Búsqueda de patrones: Análisis de la cesta de compra mediante reglas de asociación

Piensa en tu última compra impulsiva. Tal vez en la caja del supermercado compraste un paquete de chicles o una barra de chocolate. Tal vez en un viaje nocturno a comprar pañales y leche en polvo, compraste una bebida con cafeína o un paquete de seis cervezas. Es posible que incluso hayas comprado un libro por recomendación de un librero. Estas compras impulsivas no son una coincidencia, ya que los minoristas utilizan sofisticadas técnicas de análisis de datos para identificar patrones útiles para promociones de marketing e impulsar ventas adicionales a través de la colocación de productos.

En años pasados, tales recomendaciones se basaban en la intuición subjetiva de los profesionales del marketing y los gerentes de inventario. Ahora, los lectores de códigos de barras, las bases de datos de inventario y los carritos de compra en línea generan datos transaccionales que el aprendizaje automático puede utilizar para aprender patrones de compra. La práctica se conoce comúnmente como *análisis de la cesta de la compra* porque se ha aplicado con mucha frecuencia a los datos de los supermercados.

Aunque la técnica se originó con los datos de compras, también es útil en otros contextos. Cuando termines este documento, sabrás cómo aplicar las técnicas de análisis de la cesta de la compra a tus propias tareas, sean cuales sean. En general, el trabajo implica:

- Comprender las peculiaridades de los datos transaccionales
- Usar medidas de rendimiento simples para encontrar asociaciones en bases de datos grandes
- Saber cómo identificar patrones útiles y procesables

Dado que el análisis de la cesta de la compra puede descubrir fragmentos de información en muchos tipos de conjuntos de datos grandes, a medida que aplicamos la técnica, es probable que identifiques aplicaciones para tu trabajo incluso si no tienes afiliación con el sector minorista.

Comprender las reglas de asociación

Los componentes básicos de un análisis de la cesta de la compra son los artículos que pueden aparecer en cualquier transacción dada. Los grupos de uno o más artículos están rodeados por corchetes para indicar que forman un conjunto, o más específicamente, un conjunto de artículos que aparece en los datos con cierta regularidad. Las transacciones se especifican en

términos de conjuntos de elementos, como la siguiente transacción que podría encontrarse en una tienda de comestibles típica:

{pan, mantequilla de maní, mermelada}

El resultado de un análisis de la cesta de la compra es una colección de reglas de asociación que especifican patrones encontrados en las relaciones entre los elementos de los conjuntos de elementos. Las reglas de asociación siempre se componen de subconjuntos de conjuntos de elementos y se denotan relacionando un conjunto de elementos en el lado izquierdo (LHS, **Left-hand-side**) de la regla con otro conjunto de elementos en el lado derecho (RHS, **Right-hand side**) de la regla. El LHS es la condición que se debe cumplir para activar la regla, y el RHS es el resultado esperado de cumplir esa condición. Una regla identificada a partir de la transacción del ejemplo anterior podría expresarse en la forma:

{mantequilla de maní, mermelada} → {pan}

En lenguaje sencillo, esta regla de asociación establece que si se compran juntas mantequilla de maní y mermelada, entonces es probable que también se compre pan. En otras palabras, “la mantequilla de maní y la mermelada implican pan”.

Desarrolladas en el contexto de bases de datos de transacciones minoristas, las reglas de asociación no se utilizan para la predicción, sino para el descubrimiento de conocimiento no supervisado en bases de datos grandes. Esto es diferente a los algoritmos de clasificación y predicción numérica presentados en temas anteriores. Aun así, encontrarás que el resultado del aprendizaje de reglas de asociación está estrechamente relacionado y comparte muchas características con el resultado del aprendizaje de reglas de clasificación como se presentó en el tema, Divide y vencerás: Clasificación mediante árboles de decisión y reglas.

Debido a que los aprendices de reglas de asociación no están supervisados, no es necesario entrenar al algoritmo y no es necesario etiquetar los datos con anticipación. El programa simplemente se lanza en un conjunto de datos con la esperanza de encontrar asociaciones interesantes. La desventaja, por supuesto, es que no hay una manera fácil de medir objetivamente el desempeño de un aprendiz de reglas, además de evaluarlo para determinar su utilidad cualitativa, generalmente, una prueba visual de algún tipo.

Aunque las reglas de asociación se utilizan con mayor frecuencia para el análisis de la cesta de la compra, son útiles para encontrar patrones en muchos tipos diferentes de datos. Otras posibles aplicaciones incluyen:

- Búsqueda de patrones interesantes y frecuentes de secuencias de ADN y proteínas en datos sobre el cáncer

- Búsqueda de patrones de compras o reclamaciones médicas que se dan en combinación con el uso fraudulento de tarjetas de crédito o seguros
- Identificación de combinaciones de comportamiento que preceden a que los clientes cancelen su servicio de telefonía celular o actualicen su paquete de televisión por cable

El análisis de reglas de asociación se utiliza para buscar conexiones interesantes entre una gran cantidad de elementos. Los seres humanos son capaces de obtener esa información de manera bastante intuitiva, pero a menudo se necesitan conocimientos de nivel experto o una gran experiencia para hacer lo que un algoritmo de aprendizaje de reglas puede hacer en minutos o incluso segundos. Además, algunos conjuntos de datos son simplemente demasiado grandes y complejos para que un ser humano encuentre la aguja en el pajar.

El algoritmo Apriori para el aprendizaje de reglas de asociación

Así como los grandes conjuntos de datos transaccionales crean desafíos para los humanos, estos conjuntos de datos también presentan desafíos para las máquinas. Los conjuntos de datos transaccionales pueden ser grandes tanto en la cantidad de transacciones como en la cantidad de elementos o características que se registran. El problema fundamental de la búsqueda de conjuntos de elementos interesantes es que la cantidad de conjuntos de elementos potenciales crece exponencialmente con la cantidad de elementos. Dados k elementos que pueden aparecer o no en un conjunto, hay 2^k posibles conjuntos de elementos que podrían ser reglas potenciales. Un minorista que vende solo 100 artículos diferentes podría tener del orden de 2^{100} = 1.27e+30 conjuntos de elementos que un algoritmo debe evaluar, una tarea aparentemente imposible.

En lugar de evaluar cada uno de estos conjuntos de elementos uno por uno, un algoritmo de aprendizaje de reglas más inteligente aprovecha el hecho de que muchas de las posibles combinaciones de elementos rara vez se encuentran en la práctica, si es que alguna vez se encuentran. Por ejemplo, incluso si una tienda vende artículos automotrices y productos alimenticios, es probable que un conjunto de {aceite de motor, plátanos} sea extraordinariamente poco común. Al ignorar estas combinaciones raras (y quizás menos importantes), es posible limitar el alcance de la búsqueda de reglas a un tamaño más manejable.

Se ha realizado mucho trabajo para identificar algoritmos heurísticos para reducir la cantidad de conjuntos de elementos a buscar. Quizás el método más conocido para buscar reglas de manera eficiente en bases de datos grandes es el conocido como **Apriori**. Introducido en 1994 por Rakesh Agrawal y Ramakrishnan Srikant, el algoritmo Apriori se ha convertido desde entonces en sinónimo de aprendizaje de reglas de asociación, a pesar de la invención de algoritmos más nuevos y rápidos. El nombre se deriva del hecho de que el algoritmo utiliza

una creencia previa simple (es decir, *a priori*) sobre las propiedades de los conjuntos de elementos frecuentes.

Antes de analizar esto en mayor profundidad, vale la pena señalar que este algoritmo, como todos los algoritmos de aprendizaje, no está exento de fortalezas y debilidades. Algunas de ellas se enumeran a continuación:

Fortalezas	Debilidades
• Capaz de trabajar con grandes cantidades de datos transaccionales • Produce reglas que son fáciles de entender • Útil para la minería de datos y el descubrimiento de conocimiento inesperado en bases de datos	• No es muy útil para conjuntos de datos relativamente pequeños • Requiere esfuerzo separar la información verdadera del sentido común • Es fácil extraer conclusiones falsas de patrones aleatorios

Como se señaló anteriormente, el algoritmo Apriori emplea una creencia a priori simple como guía para reducir el espacio de búsqueda de reglas de asociación: todos los subconjuntos de un conjunto de elementos frecuentes también deben ser frecuentes. Esta heurística se conoce como la **propiedad Apriori**.

Usando esta astuta observación, es posible limitar drásticamente la cantidad de reglas para buscar. Por ejemplo, el conjunto {aceite de motor, plátanos} solo puede ser frecuente si tanto {aceite de motor} como {plátanos} también ocurren con frecuencia. En consecuencia, si {aceite de motor} o {plátanos} son poco frecuentes, entonces cualquier conjunto que contenga estos elementos puede excluirse de la búsqueda.

Para obtener más detalles sobre el algoritmo Apriori, consulta Fast Algorithms for Mining Association Rules, Agrawal, R., Srikant, R., Proceedings of the 20th International Conference on Very Large Databases, 1994, pp. 487-499.

Para ver cómo se puede aplicar este principio en un entorno más realista, consideremos una base de datos de transacciones simple. La siguiente figura muestra cinco transacciones completadas en la tienda de regalos de un hospital imaginario:

Al observar los conjuntos de compras, se puede inferir que hay un par de patrones de compra típicos.
Una persona que visita a un amigo o familiar enfermo tiende a comprar una tarjeta de pronta recuperación y flores, mientras que los visitantes de madres primerizas tienden a comprar ositos de peluche y globos. Estos patrones son notables porque aparecen con la suficiente frecuencia como para captar nuestro interés; simplemente aplicamos un poco de lógica y experiencia en la materia para explicar la regla.

transaction ID	items purchased
1	{flowers, get-well card, soda}
2	{plush toy bear, flowers, balloons, candy bar}
3	{get-well card, candy bar, flowers}
4	{plush toy bear, balloons, soda}
5	{flowers, get-well card, soda}

Figura 1: Conjuntos de elementos que representan cinco transacciones en la tienda de regalos de un hospital hipotético.

De manera similar, el algoritmo Apriori utiliza medidas estadísticas del "interés" de un conjunto de elementos para localizar reglas de asociación en bases de datos de transacciones mucho más grandes. En las secciones que siguen, descubriremos cómo Apriori calcula estas medidas de interés y cómo se combinan con la propiedad Apriori para reducir la cantidad de reglas que se deben aprender.

Medición del interés de las reglas - apoyo y confianza

El hecho de que una regla de asociación se considere interesante o no se determina mediante dos medidas estadísticas: apoyo y confianza. Al proporcionar umbrales mínimos para cada una de estas métricas y aplicar el principio Apriori, es fácil limitar drásticamente la cantidad de reglas informadas. Si este límite es demasiado estricto, puede hacer que solo se identifiquen las reglas más obvias o de sentido común.

Por esta razón, es importante comprender cuidadosamente los tipos de reglas que se excluyen según estos criterios para poder obtener el equilibrio adecuado.

El **soporte** de un conjunto de elementos o una regla mide la frecuencia con la que aparece en los datos. Por ejemplo, el conjunto de elementos {tarjeta de pronta recuperación, flores} tiene un soporte de 3/5 = 0.6 en los datos de la tienda de regalos del hospital. De manera similar, el soporte para {tarjeta de pronta recuperación} → {flores} también es 0.6. El soporte se puede calcular para cualquier conjunto de elementos o incluso para un solo elemento; por ejemplo, el soporte para {barra de chocolate} es 2/5 = 0.4, ya que las barras de chocolate aparecen en el 40 por ciento de las compras. Una función que define el soporte para el conjunto de elementos *X* podría definirse como:

$$\text{apoyo}(X) = \frac{\text{conteo}(X)}{N}$$

Aquí, N es el número de transacciones en la base de datos y conteo(X) es el número de transacciones que contienen el conjunto de elementos X.

La **confianza** de una regla es una medida de su poder predictivo o precisión. Se define como el soporte del conjunto de elementos que contiene tanto X como Y dividido por el soporte del conjunto de elementos que contiene solo X:

$$\text{confianza}(X \rightarrow Y) = \frac{\text{apoyo}(X, Y)}{\text{apoyo}(X)}$$

Básicamente, la confianza nos indica la proporción de transacciones en las que la presencia del elemento o conjunto de elementos X da como resultado la presencia del elemento o conjunto de elementos Y. Ten en cuenta que la confianza de que X conduce a Y no es la misma que la confianza de que Y conduce a X.

Por ejemplo, la confianza de {flores} → {tarjeta de pronta recuperación} es 0.6 / 0.8 = 0.75. En comparación, la confianza de {tarjeta de pronta recuperación} → {flores} es 0.6 / 0.6 = 1.0.

Esto significa que una compra de flores también incluye la compra de una tarjeta de pronta recuperación el 75 por ciento de las veces, mientras que una compra de una tarjeta de pronta recuperación también incluye flores el 100 por ciento de las veces. Esta información podría ser bastante útil para la administración de la tienda de regalos.

Es posible que haya notado similitudes entre el apoyo, la confianza y las reglas de probabilidad bayesianas que se tratan en el tema, Aprendizaje probabilístico: Clasificación mediante naïve Bayes. De hecho, el apoyo(A, B) es lo mismo que P(A ∩ B) y la confianza(A → B) es lo mismo que P(B | A). Es solo el contexto el que difiere.

Las reglas como {tarjeta de pronta recuperación} → {flores} se conocen como reglas fuertes porque tienen un alto nivel de apoyo y confianza. Una forma de encontrar reglas más fuertes sería examinar cada combinación posible de artículos en la tienda de regalos, medir el apoyo y la confianza, y reportar solo aquellas reglas que cumplan con ciertos niveles de interés. Sin embargo, como se señaló anteriormente, esta estrategia generalmente no es factible para nada cuando los conjuntos de datos sean más pequeños.

En la siguiente sección, verás cómo el algoritmo Apriori utiliza niveles mínimos de apoyo y confianza con el principio Apriori para encontrar reglas sólidas rápidamente al reducir la cantidad de reglas a un nivel más manejable.

Creación de un conjunto de reglas con el principio Apriori

Recuerda que el principio Apriori establece que todos los subconjuntos de un conjunto de elementos frecuentes también deben ser frecuentes.

En otras palabras, si {A, B} es frecuente, entonces {A} y {B} deben ser frecuentes. Recuerda también que, por definición, la métrica de apoyo indica la frecuencia con la que aparece un conjunto de elementos en los datos. Por lo tanto, si sabemos que {A} no cumple con un umbral de apoyo deseado, no hay razón para considerar {A, B} o cualquier otro conjunto de elementos que contenga {A}; estos no pueden ser frecuentes.

El algoritmo Apriori utiliza esta lógica para excluir posibles reglas de asociación antes de evaluarlas. El proceso de creación de reglas se produce en dos fases:

1. Identificación de todos los conjuntos de elementos que cumplen un umbral mínimo de apoyo
2. Creación de reglas a partir de estos conjuntos de elementos utilizando aquellos que cumplen un umbral mínimo de confianza

La primera fase se produce en múltiples iteraciones. Cada iteración sucesiva implica evaluar la compatibilidad de un conjunto de conjuntos de elementos cada vez más grandes. Por ejemplo, la iteración uno implica evaluar el conjunto de conjuntos de elementos de 1 elemento (conjuntos de elementos 1), la iteración dos evalúa los conjuntos de elementos de 2, y así sucesivamente. El resultado de cada iteración *i* es un conjunto de todos los conjuntos de elementos *i* que cumplen el umbral mínimo de apoyo.

