Miksi kaupat sijoittelevat tietyt tuotteet vierekkäin?

Jos saisit analysoida miljoonia ostoskuitteja, mitä yllättäviä yhdistelmiä uskoisit löytäväsi?

TAITOTALO

Assosiaatioanalyysi

Kerttuli Ratilainen

Syksy 2025

Missä käytetään

- Ostoskorianalyysi: Mitä tuotteita ostetaan usein yhdessä?
- Tuotesijoittelu: Mikä tuote B ostetaan todennäköisemmin, jos on jo ostettu tai katseltu tuotetta A?
- Tuotesuositukset: Mitä muut ovat suositelleet?

Assosiaatioanalyysi - apriori

- Soveltuu nominaaliasteikon muuttujille (esim. vihreä, punainen, sininen) -> ei voi soveltaa esimerkiksi laskutoimituksia, eivätkä luokat ole järjestyksessä
- Koostuu tapahtumista (*transactions*)
- Jokaisessa tapahtumassa on tuotteita (*items*)
- Tavoitteena on löytää tuotteita, jotka esiintyvät usein samassa tapahtumassa
- Tavoitteeseen johtavia malleja kutsutaan assosiaatiosäännöiksi tai assosiaatiosääntöjen louhinnaksi
- Assosiaatiosääntö ei takaa kausaliteettia (syy-seuraus-suhdetta) eli ensimmäinen tuote ei automaattisesti takaa toista tuotetta samassa tapahtumassa, vaikka yhteys tuotteiden välillä havaitaan.

Assosiaatiosäännöt

Jokaisella säännöllä on suunta:

- Alkuehto, LHS (left hand side), on ehto, jonka perusteella sääntöä tarkastellaan
- Seuraus, RHS (right hand side), on ehto, joka assosioituu alkuehtoon

Sääntöä luetaan vasemmalta oikealle

Esimerkki 1

”Jos asiakas ostaa leipää, hän todennäköisesti ostaa myös juustoa.”

”Jos asiakas ostaa hammasharjan ja hammastahnan, hän todennäköisesti ostaa myös hammaslankaa.”

{Leipä} -> {Juusto}

{Hammasharja, Hammastahna} -> {Hammaslanka}

Vasemmalla on asiakkaan ostama, katselema tuote

Oikealla kiinnostava tuote, ominaisuus tai toiminta

TAITOTALO

Esimerkki 2

Ostettu Tuote’:

1. Maito', 'Leipä', 'Juusto', 'Hedelmät’
2. ’Leipä', ’Juusto', 'Hedelmät',
3. 'Kahvi’, ’Leipä’, ’Juusto’
4. ’Kahvi, ’Maito', ’Leipä', ’Hedelmät',
5. 'Maito', 'Leipä', 'Juusto',
6. ’Maito’, ’Leipä’, ’Hedelmät',
7. 'Maito', 'Hedelmät'

Assosiaatiosääntöjä

- {Leipä} -> {Juusto}
- {Maito, Leipä} -> {Juusto}

Aineisto

- Seitsemän ostotapahtumaa (*transactions*)
- Ostotapahtumissa on yhteensä viisi tuotetta (*items*):
  - Maito
  - Leipä
  - Juusto
  - Hedelmät
  - Kahvi

Ostotapahtumat tuotteineen':	Maito', 'Leipä', 'Juusto', 'Hedelmät’
	’Leipä', ’Juusto', 'Hedelmät',
	'Kahvi’, ’Leipä’, ’Juusto’
	’Kahvi, ’Maito', ’Leipä', ’Hedelmät',
	'Maito', 'Leipä', 'Juusto',
	’Maito’, ’Leipä’, ’Hedelmät',
	'Maito', 'Hedelmät'

Tavoitteena

- Löytää assosiaatiosäännöt, joiden tuki ja luottamus ovat riittävän korkeat.
- Löytää oikeat parametrit algoritmille
- Tulkita tulosta: Onko tulos merkityksellinen vai sattumaa?

Tuki (support)

Ostotapahtumat:

1. Maito', 'Leipä', 'Juusto', 'Hedelmät’
2. ’Leipä', ’Juusto', 'Hedelmät',
3. 'Kahvi’, ’Leipä’, ’Juusto’
4. ’Kahvi, ’Maito', ’Leipä', ’Hedelmät',
5. 'Maito', 'Leipä', 'