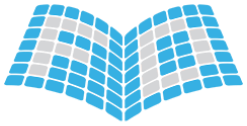


آموزش یادگیری ماشین (Machine Learning)
(تئوری – عملی) - بخش دوم

درس هشتم: کاوش قوانین انجمنی

مدرس:
فرشید شیرافکن
دانشجوی دکترای بیو انفورماتیک
دانشگاه تهران

مقدمه

- کشف قوانین انجمنی از دسته روش‌های بدون ناظر است.
- می خواهیم بدانیم که یک مجموعه اشیاء خاص بر وجود چه مجموعه اشیاء دیگر اثرگذار است.
- هدف کاوش قوانین انجمنی: پیدا کردن نظم و قوانین حاکم بر داده‌ها می‌باشد.
- کشف قوانین انجمنی، درباره علت وجود رابطه مجموعه اشیاء چیزی نمی‌گوید.

تحلیل سبد خرید
(Market Basket Analysis)

"تحلیل سبد خرید" از کاربردهای متداول در رابطه با کشف قوانین انجمنی است.

با توجه به اقلام مختلفی که مشتریان در سبد خریدشان قرار می‌دهند، عادات و رفتار خرید مشتریان مورد تحلیل قرار گرفته و الگوهای موجود در اقلام خریداری شده کشف می‌شود.

مثال: شیر ⟵ پنیر

(۱۰٪ = پشتیبانی و ۸۰٪ = اطمینان)

۱۰ ٪ مشتری‌ها پنیر و شیر را با هم خریداری می‌کنند.

مشتریانی که پنیر می‌خرند در **۸۰٪** موارد شیر نیز خریداری می‌کنند.

بیان مساله کاوش قوانین انجمنی

$$I = \{i_1, i_2, \ldots, i_n\}$$

set of transactions called the *database*

$$D = \{t_1, t_2, \ldots, t_m\}$$

$$X \Rightarrow Y \qquad X, Y \subseteq I$$

TID	items
1	{I1,I3,I4}
2	{I2,I3,I5}
3	{I1,I2,I3,I5}
4	{I2,I5}

- الگوهای مکرر (frequent pattern): الگوهایی که در یک بانک داده زیاد رخ می‌دهد.

پشتیبانی

AR : X → Y

$$Support = \frac{frq(X,Y)}{N}$$

TID	items
1	{I1,I3,I4}
2	{I2,I3,I5}
3	{I1,I2,I3,I5}
4	{I2,I5}

I1 → I3

$$s = \frac{2}{4} = 50\%$$

اطمینان

AR : X → Y

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$conf(X \text{-->} Y) = \frac{supp(X \cup Y)}{supp(X)}$$

TID	items
1	{I1,I3,I4}
2	{I2,I3,I5}
3	{I1,I2,I3,I5}
4	{I2,I5}

I1 → I3

$$c = \frac{2}{2} = 100\%$$

مثال

Tid	Items bought
1	I1, I2, I3
2	I1, I2, I4
3	I1, I2, I5
4	I3, I5, I6
5	I2, I3, I4, I5, I6

I1 → I2

$$s = \frac{3}{5} = 60\%$$

$$c = \frac{3}{3} = 100\%$$

I2 → I1

$$s = \frac{3}{5} = 60\%$$

$$c = \frac{3}{4} = 75\%$$

مثال

Tid	Items bought
1	I1, I2
2	I1, I3, I4, I5
3	I2, I3, I4, I6
4	I1, I2, I3, I4
5	I1, I2, I3, I6

{I2 , I3} → I4

$$s = \frac{2}{5} = 40\%$$

$$c = \frac{2}{3} = 0.67\%$$

قانون قوی

TID	items
1	{I1, I3, I4}
2	{I2, I3, I5}
3	{I1, I2, I3, I5}
4	{I2, I5}

minsup = 50%

minconf = 75%

{I2, I3} → I5

$s = \frac{2}{4} = 50\%$

$c = \frac{2}{2} = 100\%$

{I2} → {I3, I5}

$s = \frac{2}{4} = 50\%$

$c = \frac{2}{3} = 66\% < 75\%$

الگوریتم Apriori

الگوریتم Apriori ، روشی برای کشف قوانین انجمنی است که شامل دو مرحله است:

مرحله اول :

تولید مجموعه اشیاء مکرر با روش تکراری

s >= minsup

مرحله دوم :

c >= minconf

تولید قانون (ساختن تمام زیرمجموعه‌های ممکن قوانین ، به جز مجموعه‌های تهی)

الگوریتم Apriori

۱- k=1

۲- ایجاد مجموعه اشیاء مکرر با طول یک.

۳- تکرار مراحل زیر تا زمانی که هیچ مجموعه شیء مکرر پیدا نشود:

- k=k+1
- پیدا کردن مجموعه اشیاء کاندید با طول k از مجموعه اشیاء مکرر با طول k-1.
- محاسبه ساپورت (s) هر کاندید با اسکن بانک داده.
- هرس کاندیدهای نامکرر.

مثال

Tid	Items
1	A,B,E
2	B,D
3	B,C
4	A,B,D
5	A,C
6	B,C
7	A,C
8	A,B,C,E
9	A,B,C

minsup = 2

minconf = 70%

مرحله اول: تولید مجموعه اشیاء مکرر

Tid	Items
1	A,B,E
2	B,D
3	B,C
4	A,B,D
5	A,C
6	B,C
7	A,C
8	A,B,C,E
9	A,B,C

Itemset	sup
{A}	6
{B}	7
{C}	6
{D}	2
{E}	2

Itemset
{A}
{B}
{C}
{D}
{E}

$k=1$

Itemset	sup
{A, B}	4
{A, C}	4
{A, D}	1
{A, E}	2
{B,C}	4
{B,D}	2
{B,E}	2
{C,D}	0
{C,E}	1
{D,E}	0

Itemset
{A, B}
{A, C}
{A, E}
{B,C}
{B,D}
{B,E}

$k=2$

Itemset	sup
{A,B,C}	2
{A,B,D}	1
{A,B,E}	2
{A,C,D}	0
{A,C,E}	1
{A,D,E}	0
{B,C,D}	0
{B,C,E}	1
{B,D,E}	0

$k=3$

Itemset
{A,B,C}
{A,B,E}

مرحله دوم : تولید قانون

Tid	Items
1	A,B,E
2	B,D
3	B,C
4	A,B,D
5	A,C
6	B,C
7	A,C
8	A,B,C,E
9	A,B,C

{A,B,E}:

$2^3 - 2 = 6$

1 A → {B,E} → $c = \frac{2}{6}$

2 B → {A,E} → $c = \frac{2}{7}$

3 E → {A,B} → $c = \frac{2}{2}$

4 {A,B} → E → $c = \frac{2}{4}$

5 {A,E} → B → $c = \frac{2}{2}$

6 {B,E} → A → $c = \frac{2}{2}$

مثال

Tid	Items
1	A, C, D
2	B, C, E
3	A, B, C, E
4	B, E

minsup = 2

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

Itemset
{A}
{B}
{C}
{E}

K=1

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

Itemset
{A, C}
{B, C}
{B, E}
{C, E}

K=2

Itemset	sup
{B, C, E}	2

Itemset
{B, C, E}

K=3

1. $B \rightarrow CE$
2. $C \rightarrow BE$
3. $E \rightarrow BC$
4. $CE \rightarrow B$
5. $CB \rightarrow E$
6. $BE \rightarrow C$

مثال

Tid	Items
1	M1,M2,M5
2	M2,M4
3	M2,M3
4	M1,M2,M4
5	M1,M3
6	M2,M3
7	M1,M3
8	M1,M2,M3,M5
9	M1,M2,M3

$\{M_1, M_5\} \rightarrow M_2$

$\{M_2, M_5\} \rightarrow M_1$

برای $\begin{cases} S = \frac{2}{9} \\ C = \frac{2}{9} \end{cases}$

تمرین

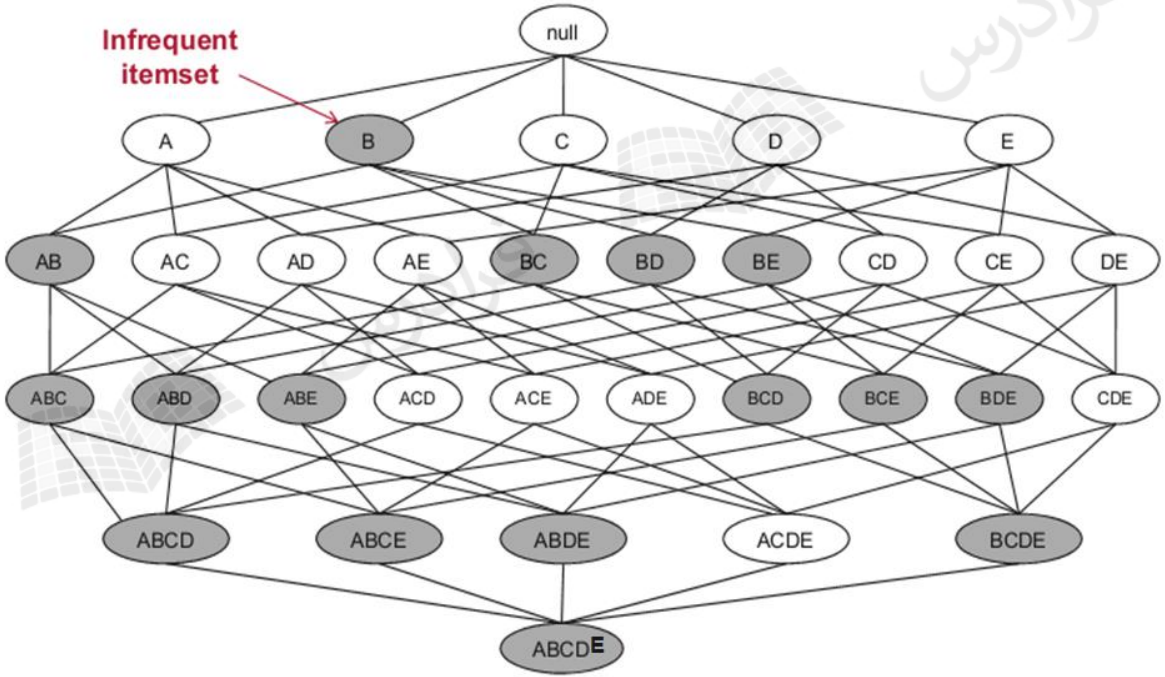
t1: Beef, Chicken, Milk
t2: Beef, Cheese
t3: Cheese, Boots
t4: Beef, Chicken, Cheese
t5: Beef, Chicken, Clothes, Cheese, Milk
t6: Chicken, Clothes, Milk
t7: Chicken, Milk, Clothes

Tid	Items
t1	A,C,E
t2	A,B
t3	B,F
t4	A,C,B
t5	A,C,D,B,E
t6	C,D,E
t7	C,E,D

minsup = 30 %
minconf = 80%

هرس

- همه ابر مجموعه‌های مربوط به مجموعه شیء نامکرر (infrequent) از شبکه مجموعه اشیاء حذف می‌شوند.

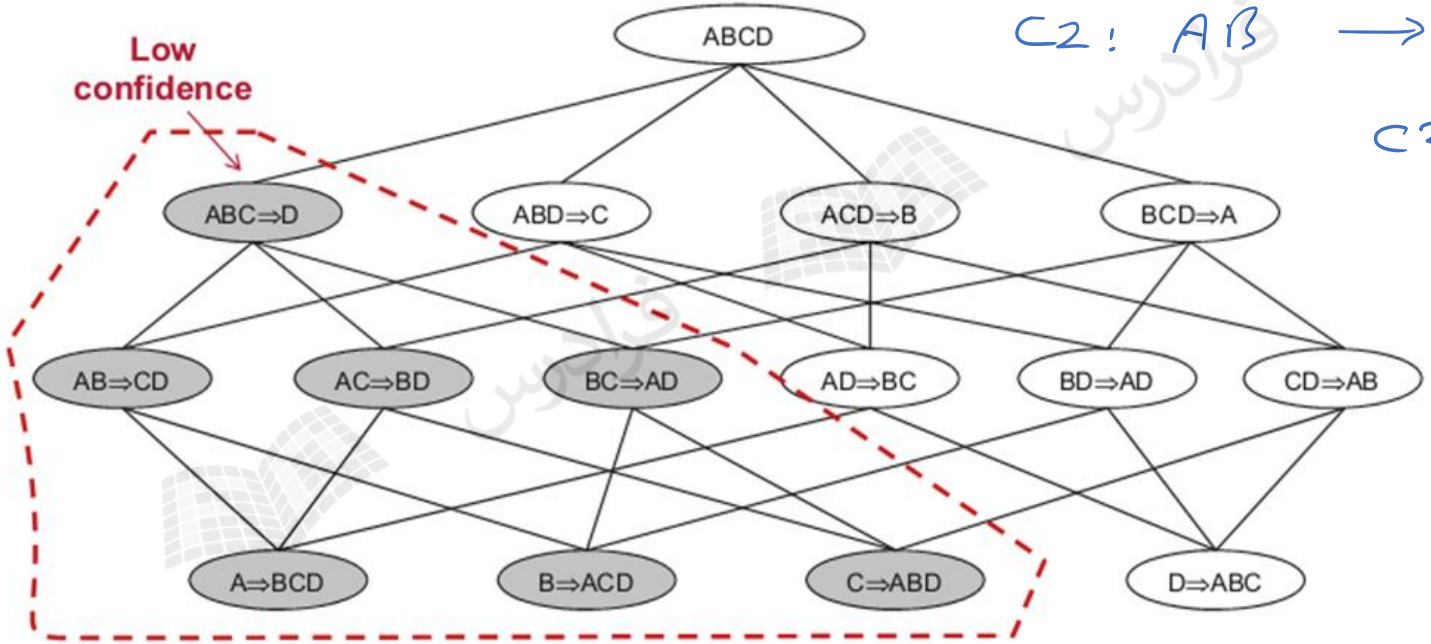


هرس قانون

C1 : ABC → D

C2 : AB → CD

C2 < C1



مزایا و معایب الگوریتم Apriori

مزیت:

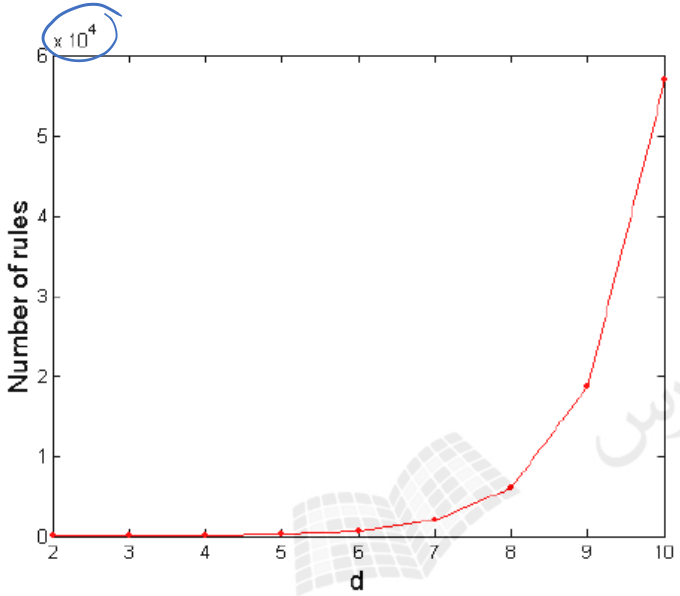
- پیاده‌سازی آن ساده است.

معایب:

- در هر دور اجرای الگوریتم، کل تراکنش‌ها پیمایش می شود.
- تراکنش‌ها در حافظه اصلی ذخیره می شود.

تعداد کل قوانین انجمنی قابل استخراج

$$\binom{3}{1}\left[\binom{2}{1}+\binom{2}{2}\right]+\binom{3}{2}\left[\binom{1}{1}\right]=12$$



$$R=\sum_{k=1}^{d-1}\left[\binom{d}{k}\times\sum_{j=1}^{d-k}\binom{d-k}{j}\right]=3^d-2^{d+1}+1$$

$d=3$ $\{A,B,C\}$

d	
3	12
6	
10	

1. $A\rightarrow B$
2. $B\rightarrow C$
3. $C\rightarrow A$
4. $B\rightarrow A$
5. $C\rightarrow B$
6. $A\rightarrow C$
7. $A\rightarrow BC$
8. $B\rightarrow AC$
9. $C\rightarrow AB$
10. $BC\rightarrow A$
11. $AC\rightarrow B$
12. $AB\rightarrow C$

d
2^d-2

ارزیابی قوانین انجمنی

گاهی معیار درصد پشتیبانی و اطمینان مناسب نیستند.

	Basketball	Not Basketball	
Milk	2000	1750	3750
Not Milk	1000	250	1250
	3000	2000	5000

S
C
→ lift

$\frac{3750}{5000} \simeq 75\%$

B → M

$S = \frac{2000}{5000} = 40\%$

$C = \frac{2000}{3000} \simeq 66\%$

B → ~M

$S = \frac{1000}{5000} = 20\%$

$C = \frac{1000}{3000} \simeq 33\%$

معیار میزان وابستگی

$A \rightarrow B$

$$lift = \frac{P(A \cup B)}{P(A)P(B)}$$

- نشان دهنده تاثیر بالا رفتن یکی در بالا رفتن دیگری

- Lift > 1: وابستگی مثبت (رخداد یکی باعث رخداد دیگری است)
- Lift < 1: وابستگی منفی (رخداد یکی باعث رخ ندادن دیگری است.)
- Lift = 1: یعنی A و B از هم مستقل هستند.

مثال

	Basketball	Not Basketball
Milk	2000	1750
Not Milk	1000	250

B → M

$$lift(B \rightarrow M) = \frac{P(B \cup M)}{P(B) \cdot P(M)} = \frac{\frac{2000}{5000}}{\frac{3000}{5000} \times \frac{3750}{5000}}$$

B → ~M

$$lift(B \rightarrow \neg M) = \frac{\frac{1000}{5000}}{\frac{3000}{5000} \times \frac{1250}{5000}} = 0.89 < 1$$ وابستگی منفی

$$= 1.33 > 1$$ وابستگی مثبت

مثال

قانون A → B را ارزیابی کنید.

- تعداد تراکنش: **100**
- (A): خرید بازی کامپیوتری: **60**
- (B): خرید کارت گرافیک: **75**
- شامل هر دو: **40**

$$\begin{cases} S = \frac{40}{100} = 40\% \\ c = \frac{40}{60} = 66\% \end{cases}$$

$$\frac{75}{100} = 75\%$$

$$lift = \frac{\frac{40}{100}}{\frac{60}{100} \times \frac{75}{100}} < 1$$

وابستگی منفی

خرید یکی باعث کاهش احتمال خرید دیگری است.

minsup = 30%

minconf = 60%

مشاوره فرشید شیرافکن:

۰۹۱۲۱۹۷۲۰۲۸

این اسلایدها بر مبنای نکات مطرح شده در فرادرس
«آموزش یادگیری ماشین (Machine Learning) (تئوری - عملی) – بخش دوم»
تهیه شده است.

برای کسب اطلاعات بیشتر در مورد این آموزش به لینک زیر مراجعه نمایید.

faradars.org/fvdm94062