descartando otros que no aportan información relevante para el problema a resolver. Por otra parte, la **extracción de atributos** se trata de calcular nuevos atributos a partir de los existentes, de tal forma que los nuevos atributos resuman mejor la información que contienen, capturando la naturaleza de la estructura subyacente en los datos.

Existen métodos automáticos para la selección y extracción de atributos, no obstante, ambos métodos también puede hacerse de forma manual. Es importante subrayar que la selección o extracción manual de atributos requiere de un experto en el dominio que analice y escoja los atributos más relevantes para el problema que se intenta resolver en cada caso. Este es un proceso *adhoc* que requiere gran conocimiento del dominio del problema, así como de los datos que se utilizarán para el proceso de minería de datos.

Por ejemplo, el índice de masa corporal, que se define como el peso de una persona en kilogramos dividido por el cuadrado de su altura en metros, informa mejor del grado de obesidad de una persona que las dos variables originales por separado.

Este tipo de conocimiento proviene del contexto o dominio del problema, y no debe ser descartado. No obstante, en la mayoría de casos será necesario recurrir a los métodos automáticos para extraer características de un conjunto de datos.

En este capítulo nos centraremos en los métodos automáticos de selección y extracción de atributos.

4.1. Selección de atributos

Los métodos de selección de características o atributos (*feature selection*) permiten identificar los atributos que aportan información relevante para el proceso de minería de datos, o al revés, los atributos redundantes que no aportan información relevante a este proceso. En ambos casos el objetivo es el mismo, elegir qué subconjunto de atributos es más beneficioso para resolver el problema en cuestión.

Dependiendo de si la selección de características usa o no información del método de clasificación posterior, podemos definir la siguiente taxonomía:

- Los algoritmos filtro (*filter*), donde los atributos o subconjuntos de atributos son evaluados de forma independiente del método de clasificación que se utilizará posteriormente.

- Los algoritmos empotrados (*wrappers*), donde el método de selección de características utiliza el clasificador que se usará posteriormente para evaluar qué característica o subconjunto de características es el más adecuado.

Los algoritmos empotrados, aunque suelen tener un buen rendimiento, pueden tener más tendencia a sobreaprender el conjunto de entrenamiento, perdiendo capacidad de generalización. Los algoritmos de selección de características han sido ámpliamente estudiados, y se han desarrollado multitud de algoritmos, algunos específicos para determinados problemas (Guyon & Elisseeff, 2003).

Lectura complementaria

I. Guyon, A. Elisseeff. (2003). "An Introduction to Variable and Feature Selection". J. Mach.
Learn. Res., Vol. 3:1157-1182.
`http://dl.acm.org/citation.cfm?id=944919`

En primer lugar, veremos brevemente algunos de los métodos de selección de atributos individuales, llamados **algoritmos univariantes**, más empleados para la selección de atributos:

- Selección de máxima relevancia (*Maximum relevance selection*). Utiliza el coeficiente de correlación entre cada característica y los resultados de clasificar un determinado conjunto de entrenamiento, obteniendo una lista ordenada de las características que mejor diferencian los datos.

- Selección basada en la información mutua. Mide la información mutua entre las variables aleatorias que modelan cada característica y las etiquetas de clasificación, escogiendo las características que maximizan esta información mutua.

- Métodos basados en tests estadísticos. Aplicación de tests estadísticos de hipótesis sobre los datos, como por ejemplo el *T-statstic* o el *chi-square*.

En segundo lugar, encontramos los métodos de selección de subconjuntos de atributos, llamados **algoritmos multivariantes**:

- Búsqueda exhaustiva (*Exhaustive search*). Consiste en definir un espacio de búsqueda y evaluar, mediante una función de coste, todas las posibles combinaciones

de atributos. Solo es aplicable a problemas de dimensionalidad pequeña.

- Selección paso a paso (*StepWise selection*). Consiste en iterar un algoritmo en el cual a cada paso o bien se añade al conjunto de atributos seleccionados aquel atributo que aumenta el rendimiento global del conjunto, o bien se elimina aquel atributo que hace que el rendimiento del subconjunto empeore.

- Ramificación y poda (*Branch and bound*). Consiste en aplicar la técnica de búsqueda de *branch and bound* en el espacio de las posibles combinaciones de características. Esta técnica reduce de forma muy notable la búsqueda exhaustiva de la solución.

4.2. Extracción de atributos

El objetivo de la extracción de características es obtener un espacio de dimensionalidad inferior, que preserve al máximo posible los datos útiles y elimine la información redundante. A diferencia de la selección de atributos, en la extracción de atributos se pueden crear nuevos atributos a partir de los existentes en el conjunto de datos inicial.

Las técnicas que se describen en este apartado son conocidas como **técnicas de factorización matricial**, puesto que descomponen una matriz de datos como el producto de matrices más simples. En cualquier caso, la factorización de una matriz no es única, y cada técnica pone de manifiesto aspectos diferentes de la información contenida en los datos originales.

4.2.1. Análisis de Componentes Principales (PCA)

El análisis de componentes principales (*Principal Component Analysis*, PCA) nos puede ayudar a solucionar problemas de reducción de dimensionalidad y extracción de características en nuestros datos de forma automática. El PCA es un algoritmo muy conocido y usado en el análisis de datos, y tiene muchas posibles aplicaciones diferentes. Informalmente, se puede definir como la técnica que intenta conseguir una representación de un conjunto de datos en un espacio de dimensionalidad más reducida, minimizando el error cuadrático cometido.

El algoritmo PCA reduce la dimensionalidad a través de una transformación lineal que escoge un nuevo sistema de coordenadas para el conjunto original de datos (es decir, realiza una rotación del espacio d-dimensional), en el cual la varianza de mayor tamaño del conjunto de datos se recoge en el primer eje (llamado el primer componente principal), la segunda varianza más grande en el segundo eje, y así sucesivamente. Si los datos presentan alguna "inclinación" (en sentido geométrico), los ejes calculados por el PCA siguen dicha inclinación para capturar la misma varianza con un menor volumen del hiperparalelepípedo que contiene todos los datos.

Este algoritmo se basa en la matriz de covarianzas o correlaciones de los datos originales, de forma que calcula los vectores propios de esta matriz y se aplica a los datos originales para conseguir la transformación lineal. Generalmente se utiliza la matriz de correlación cuando los datos no son dimensionalmente homogéneos o el orden de magnitud de las variables aleatorias medidas no es el mismo. La matriz de covarianzas se usará cuando los datos son dimensionalmente homogéneos y presentan valores medios similares.

Podemos expresar el resultado del proceso de extracción de características como el producto matricial:

$$S_{q,n} = P_{q,m} \cdot D^T_{m,n} \tag{26}$$

donde D es el conjunto de datos original, S es la matriz de los resultados (donde $q << m$), y P es la matriz de proyección, que permite realizar la extracción de características. Es importante notar que la matriz de datos se presenta de forma traspuesta a la forma en que la hemos definido en este texto, es decir, las filas representan los atributos y las columnas las instancias.

El objetivo consiste en encontrar la matriz P que reduce la dimensionalidad de los datos, minimizando el error cuadrático cometido en este proceso de reducción de la dimensionalidad.

El proceso para el cálculo de la matriz de soluciones D es el siguiente:

1) Calcular la media de los datos de cada columna de la matriz D, para obtener un conjunto de datos centrados en su origen $\hat{D}$.

2) Calcular la matriz de covarianza C de los datos como:

$$C = \frac{1}{N}\hat{D} \times \hat{D}^T \tag{27}$$

3) Coger como matriz P, los q primeros vectores propios con mayor valor propio asociado de la matriz de covarianza C.

Aplicando el PCA obtenemos una serie de vectores $S = \{s_1, \ldots, s_q, \ldots, s_m\}$ donde m es el número de atributos. En un problema de reducción de dimensionalidad nos quedaríamos con los q primeros vectores que explicasen la mayor parte de la varianza, y que escogemos en el tercer paso del algoritmo detallado anteriormente. Los coeficientes de los vectores asociados a cada componente marcan la importancia de cada

uno de los nuevos atributos, de forma que los valores altos serán más importantes, mientras que podemos prescindir del resto.

4.2.2. Descomposición en Valores Singulares (SVD)

El método de la descomposición en valores singulares (*Singular Value Descomposition*, SVD) es una herramienta de álgebra lineal que permite descomponer una matriz de datos expresándola como la suma ponderada de sus vectores propios.

El resultado de la descomposición factorial de la matriz D de tamaño $m \times n$ se representa mediante:

$$D_{m,n} = U_{m,m} \cdot S_{m,n} \cdot V^T_{n,n} \tag{28}$$

donde S es una matriz diagonal de tamaño $m \times n$ con componentes no negativos y U, V son matrices unitarias de tamaño $m \times m$ y $n \times n$ respectivamente.

A los componentes diagonales no nulos de la matriz S se les conoce como valores singulares, mientras que las m columnas de la matriz U y las n columnas de la matriz V se denominan vectores singulares por la izquierda y por la derecha, respectivamente. El número máximo de valores singulares diferentes de la matriz D está limitado por el rango máximo de dicha matriz, $r = min\{m, n\}$.

Aplicando este procedimiento a una matriz de datos obtendremos una representación de la información en función de unos pocos vectores singulares, y por lo tanto dispondremos de una representación de los datos en un espacio de dimensionalidad reducida.

4.2.3. Factorización de Matrices No-Negativas (NNMF)

El método de factorización de matrices no-negativas (*Non-Negative Matrix Factorization*, NMF o NNMF) realiza una descomposición en factores una matriz de datos no-negativa D de la siguiente forma:

$$D_{n,m} \approx W_{n,q} \cdot F_{q,m} \tag{29}$$

donde las matrices no-negativas F y W son conocidas como matriz de características y matriz de pesos, respectivamente.

La matriz de características F, de tamaño $q \times m$, define q características y pondera cada una de ellas mediante la importancia que tiene cada uno de los m atributos. Las características corresponden a temas genéricos que han sido definidos a partir de la agrupación de atributos en diferentes instancias. La componente i, j de F representa la importancia del atributo j en la característica i.

Por otra parte, la matriz de pesos W le atribuye un peso a cada una de las características en función de su importancia en cada una de las n instancias. Sus q columnas corresponden a las características y sus n filas a las instancias del conjunto de datos analizados, de forma que la componente i, j de W indica la importancia que tiene la característica j en la instancia i.

Uno de los parámetros que deben definirse para aplicar NNMF es el número de características q que se desean. Un valor demasiado alto resulta en un etiquetado demasiado específico de las instancias, mientras que un valor excesivamente pequeño agrupa las instancias en unas pocas características demasiado genéricas y por tanto carentes de información relevante. La matriz de características F es equivalente a la matriz de vectores propios que obteníamos con otras técnicas de factorización de datos, como por ejemplo PCA. La diferencia principal entre NNMF y PCA radica en que en NNMF la matriz F es no-negativa y, por tanto, la factorización de D puede interpretarse como una descomposición acumulativa de los datos originales.

Conviene precisar que la descomposición NNMF no es única, es decir, dado un número de características q hay más de una forma de descomponer los datos en la forma de la Ecuación 29. Lo ideal es aplicar el procedimiento e interpretar los resultados para obtener información adicional de los datos. El algoritmo que se encarga de realizar esta factorización no es trivial y escapa al alcance de este libro.

5. Conjuntos desbalanceados de datos

XXX

Resumen

XXX

Glosario

Aprendizaje automático El aprendizaje automático (más conocido por su denominación en inglés, *machine learning*) es el conjunto de técnicas, métodos y algoritmos que permiten a una máquina aprender de manera automática en base a experiencias pasadas.

Aprendizaje no supervisado El aprendizaje no supervisado se basa en el descubrimiento de patrones, características y correlaciones en los datos de entrada, sin intervención externa.

Aprendizaje supervisado El aprendizaje supervisado se basa en el uso de un componente externo, llamado supervisor, que compara los datos obtenidos por el modelo con los datos esperados por éste, y proporciona retroalimentación al modelo para que vaya ajustándose y mejorando las predicciones.

Bibliografía

Y. Bengio, I. Goodfellow, A. Courville (2016). *Deep Learning*. Cambridge, MA: MIT Press.

S. O. Haykin (2009). *Neural Networks and Learning Machines,3rd Edition*. Pearson.

J. Gironés Roig, J. Casas Roma, J. Minguillón Alfonso, R. Caihuelas Quiles (2017). *Minería de datos: Modelos y algoritmos*. Barcelona: Editorial UOC.