Métodos no supervisados y detección de anomalías: el problema de agrupamiento

UCO ONLINE

El problema de agrupamiento

1. Introducción

Los objetos que nos rodean presentan ciertas propiedades que nos hacen ser capaces de identificarlos y clasificarlos. Por ejemplo, somos capaces de diferenciar animales de plantas en base a propiedades como su forma o su hábitat. A su vez, podemos clasificar ambos tipos de seres vivos en otras categorías. Por ejemplo, en el caso de los animales, existen clasificaciones según el reino al que pertenecen, la familia, o la especie. Los animales de una misma especie son muy parecidos entre sí, por ejemplo, pues tienen una fisionomía similar, presentan pelaje o no, etc. A medida que subimos en la clasificación, los animales van perdiendo rasgos en común, pero mantienen otros a un nivel más general, como puede ser el número de extremidades o el tipo de alimentación que siguen.

Conforme los humanos crecemos y aprendemos conceptos, estos sistemas de clasificación basados en propiedades nos ayudan a identificar nuevos objetos y categorizarlos, aunque no sepamos exactamente cómo denominarlo. Por ejemplo, si vemos un pequeño animal que vuela y construye nidos en los árboles, sabremos decir que es un pájaro, aunque no sepamos a qué especie en particular pertenece. Nuestro cerebro es capaz de abstraer las propiedades más relevantes que conoce de un concepto y analizar nuevos objetos en busca de esas propiedades que nos ayuden a entender nuevos objetos.

Este mismo proceso es el que tratan de simular los sistemas basados en aprendizaje automático. Cuando el número y tipo de concepto a aprender está perfectamente delimitado, estamos ante un problema de clasificación que se resuelve por medio de aprendizaje supervisado. Tenemos un conjunto de ejemplos (instancias) cuya clase conocemos, y tratamos de entrenar un modelo que nos diga cuándo una instancia es de una clase u otra. Sin embargo, puede darse la situación de no disponer de ejemplos etiquetados, esto es, sabemos que los ejemplos comparten ciertas propiedades, pero nadie ha indicado a qué clase se asocia cada ejemplo. Estamos ante un problema que requiere un enfoque no supervisado. Además, el objetivo como tal de este nuevo problema no es tanto predecir la clase a la que pertenece un nuevo objeto desconocido, sino analizar las características comunes que explican una posible clasificación de las instancias. A continuación, vamos a definir formalmente este problema, al que denominamos *problema de agrupamiento*.

2. Definición del problema de agrupamiento

El problema de agrupamiento consiste en separar o dividir un conjunto de instancias no etiquetadas en un número finito y discreto de *grupos*. Cada uno de estos grupos se corresponde con un concepto o categoría de forma tal que las instancias agrupadas son muy similares entre sí, y a su vez distintas de las instancias asignadas al resto de grupos. Tal y como se profundizará en el siguiente apartado, la definición de *similitud* (o en términos más generales, *proximidad*), es clave a la hora de abordar el problema de agrupamiento. El proceso de agrupamiento, también llamado análisis de grupos (*cluster analysis*) consiste, por tanto, en determinar cuál es la mejor asignación de instancias a grupos, de forma que se maximice la similitud entre las instancias agrupadas conjuntamente, mientras que se maximiza la disimilitud entre instancias asignadas a distintos grupos. A menudo se trata de un proceso iterativo en el cual se van reasignando instancias a grupos hasta que la asignación se considera "estable" o se alcanza un número máximo de iteraciones. Según el método elegido, puede ser necesario indicar cuántos grupos han de descubrirse o puede ser el propio método el que lo determine durante el proceso. El análisis de grupos es un problema antiguo, pero que ha adquirido mayor repercusión con la expansión del aprendizaje automático y su capacidad para analizar grandes conjuntos de datos. Según Aldenderfer y Blashfield, el análisis de grupos tiene cuatro objetivos principales:

- Desarrollar una clasificación de conceptos
- Investigar esquemas conceptuales que sean útiles a la hora de agrupar entidades
- Generar hipótesis sobre los datos a través de su exploración
- Tratar de determinar si algún concepto, definido por otro medio, está presente en los datos

