

Inteligencia Artificial & Machine Learning

Aplicaciones en movilidad



Reglas de asociación

Algoritmo Apriori



Reglas de asociación

- Encontrar asociaciones o correlaciones entre los elementos u objetos de bases de datos transaccionales
- Las reglas de asociación tienen diversas aplicaciones como:
 - Análisis de información de ventas
 - Distribución de mercancías en tiendas
 - Segmentación de clientes con base en patrones de compra

Reglas de asociación - Market Basket Analysis / Análisis de Canasta de Mercado



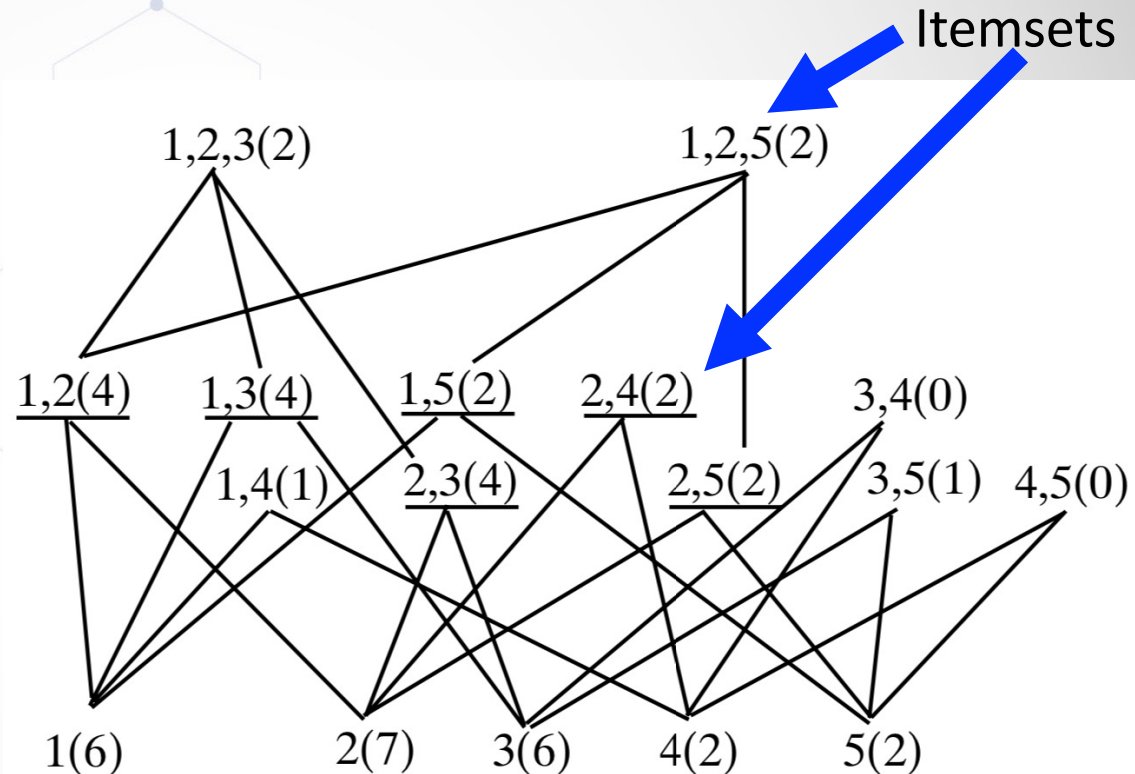
- Encontrar asociaciones entre los productos adquiridos por los clientes, que pueden impactar a las estrategias de mercadotecnia

Reglas de asociación - Idea general

id1	p1,p2,p5
id2	p2,p4
id3	p2,p3
id4	p1,p2,p4
id5	p1,p3
id6	p2,p3
id7	p1,p3
id8	p1,p2,p3,p5
id9	p1,p2,p3

Reglas de asociación - Idea general

id1	p1,p2,p5
id2	p2,p4
id3	p2,p3
id4	p1,p2,p4
id5	p1,p3
id6	p2,p3
id7	p1,p3
id8	p1,p2,p3,p5
id9	p1,p2,p3



Encontrar los itemsets más frecuentes

Reglas de asociación – Conceptos básicos

- Conjunto de transacciones \mathbb{D}

Transacción	Elementos Comprados
1	A,B,C
2	A,C
3	A,D
4	B,E,F

$$C \Rightarrow A$$

Si **C** Entonces **A**

Reglas de asociación – Conceptos básicos

Transacción	Elementos Comprados
1	A,B,C
2	A,C
3	A,D
4	B,E,F

$$C \Rightarrow A$$

- Una regla de asociación es una expresión de la forma $X \Rightarrow Z$
- El significado intuitivo:
 - Las transacciones de la base de datos que contienen X tienden a contener Z