Todos los conjuntos de elementos de la iteración *i* se combinan para generar conjuntos de elementos candidatos para su evaluación en la iteración *i* + 1. Pero el principio Apriori puede eliminar algunos de ellos incluso antes de que comience la siguiente ronda. Si {A}, {B} y {C} son frecuentes en la iteración uno, mientras que {D} no lo es, entonces la iteración dos considerará solo {A, B}, {A, C} y {B, C}. Por lo tanto, el algoritmo necesita evaluar solo tres conjuntos de elementos en lugar de los seis conjuntos de elementos de 2 elementos que habrían sido necesarios evaluar si los conjuntos que contienen D no se hubieran eliminado a priori.

Continuando con este pensamiento, supongamos que durante la iteración dos se descubre que {A, B} y {B, C} son frecuentes, pero {A, C} no lo es. Aunque la iteración tres normalmente comenzaría evaluando el apoyo para el conjunto de elementos de 3 elementos {A, B, C}, este paso no es necesario. ¿Por qué no? El principio Apriori establece que {A, B, C} no puede ser frecuente, ya que el subconjunto {A, C} no lo es. Por lo tanto, al no haber generado nuevos conjuntos de elementos en la iteración tres, el algoritmo puede detenerse.

iteration	must evaluate	frequent itemsets	infrequent itemsets
1	{A}, {B}, {C}, {D}	{A}, {B}, {C}	{D}
2	{A, B}, {A, C}, {B, C} ~~{A, D}, {B, D}, {C, D}~~	{A, B}, {B, C}	{A, C}
3	~~{A, B, C}, {A, B, D}~~ ~~{A, C, D}, {B, C, D}~~		
4	~~{A, B, C, D}~~		

Figura 2: En este ejemplo, el algoritmo Apriori solo evaluó 7 de los 15 conjuntos de elementos potenciales que pueden aparecer en datos transaccionales para cuatro elementos (el conjunto de elementos de 0 elementos no se muestra).

En este punto, puede comenzar la segunda fase del algoritmo Apriori. Dado el conjunto de conjuntos de elementos frecuentes, se generan reglas de asociación a partir de todos los subconjuntos posibles. Por ejemplo, {A, B} daría como resultado reglas candidatas para {A} → {B} y {B} → {A}. Estas se evalúan en relación con un umbral de confianza mínimo y se elimina cualquier regla que no cumpla con el nivel de confianza deseado.

Analizaremos un ejemplo para entender el algoritmo.

Resumen

Las reglas de asociación se utilizan para obtener información sobre las enormes bases de datos de transacciones de los grandes minoristas.

Como proceso de aprendizaje no supervisado, los aprendices de reglas de asociación son capaces de extraer conocimiento de grandes bases de datos sin ningún conocimiento previo de qué patrones buscar. El problema es que se necesita un cierto esfuerzo para reducir la riqueza de información a un conjunto de resultados más pequeño y manejable. El algoritmo Apriori, que estudiamos en este documento (después de revisar el ejemplo), lo hace estableciendo umbrales mínimos de interés e informando solo las asociaciones que cumplen estos criterios.

Pusimos en práctica el algoritmo Apriori mientras realizábamos un análisis de la cesta de la compra de un mes de transacciones en un supermercado de tamaño modesto. Incluso en este pequeño ejemplo, se identificó una gran cantidad de asociaciones. Entre ellas, notamos varios patrones que pueden ser útiles para futuras campañas de marketing. Los mismos métodos que aplicamos se utilizan en minoristas mucho más grandes en bases de datos de muchas veces este tamaño, y también se pueden aplicar a proyectos fuera del ámbito minorista.

En el próximo documento, examinaremos otro algoritmo de aprendizaje no supervisado. Al igual que las reglas de asociación, su finalidad es encontrar patrones dentro de los datos. Pero a diferencia de las reglas de asociación que buscan grupos de elementos o características relacionadas, los métodos del siguiente documento se ocupan de encontrar conexiones y relaciones entre los ejemplos.