Aunque no existe una definición universal de qué es un *grupo*, diferentes autores han ido estableciendo sus características. De la definición del problema de agrupamiento se infiere que un grupo es un conjunto de instancias. Si trasladamos esta idea a un espacio de n dimensiones, un grupo sería una región continua de ese espacio donde se concentran una serie de puntos que están próximos entre sí y alejados del resto. Aquí entra en juego de nuevo el concepto de proximidad, que se estudiará más adelante. Por ahora, basta saber que dentro de un grupo las instancias deben tener una alta proximidad, lo que podemos asociar con que presenten poca distancia entre ellas en el plano n dimensional. En concreto, la distancia entre dos puntos cualesquiera que pertenecen a un mismo grupo debería ser menor a la distancia entre cualquiera de estos puntos y otro punto ubicado en otro grupo. Esto es lo que se conoce como *homogeneidad interna* (los puntos son similares entre sí) y *separación externa* (los puntos son distantes al resto de grupos). A lo largo del curso se irá revisitando esta definición para analizar otras propiedades que se le pueden exigir a los grupos, como que sean solapados o no, o se permita componer unos en base a otros.

3. El proceso de análisis de grupos

Como todo proceso de aprendizaje automático, el análisis de grupos parte de un conjunto de datos a analizar, sobre el cual se quiere extraer algún tipo de conocimiento (los grupos en este caso). El proceso habitual de análisis de grupos consta de los siguientes cuatro pasos:

1. Selección o extracción de características. Previo a la ejecución de un algoritmo de agrupamiento, es necesario decidir las características que nos van a ayudar a determinar los grupos. Esta selección es de gran importancia, porque un algoritmo de agrupamiento se va a basar en "medir" la distancia en base a esas características, lo cual puede ser un proceso muy costoso computacionalmente. Idealmente, se deben elegir aquellas propiedades que mejor caracterizan a los distintos conceptos, pero esto es algo que no tiene por qué saberse a priori. Distintos conjuntos de características pueden dar lugar a distintos grupos. En el caso de la agrupación de animales, obtendremos diferentes resultados si las características nos hablan de los hábitos alimenticios (lo que dará lugar a clasificar como carnívoros, herbívoros y omnívoros) o nos hablan de la forma de gestación (lo que dará lugar a clasificar como ovíparos, ovovivíparos y vivíparos).

2. Elección del algoritmo de agrupamiento. Este paso comienza con la definición de la medida de proximidad que va a utilizarse. Esta medida va a depender del tipo de datos que tengamos y de qué entendamos por proximidad para el problema en concreto a resolver. Sea cual sea la formulación elegida, esta medida debe ser capaz de determinar en qué grado dos instancias son similares entre sí. En base a ese criterio, el algoritmo elegido distribuirá las instancias entre grupos tratando de optimizar dicho criterio. Esto es, agrupar juntas las instancias más similares y separar en grupos las instancias menos similares. Existen multitud de métodos de agrupamiento, y no todos ellos se rigen por las mismas "normas" al construir los grupos. Como se verá más adelante, existen métodos que particionan el espacio en un número determinado de grupos, de forma que un punto solo puede pertenecer a un grupo. Otros métodos relajan esta suposición y permiten que un punto pertenezca a varios grupos según una probabilidad. Es importante conocer bien el tipo de problema de agrupamiento a resolver para elegir la técnica más adecuada para su resolución.

3. Validación de los grupos. Al ejecutar un algoritmo de agrupamiento, obtenemos como resultado un conjunto de grupos. Sin embargo, esto no significa que ese conjunto sea el que explique la estructura de los datos o identifique perfectamente los conceptos que subyacen en esos datos. El conjunto de grupos obtenido puede verse afectado por los parámetros del algoritmo y, sobre todo, por la definición del criterio de proximidad sobre las características seleccionadas para el análisis. Dado que no disponemos de las etiquetas "reales" a las que pertenecen las instancias, evaluar la calidad del agrupamiento obtenido es algo más complejo que en aprendizaje supervisado.
Por las particularidades del problema de agrupamiento, una primera forma de validación en la que podemos pensar es la visualización de los grupos resultantes. En un espacio de dos o tres dimensiones, podemos representar las instancias y su agrupación, y comprobar si responden a los criterios de homogeneidad interna y separación externa. No obstante, esta solución no deja de ser subjetiva, porque distintas personas pueden entender la proximidad de forma diferente. Son necesarios criterios objetivos de evaluación, tanto para comprobar si el número de grupos es el adecuado como para comparar la asignación de instancias a grupos según el algoritmo elegido. La validación de grupos se abordará en profundidad en la segunda semana del curso.

4. Interpretación de los resultados. El objetivo final del análisis de grupos es ayudar al usuario a entender los datos, por lo que una parte importante del proceso consiste en estudiar la composición de grupos obtenida y decidir qué puede estar representando cada grupo. Como se ha mencionado antes, una distribución de grupos puede no ser la que mejor explica los datos, tan solo una visión particular dado un tipo de algoritmo y definición de proximidad. Queda por tanto a criterio del usuario decidir si el resultado es acorde a lo esperado o se necesita otro tipo de análisis. Los pasos descritos son habitualmente realizados en secuencia, pero fruto de alguno de ellos puede ser necesario volver atrás para, por ejemplo, ajustar el número de grupos si el resultado no es coherente, o cambiar la selección de las características. La validación nos da información valiosa para entender cómo de bien está funcionamiento el análisis de grupos, y decidir si es necesario hacer algún cambio.

4. El concepto de proximidad

La división de las instancias en grupos requiere de un criterio que nos indique si dos o más instancias deben ser agrupadas juntas o no. Dicho criterio debe darnos una forma objetiva de medir si dos instancias son próximas considerando los valores de sus características. Aparte de medir la proximidad o similitud entre instancias, también puede ser necesario medir la proximidad entre una instancia y un grupo, o entre dos grupos. Este tipo de medidas facilitan la validación de los grupos, por lo que se estudiarán más adelante. A continuación, se presentan las propiedades matemáticas que debe cumplir una medida de proximidad. El resto del apartado se centra en definir las medidas de proximidad más habituales según la naturaleza de los datos.

En este punto, cabe recordar que cada instancia de un conjunto de datos está descrita en base a un número determinado de características (variables). Si disponemos de N instancias y M características, nuestro conjunto de datos queda representado como una matriz de tamaño NxM. A la hora de explicar las medidas, se utilizarán los términos *punto* o *vector* como sinónimo de instancia, pues son conceptos más ligados a la definición matemática de dichas funciones.

4.1. Propiedades de las medidas de proximidad

Una medida de proximidad es una generalización que engloba tanto a medidas de similitud como de disimilitud. A menudo también se utiliza el término medida de distancia, aunque existe una diferencia sutil entre distancia y similitud. Matemáticamente, una *medida de distancia* debe cumplir las siguientes propiedades:

1. Simetría. Dados dos puntos x_i y x_j, la distancia entre ellos es la misma con independencia del orden con el que se aplique la función:

$$d(x_i, x_j) = d(x_j, x_i) \tag{1}$$

2. Positividad. El resultado de una medida de distancia siempre es mayor o igual a cero.

$$d(x_i, x_j) \geq 0 \qquad \forall x_i, x_j \tag{2}$$

3. Desigualdad triangular. Expresa la relación entre tres puntos, de forma que la distancia entre los dos más lejanos es igual o inferior a la suma de las distancias entre cada uno de ellos y un punto intermedio:

$$d(x_i, x_j) \leq d(x_i, x_k) + d(x_k, x_j) \qquad \forall x_i, x_j, x_k \tag{3}$$

4. Reflexión. La distancia de un punto a sí mismo es igual a 0:

$$d(x_i, x_j) = 0 \leftrightarrow x_i = x_j \quad (4)$$

Una *medida de similitud* debe cumplir también con las propiedades anteriores, pero sus valores posibles se restringen al rango [0,1]. Por tanto, las propiedades anteriores son definidas como sigue:

1. Simetría: $s(x_i, x_j) = s(x_j, x_i)$ (5)

2. Positividad: $0 \leq s(x_i, x_j) \leq 1 \quad \forall x_i, x_j$ (6)

3. Similitud entre tres puntos: $s(x_i, x_j)s(x_j, x_k) \leq [s(x_i, x_j) + s(x_j, x_k)]s(x_i, x_k) \; \forall x_i, x_j, x_k$ (7)

4. Reflexión: $s(x_i, x_j) = 1 \leftrightarrow x_i = x_j$ (8)

Por tanto, dados *N* puntos en un espacio de *M* dimensiones (nuestro conjunto de datos), podemos definir una matriz de proximidad (o una matriz de distancia) simétrica de tamaño NxN. La distancia entre el punto i y el punto j será el valor de la medida de similitud (o distancia) entre ellos, y ocupará las posiciones *(i, j)* y *(j, i)* de la matriz. La construcción de esta matriz es la base de muchos de los métodos de agrupamiento que se estudiarán en este curso.

4.2. Medidas de proximidad para variables continuas

Las características numéricas como una temperatura, una longitud o un volumen son representadas como variables continuas. Para este tipo de variables, la distancia más conocida es la *distancia Euclídea*, o norma L_2, la cual se calcula como:

$$d(x_i, x_j) = \sum_{m=1}^{M} \sqrt{(x_{i,m} - x_{j,m})^2} \quad (9)$$

La distancia Euclídea es invariante a traslaciones y rotaciones de los puntos en el espacio y, cuando es aplicada al problema de agrupamiento, tiende a formar grupos de forma esférica. Un inconveniente de esta distancia es que es sensible a la escala de las variables. Aquellas que varían en un rango más grande de valores tendrán más influencia en el cómputo total de la distancia. No obstante, este problema se puede resolver escalando las variables antes de calcular la distancia. Lo más habitual es escalar en base a la media y la desviación estándar (conocida como estandarización o normalización), o en base a los valores máximos y mínimos (normalización basada en la unidad). La distancia Euclídea es un caso particular de las llamadas normas L_p (p=2, en este caso). De forma general, las distancias que siguen la norma L_p (llamada distancia de Minkowski), tienen la siguiente formulación:

$$d(x_i, x_j) = \left(\sum_{m=1}^{M} |x_{i,m} - x_{j,m}|^{1/p}\right)^p \quad (10)$$

Otra distancia de esta misma familia que también se utiliza frecuentemente en problemas de agrupamiento es la *distancia de Manhattan*, que se corresponde con la norma L_1 (p=1):

$$d(x_i, x_j) = \sum_{m=1}^{M} |x_{i,m} - x_{j,m}| \quad (11)$$

Otra medida de distancia que no tiene en cuenta solo la "cercanía" de los valores sino también la correlación entre variables es la *distancia de Mahalanobis*. Para su cómputo es necesario considerar la matriz de covarianza C, así como garantizar que las dos variables siguen la misma distribución de probabilidad. La distancia se calcula con la siguiente expresión:

$$d(x_i, x_j) = (x_i - x_j)^T C^{-1} (x_i - x_j) \quad (12)$$

La distancia de Mahalanobis es también insensible a transformaciones lineales en los puntos, pero a diferencia de la distancia Euclídea, tiende a generar grupos con forma elíptica.
Todas las medidas anteriores se basan en el concepto de distancia, de forma que expresan la proximidad en el espacio multidimensional. Otro tipo de medidas que podemos utilizar son las medidas puramente basadas en correlación. Con este tipo de medidas, estaremos expresando que existe una similitud entre los puntos que es debida a cómo se relacionan sus variables, aunque los valores de las variables difieran considerablemente entre los dos puntos. La primera medida de correlación en la que nos podemos basar es el coeficiente de *correlación de Pearson:*

$$pearson(x_i, x_j) = \frac{\sum_{m=1}^{M}(x_{i,m}-\bar{x}_i)(x_{j,m}-\bar{x}_j)}{\sqrt{\sum_{m=1}^{M}(x_{i,m}-\bar{x}_i)^2 \sum_{m=1}^{M}(x_{j,m}-\bar{x}_j)^2}} \quad (13)$$

Este coeficiente varía entre -1 (máxima correlación negativa) y 1 (máxima correlación positiva). Para utilizarla como medida de distancia, esto es que cumpla la condición de que su valor sea siempre mayor a cero, basta con aplicar una sencilla transformación:

$$d(x_i, x_j) = (1 - pearson(x_i, x_j))/2 \quad (14)$$

Otros coeficientes de correlación que pueden utilizarse con el mismo propósito son los coeficientes de Spearman y de Kendall. El *coeficiente de Spearman* no se basa en los valores como tal de las variables, sino que calcula la correlación entre los rankings que se obtienen al ordenar dichos valores:

$$spearman(x_i, x_j) = 1 - \frac{\sum_{m=1}^{M}(rank(x_{i,m})-rank(\bar{x}_i))(rank(x_{j,m})-rank(\bar{x}_j))}{\sqrt{\sum_{m=1}^{M}(rank(x_{i,m})-rank(\bar{x}_i))^2 \sum_{m=1}^{M}(rank(x_{j,m})-rank(\bar{x}_j))^2}} \quad (15)$$

La medida de *correlación de Kendall* también se basa en los valores ordenados en rankings de las variables, pero en este caso, cuenta el número de pares de valores que son concordantes y discordantes. Dado un par de valores $(x_{i,1}, x_{i,2})$ y $(x_{j,1}, x_{j,2})$, se dice que se trata de un par concordante si al ordenar sus valores en un ranking, se mantiene el orden de las posiciones. Esto sucede en dos situaciones: 1) $x_{i,1} > x_{j,1} \wedge x_{i,2} > x_{j,2}$ o 2) $x_{i,1} < x_{j,1} \wedge x_{i,2} < x_{j,2}$.
Si n_c es el número de pares concordantes, y n_d el número de pares discordantes, el coeficiente de correlación de Kendall se calcula como sigue:

$$kendall(x_i, x_j) = 1 - \frac{n_c - n_d}{\frac{1}{2}M(M-1)} \quad (16)$$

Por último, una medida de similitud es la *similitud del coseno*. Cuanto más similares sean los dos vectores en el espacio M dimensional, formarán un ángulo más "paralelo", de forma que el coseno de ese ángulo será más grande. Su formulación es la siguiente:

$$cosine(x_i, x_j) = \frac{x_i \cdot x_j}{\|x_i\|\|x_j\|} \quad (17)$$

Si quisiéramos utilizar esta medida de similitud como medida de distancia, se podría hacer una transformación simple:

$$d(x_i, x_j) = 1 - cosine(x_i, x_j) \tag{18}$$

4.3. Medidas de proximidad para variables discretas

Las medidas de proximidad para variables discretas se pueden dividir en dos grupos: las definidas para variables binarias, esto es, que solo toman dos valores posibles; y las definidas para variables que pueden tomar más de dos valores.

En el caso de las variables binarias, la medida más habitual es la *distancia de Hamming*. Dados dos vectores que toman los valores 0 y 1 en sus M posiciones, la distancia de Hamming consiste en calcular en cuántas posiciones los valores de los dos vectores difieren. Matemáticamente, se puede expresar de la siguiente forma:

$$d(x_i, x_j) = \frac{n_{10}+n_{01}}{n_{11}+n_{00}+n_{10}+n_{01}} \tag{19}$$

La distancia de Hamming tiene relación con una formulación general de medidas de similitud para variables binarias que se siguen la siguiente formulación general:

$$s(x_i, x_j) = \frac{n_{11}+n_{00}}{n_{11}+n_{00}+w(n_{10}+n_{01})} \tag{20}$$

Es sencillo comprobar que la distancia de Hamming es equivalente a $1 - s(x_1, x_2)$ cuando $w = 1$. El peso w permite ponderar la importancia que se le concede a los pares discordantes.

Otra familia de medidas de similitud binarias solo considera relevante la coincidencia en del valor 1. En esta familia, cuando $w = 1$, se obtiene el coeficiente o *índice de Jaccard*:

$$jaccard(x_i, x_j) = \frac{n_{11}}{n_{11}+w(n_{10}+n_{01})} \tag{21}$$

No obstante, es más sencillo interpretarlo en base a los operadores de unión e intersección. Para dos conjuntos, el índice de Jaccard calcula la ratio entre los elementos que están presentes en ambos conjuntos, entre el número total de elementos entre ambos conjuntos:

$$jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{22}$$

Esta medida varía entre 0 y 1, por lo que no es necesaria ninguna transformación para utilizarla como medida de distancia.

Cuando el número de valores posibles es mayor a dos, una primera solución al cálculo de la similitud es transformar cada variable discreta en tantas variables binarias como valores pueda tomar. Solo una de esas variables tomará el valor 1. Por ejemplo, si la variable original podía tomar los valores "A", "B" y "C", podemos reemplazarla por las siguientes variables binarias: "esA", "esB", "esC". Si una instancia tenía el valor "A", su codificación en las variables binarias será: esA=1, esB=0 y esC=0. Esta transformación aumenta el número de variables, lo cual puede complicar su posterior análisis, pero es una solución habitual en aprendizaje automático cuando las implementaciones de los algoritmos solo pueden trabajar con variables numéricas (como ocurren en *scikit-learn*, por ejemplo).

Si queremos mantener el número de variables original y nuestra implementación no presenta ese tipo de restricción, podemos adaptar la idea del mapeo binario para contabilizar cuántas veces los vectores coinciden en los valores de cada característica. Se trata de una simple medida de correspondencia, que matemáticamente se expresa como sigue:

$$s(x_i, x_j) = \frac{1}{m}\sum_{m=1}^{M} s_m(x_{i,m}, x_{j,m}) \qquad (23)$$

$$s_m(x_{i,m}, x_{j,m}) = \begin{cases} 0 & si\ x_{i,m} \neq x_{j,m} \\ w & si\ x_{i,m} = x_{j,m} \end{cases}$$

Habitualmente, $w = 1$, pero podrían considerarse otros valores mayores a 1 cuando el número de valores que puede tomar una variable sea mayor. De esta forma se le puede dar más peso a la coincidencia en dicha variable frente a otras variables que tienen menos valores, donde la coincidencia es más probable.

5. Ejemplos

En este apartado se presentan algunos ejemplos para poner en práctica el cálculo de medidas de similitud. En primer lugar, se trabajará con variables continuas. A continuación, se realizan los ejemplos para variables discretas, incluyendo variables binarias.

5.1. Cálculo de proximidad con variables continuas

Para estos ejemplos, vamos a partir de tres supuestas instancias con cuatro características numéricas:

$x_1 = (3.2, 4.6, 1.9, 8.4)$
$x_2 = (12.1, 16.2, 4.3, 20.2)$
$x_3 = (2.9, 1.4, 2.5, 6.5)$

En primer lugar, vamos a calcular la distancia Euclídea entre la primera instancia y el resto siguiendo la fórmula de la ecuación (9):

$$d(x_1, x_2) = \sum_{m=1}^{4} \sqrt{(x_{1,m} - x_{2,m})^2} = \sqrt{(3.2 - 12.1)^2 + (4.6 - 16.2)^2 + (1.9 - 4.3)^2 + (8.4 - 20.2)^2} = 18.94$$

$$d(x_1, x_3) = \sum_{m=1}^{4} \sqrt{(x_{1,m} - x_{3,m})^2} = \sqrt{(3.2 - 2.9)^2 + (4.6 - 1.4)^2 + (1.9 - 2.5)^2 + (8.4 - 6.5)^2} = 3.78$$

Según esta medida de distancia espacial, diríamos que x_1 es más próxima a x_3 que a x_2, porque la distancia a x_3 es menor que a x_2. Vamos a calcular a continuación la distancia de Manhattan (ecuación 11) para comprobar si se mantiene esta relación entre las tres instancias:

$$d(x_1, x_2) = \sum_{m=1}^{4} |x_{1,m} - x_{2,m}|^2 = |3.2 - 12.1|^2 + |4.6 - 16.2|^2 + |1.9 - 4.3|^2 + |8.4 - 20.2|^2 = 34.70$$

$$d(x_1, x_3) = \sum_{m=1}^{4} |x_{1,m} - x_{3,m}|^2 = |3.2 - 2.9|^2 + |4.6 - 1.4|^2 + |1.9 - 2.5|^2 + |8.4 - 6.5|^2 = 6.0$$

Puesto que se trata de otra medida de distancia espacial, la conclusión que obtenemos respecto a la proximidad entre las instancias es la misma, solo cambia la magnitud de los valores obtenidos. Veamos qué ocurre si consideramos una medida de distancia basada en correlación en su lugar. En este caso en particular, utilizaremos el coeficiente de Pearson (ver ecuaciones 13 y 14):

En primer lugar, necesitaremos las medias de cada instancia:

$$\bar{x}_1 = \frac{3.2 + 4.6 + 1.9 + 8.4}{4} = 4.525$$
$$\bar{x}_2 = \frac{12.1 + 16.2 + 4.3 + 20.2}{4} = 13.2$$
$$\bar{x}_3 = \frac{2.9 + 1.4 + 2.5 + 6.5}{4} = 3.325$$

A continuación, podemos calcular el coeficiente de correlación de Pearson para los dos pares de instancias:

$$pearson(x_1, x_2) = \frac{\sum_{m=1}^{4}(x_{1,m} - \bar{x}_1)(x_{2,m} - \bar{x}_2)}{\sqrt{\sum_{m=1}^{4}(x_{1,m} - \bar{x}_1)^2 \cdot \sum_{m=1}^{4}(x_{2,m} - \bar{x}_2)^2}} =$$
$$= \frac{(3.2 - 4.525)(12.1 - 13.2) + (4.6 - 4.525)(16.2 - 13.2) + (1.9 - 4.525)(4.3 - 13.2) + (8.4 - 4.525)(20.2 - 13.2)}{\sqrt{[(3.2 - 4.525)^2 + (4.6 - 4.525)^2 + (1.9 - 4.525)^2 + (8.4 - 4.525)^2] \cdot [(12.1 - 13.2)^2 + (16.2 - 13.2)^2 + (4.31 - 13.2)^2 + (20.2 - 13.2)^2]}}$$
$$= 0.9115$$

$$pearson(x_1, x_3) = \frac{\sum_{m=1}^{4}(x_{1,m} - \bar{x}_1)(x_{3,m} - \bar{x}_3)}{\sqrt{\sum_{m=1}^{4}(x_{1,m} - \bar{x}_1)^2 \cdot \sum_{m=1}^{4}(x_{3,m} - \bar{x}_3)^2}} =$$
$$= \frac{(3.2 - 4.525)(2.9 - 3.325) + (4.6 - 4.525)(1.4 - 3.325) + (1.9 - 4.525)(2.5 - 3.325) + (8.4 - 4.525)(6.5 - 3.325)}{\sqrt{[(3.2 - 4.525)^2 + (4.6 - 4.525)^2 + (1.9 - 4.525)^2 + (8.4 - 4.525)^2] \cdot [(2.9 - 3.325)^2 + (1.4 - 3.325)^2 + (2.5 - 3.325)^2 + (6.5 - 3.325)^2]}}$$
$$= 0.7996$$

Finalmente, obtenemos las distancias:

$$d(x_1, x_2) = \frac{1 - 0.9115}{2} = 0.0443$$
$$d(x_1, x_3) = \frac{1 - 0.7996}{2} = 0.1002$$

Vemos que, según esta otra formulación de distancia, la instancia x_1 es más similar a x_2 que a x_3. Esto se debe a que se está evaluando la distribución de las características, en lugar de sus valores concretos. Así, podemos ver que x_1 y x_2 comparten que los valores de las características (m) siguen el siguiente orden creciente: $m_3 < m_2 < m_1 < m_4$. Sin embargo, en x_2 el orden es: $m_2 < m_3 < m_1 < m_4$. Al independizar de la escala por restar a cada instancia su media, es la distribución respecto a la media (valores superiores o inferiores) la que afecta al cómputo de proximidad.

5.2. Cálculo de proximidad con variables discretas

En este apartado se presentan dos ejemplos del cálculo de medidas de similitud, uno para variables binarias utilizando la distancia de Hamming, y otro para variables discretas utilizando el índice de Jaccard. Para el caso de variables binarias, consideraremos las siguientes dos instancias con diez características que toman el valor 0 o 1:

$x_1 = (1, 0, 0, 0, 1, 1, 0, 1, 1, 1)$
$x_2 = (0, 1, 0, 0, 1, 1, 1, 0, 1, 1)$

Para calcular la distancia de Hamming, necesitamos contabilizar el número de posiciones en las cuales coinciden o no las dos instancias. Con ello determinamos los valores n_{11}, n_{10}, n_{01} y n_{00}. Por ejemplo, n_{11} contabiliza el número de posiciones en las que ambas instancias tienen asignado el valor 1. Vemos que esto ocurre en 4 ocasiones. De igual forma se determina el resto de los casos:

$n_{11} = 4$
$n_{10} = 2$
$n_{01} = 2$
$n_{00} = 2$

A continuación, calculamos la distancia según la ecuación 19:

$$d(x_1, x_2) = \frac{n_{10} + n_{01}}{n_{11} + n_{00} + n_{10} + n_{01}} = 1 - \frac{2 + 2}{4 + 2 + 2 + 2} = 0.4$$

Otra medida que podemos aplicar en este ejemplo es el índice de Jaccard, pues también se basa en el número de coincidencias. En este caso, aplicaremos la fórmula de la ecuación 21, con $w = 1$:

$$jaccard(x_1, x_2) = \frac{n_{11}}{n_{11} + n_{10} + n_{01}} = \frac{4}{4 + 2 + 2} = 0.5$$

No obstante, la interpretación de este índice es más intuitiva cuando las instancias representan conjuntos de variables discretas (no necesariamente binarias). Imaginemos por ejemplo el siguiente caso: $A = \{'a', 'b', 'c', 'd', 'e', 'f'\}$ y $B = \{'a', 'e', 'i', 'o', 'u'\}$. El índice de Jaccard se obtiene fácilmente a partir del cálculo del conjunto unión e intersección según la ecuación 22:

$$jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|\{'a', 'e'\}|}{|\{'a', 'b', 'c', 'd', 'e', 'f', 'i', 'o', 'u'\}|} = \frac{2}{9} = 0.22$$

Referencias

C. C. Aggarwal, C. K. Reddy (eds.). "Data Clustering: Algorithms and Applications". Chapman & Hall / CRC Press, 1ª edición, 652 páginas. 2014.

T. Hastie, R. Tibshirani, J. Friedman. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Springer Series in Statistics, 2ª edición, 745 páginas. 2017.

G. James, D. Witten, R. Tibshirani, T. Hastie. "An Introduction to Statistical Learning with Applications in R". Springer Texts in Statistics, 1ª edición (7ª impresión), 426 páginas. 2017. Disponible en: https://www.statlearning.com/

A. Kassambara. "Practical Guide to Cluster Analysis in R". STHDA, 187 páginas. 2017.

S. Shalev-Shwartz, S. Ben-David. "Understanding Machine Learning: From Theory to Algorithms". Cambridge University Press, 1ª edición, 449 páginas. 2014.

R. Xu, D.C. Wunsch II. "Clustering". Wiley/IEEE Press, 1ª edición, 363 páginas. 2009.