Reglas de asociación – Soporte y confianza

- Soporte:
 - $A \Rightarrow C$: Porcentaje de las transacciones en que aparece el conjunto $\{A, C\}$
- Confianza:
 - $A \Rightarrow C$: De las transacciones en las que aparece A, porcentaje en las que también aparece C
 - $C \Rightarrow A$: De las transacciones en las que aparece C, porcentaje en las que también aparece A

Transacción	Elementos Comprados
1	A,B,C
2	A,C
3	A,D
4	B,E,F

$A \Rightarrow C$ (50% , 66%)

$C \Rightarrow A$ (50% , 100%)



soporte



confianza

Algoritmo Apriori

Notación

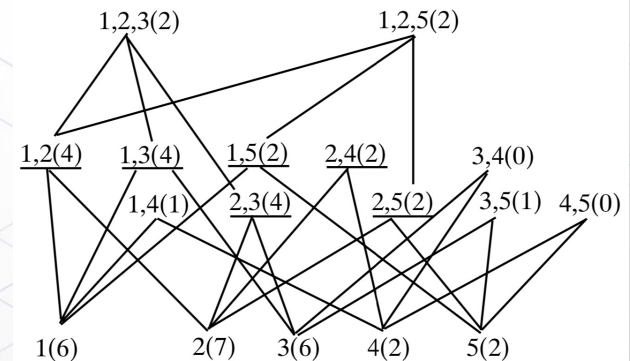
- $X = \{i_1, i_2, \dots, i_m\}$: conjunto (universo) de ítems
- D : Conjunto de transacciones
- T : Transacción que consiste de un grupo de ítems, i.e. $T \subseteq X$
- k -itemset: Un conjunto de k ítems
- C_k : Conjunto de k -itemsets candidatos
- L_k : Conjunto de k -itemsets “finales”

Fast Algorithms for Mining Association Rules

Rakesh Agrawal

Ramakrishnan Srikant*

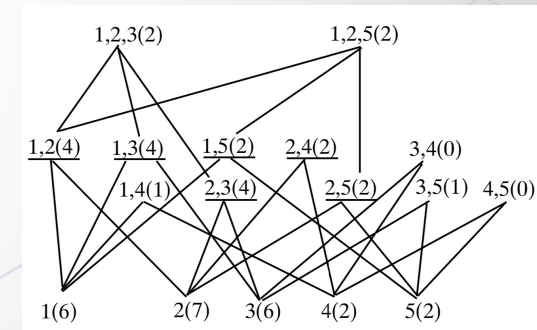
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120



Idea general

1. Genera los **itemsets**

- Genera todos los itemsets con **un elemento**
 - Toma todos los que cumplen con el **mínimo soporte**
- Usa esos itemsets para generar los de **dos elementos**,
 - Toma todos los que cumplen con el **mínimo soporte**
- Usa los nuevos itemsets para generar los de tres elementos y así sucesivamente



2. Genera las **reglas** revisando que cumplan con el criterio **mínimo de confianza**

Ejemplo

 \mathbb{D}

ID	Items
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E
T5	A, C, E

 $X = \{???\}$

Ejemplo - Generación de itemsets

 \mathbb{D}

ID	Items
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E
T5	A, C, E

 C_1

Itemset	Soporte
{A}	3/5
{B}	3/5
{C}	4/5
{D}	1/5
{E}	4/5

Ejemplo - Generación de itemsets

 \mathbb{D}

ID	Items
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E
T5	A, C, E

 C_1

Itemset	Soporte
{A}	3/5
{B}	3/5
{C}	4/5
{D}	1/5
{E}	4/5

 L_1

Itemset	Soporte
{A}	3/5
{B}	3/5
{C}	4/5
{E}	4/5

Soporte mínimo = 2/5

Ejemplo - Generación de itemsets

 \mathbb{D}

ID	Items
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E
T5	A, C, E

 C_2

Itemset	Soporte
{A,B}	1/5
{A,C}	3/5
{A,E}	2/5
{B,C}	2/5
{B,E}	3/5
{C,E}	3/5

 L_2

Itemset	Soporte
{A,C}	3/5
{A,E}	2/5
{B,C}	2/5
{B,E}	3/5
{C,E}	3/5

Ejemplo - Generación de itemsets

 \mathbb{D}

ID	Items
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E
T5	A, C, E

 \mathcal{C}_3

Itemset	Soporte
{A,B,C}	1/5
{A,B,E}	1/5
{A,C,E}	2/5
{B,C,E}	2/5

 L_3

Itemset	Soporte
{A,C,E}	2/5
{B,C,E}	2/5

Ejemplo - Generación de itemsets

 \mathbb{D}

ID	Items
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E
T5	A, C, E

 \mathcal{C}_4

Itemset	Soporte
{A,B,C,E}	1/5

 L_4

Itemset	Soporte

Ejemplo - Generación de itemsets

 \mathbb{D}

ID	Items
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E
T5	A, C, E

 L_3

Itemset	Soporte
{A,C,E}	2/5
{B,C,E}	2/5

Ejemplo - Generación de reglas

- Para cada conjunto **I** de items, genera todos sus subconjuntos **S**
- Para cada subconjunto $\mathbf{s} \subset \mathbf{I}$, genera una regla: $\mathbf{s} \Rightarrow (\mathbf{I} - \mathbf{s})$ si

$$\frac{\text{soporte}(l)}{\text{soporte}(s)} \geq \text{nivel de confianza}$$

- Todas las reglas satisfacen los niveles mínimos de soporte

Ejemplo - Generación de reglas

 \mathbb{D}

ID	Items
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E
T5	A, C, E

 L_{final}

Itemset	Soporte
{A,C,E}	2/5
{B,C,E}	2/5

$$\frac{\text{soporte}(l)}{\text{soporte}(s)} \geq 60\%$$

Generación de reglas – primer itemset

 \mathbb{D}

ID	Items
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E
T5	A, C, E

Itemset	Subconjuntos
{A,C,E}	{A,C}, {A,E}, {C,E}, {A}, {C}, {E}

- Regla 1: $\{A,C\} \Rightarrow (\{A,C,E\} - \{A,C\})$
 - $\{A,C\} \Rightarrow E$
 - Confianza = soporte (A,C,E) / soporte (A,C) = 2 / 3 = 66.66%
- Regla 2: $\{A,E\} \Rightarrow (\{A,C,E\} - \{A,E\})$
 - $\{A,E\} \Rightarrow C$
 - Confianza = soporte (A,C,E) / soporte (A,E) = 2 / 2 = 100%
- Regla 3: $\{C,E\} \Rightarrow (\{A,C,E\} - \{C,E\})$
 - $\{C,E\} \Rightarrow A$
 - Confianza = soporte (A,C,E) / soporte (C,E) = 2 / 3 = 66.66%

Generación de reglas – primer itemset

 \mathbb{D}

ID	Items
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E
T5	A, C, E

Itemset	Subconjuntos
{A,C,E}	{A,C}, {A,E}, {C,E}, {A}, {C}, {E}

- Regla 4: $\{A\} \Rightarrow (\{A,C,E\} - \{A\})$
 - $\{A\} \Rightarrow \{C,E\}$
 - Confianza = soporte (A,C,E) / soporte (A) = 2 / 3 = 66.66%
- Regla 5: $\{C\} \Rightarrow (\{A,C,E\} - \{C\})$
 - $\{C\} \Rightarrow \{A,E\}$
 - Confianza = soporte (A,C,E) / soporte (C) = 2 / 4 = 50%
- Regla 6: $\{E\} \Rightarrow (\{A,C,E\} - \{E\})$
 - $\{E\} \Rightarrow \{A,C\}$
 - Confianza = soporte (A,C,E) / soporte (E) = 2 / 4 = 50%

Generación de reglas – segundo itemset

 \mathbb{D}

ID	Items
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E
T5	A, C, E

Itemset	Subconjuntos
{B,C,E}	{B,C}, {B,E}, {C,E}, {B}, {C}, {E}

- Regla 7: $\{B,C\} \Rightarrow (\{B,C,E\} - \{B,C\})$
 - $\{B,C\} \Rightarrow E$
 - Confianza = soporte (B,C,E) / soporte (B,C) = 2 / 2 = 100%
- Regla 8: $\{B,E\} \Rightarrow (\{B,C,E\} - \{B,E\})$
 - $\{B,E\} \Rightarrow C$
 - Confianza = soporte (B,C,E) / soporte (B,E) = 2 / 3 = 66.66%
- Regla 9: $\{C,E\} \Rightarrow (\{B,C,E\} - \{C,E\})$
 - $\{C,E\} \Rightarrow B$
 - Confianza = soporte (B,C,E) / soporte (C,E) = 2 / 3 = 66.66%

Generación de reglas – segundo itemset

 \mathbb{D}

ID	Items
T1	A, C, D
T2	B, C, E
T3	A, B, C, E
T4	B, E
T5	A, C, E

Itemset	Subconjuntos
{B,C,E}	{B,C}, {B,E}, {C,E}, {B}, {C}, {E}

- Regla 10: $\{B\} \Rightarrow (\{A,C,E\} - \{B\})$
 - $\{B\} \Rightarrow \{C,E\}$
 - Confianza = soporte (B,C,E) / soporte (B) = $2 / 3 = 66.66\%$
- Regla 11: $\{C\} \Rightarrow (\{B,C,E\} - \{C\})$
 - $\{C\} \Rightarrow \{B,E\}$
 - Confianza = soporte (B,C,E) / soporte (C) = $2 / 4 = 50\%$
- Regla 12: $\{E\} \Rightarrow (\{B,C,E\} - \{E\})$
 - $\{E\} \Rightarrow \{B,C\}$
 - Confianza = soporte (B,C,E) / soporte (E) = $2 / 4 = 50\%$

Ejemplo - Python

```
import pandas as pd
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori, association_rules
```

Datos

```
dataset = [['A', 'C', 'D'],
           ['B', 'C', 'E'],
           ['A', 'B', 'C', 'E'],
           ['B', 'E'],
           ['A', 'C', 'E']]
```

```
te = TransactionEncoder()
te_ary = te.fit(dataset).transform(dataset)
df = pd.DataFrame(te_ary, columns=te.columns_)
```

Genera los itemsets

```
items_frecuentes = apriori(df, min_support=0.4, use_colnames=True)
```

Genera las reglas

```
reglas = association_rules(items_frecuentes, metric="confidence", min_threshold = 0.6)
reglas = reglas.sort_values(['confidence'], ascending = [False])
reglas.head()
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(A)	(C)	0.6	0.8	0.6	1.00	1.25	0.12	inf	0.500000
5	(B)	(E)	0.6	0.8	0.6	1.00	1.25	0.12	inf	0.500000
8	(E, A)	(C)	0.4	0.8	0.4	1.00	1.25	0.08	inf	0.333333
14	(B, C)	(E)	0.4	0.8	0.4	1.00	1.25	0.08	inf	0.333333
1	(C)	(A)	0.8	0.6	0.6	0.75	1.25	0.12	1.6	1.000000

Código



Ejemplo - Python

items_frecuentes

	support	itemsets
0	0.6	(A)
1	0.6	(B)
2	0.8	(C)
3	0.8	(E)
4	0.6	(A, C)
5	0.4	(E, A)
6	0.4	(B, C)
7	0.6	(E, B)
8	0.6	(E, C)
9	0.4	(E, A, C)
10	0.4	(E, B, C)

Ejemplo - Python

reglas

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(A)	(C)	0.6	0.8	0.6	1.000000	1.250000	0.12	inf	0.500000
5	(B)	(E)	0.6	0.8	0.6	1.000000	1.250000	0.12	inf	0.500000
8	(E, A)	(C)	0.4	0.8	0.4	1.000000	1.250000	0.08	inf	0.333333
14	(B, C)	(E)	0.4	0.8	0.4	1.000000	1.250000	0.08	inf	0.333333
1	(C)	(A)	0.8	0.6	0.6	0.750000	1.250000	0.12	1.6	1.000000
4	(E)	(B)	0.8	0.6	0.6	0.750000	1.250000	0.12	1.6	1.000000
6	(E)	(C)	0.8	0.8	0.6	0.750000	0.937500	-0.04	0.8	-0.250000
7	(C)	(E)	0.8	0.8	0.6	0.750000	0.937500	-0.04	0.8	-0.250000
2	(A)	(E)	0.6	0.8	0.4	0.666667	0.833333	-0.08	0.6	-0.333333
3	(B)	(C)	0.6	0.8	0.4	0.666667	0.833333	-0.08	0.6	-0.333333
9	(E, C)	(A)	0.6	0.6	0.4	0.666667	1.111111	0.04	1.2	0.250000
10	(A, C)	(E)	0.6	0.8	0.4	0.666667	0.833333	-0.08	0.6	-0.333333
11	(A)	(E, C)	0.6	0.6	0.4	0.666667	1.111111	0.04	1.2	0.250000
12	(E, B)	(C)	0.6	0.8	0.4	0.666667	0.833333	-0.08	0.6	-0.333333
13	(E, C)	(B)	0.6	0.6	0.4	0.666667	1.111111	0.04	1.2	0.250000
15	(B)	(E, C)	0.6	0.6	0.4	0.666667	1.111111	0.04	1.2	0.250000

Lift

$$\text{lift}(X \rightarrow Y) = \frac{\text{confianza}(X \rightarrow Y)}{\text{soporte}(Y)}$$

1. Cuando lift es mayor a 1, Y tiene alta probabilidad de ser comprado cuando se adquiere X
2. Cuando lift es menor a 1, Y tiene poca probabilidad de ser comprado cuando se adquiere X
3. Cuando lift es igual a 1, Y y X son independientes

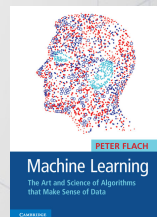
Algunas de las razones del porqué funciona Apriori

- Se enfoca en los (sub)conjuntos más frecuentes
- Si un subconjunto es poco frecuente, los super-conjuntos de dicho subconjunto también serán poco frecuentes
- Las transacciones suelen seguir leyes de potencia
 - P.ej. Google, Amazon,
 - Regla 80 : 20
 - El 80% de tus clientes está interesado en el 20% de tus productos / servicios
 - "80% of sales come from 20% of clients"
 - "Microsoft noted that by fixing the top 20% of the most-reported bugs, 80% of the related errors and crashes in a given system would be eliminated"

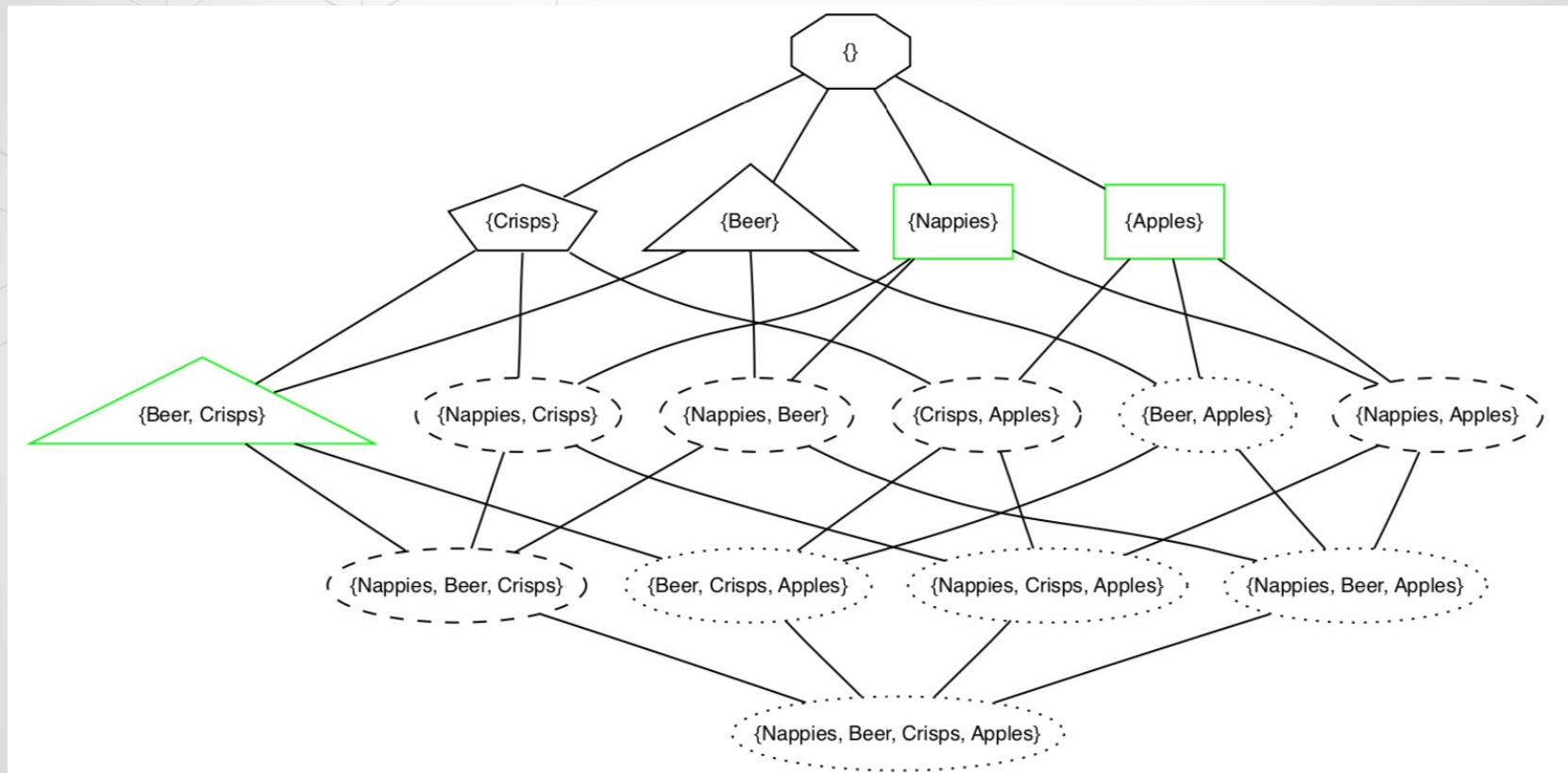
Reglas de asociación - Wrap up

<i>Transaction</i>	<i>Items</i>
1	nappies
2	beer, crisps
3	apples, nappies
4	beer, crisps, nappies
5	apples
6	apples, beer, crisps, nappies
7	apples, crisps
8	crisps

<http://people.cs.bris.ac.uk/~flach/mlbook/>



Reglas de asociación - Wrap up



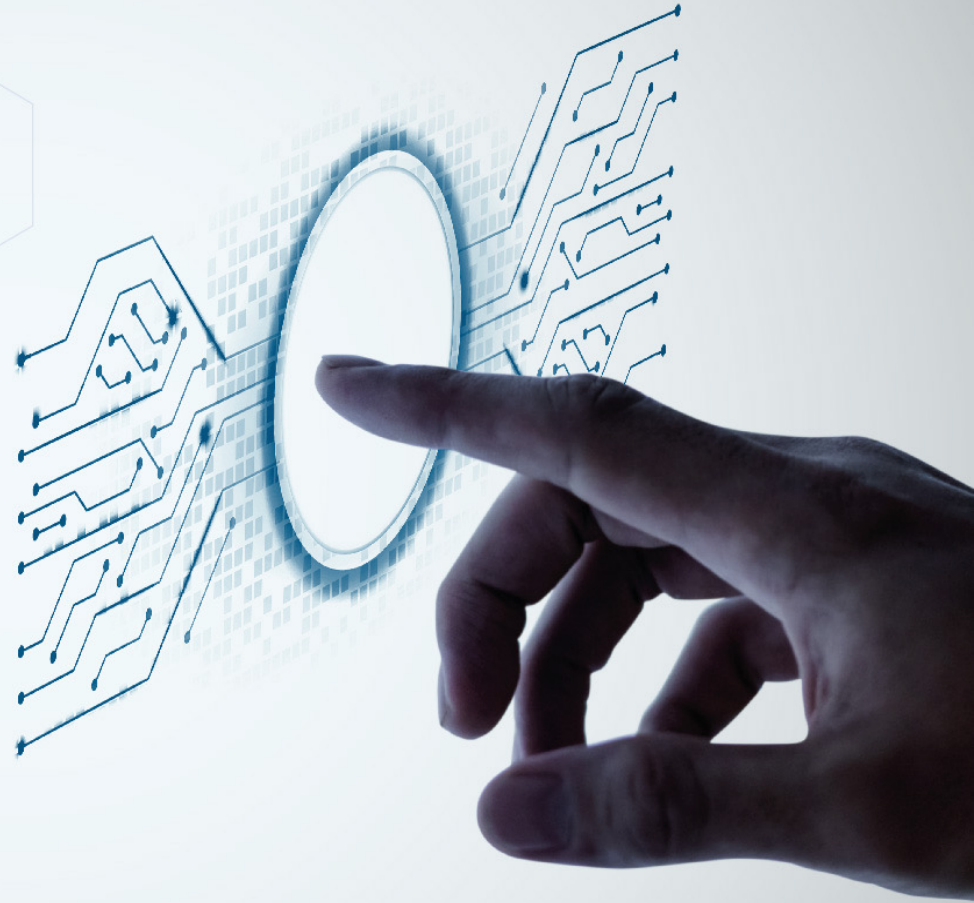
Código



EngineeringX

Founded by the Royal Academy of Engineering
and Lloyd's Register Foundation

GRACIAS



<https://hubiq.mx/>

 HUBIQRO  HUBIQ  HUBIQRO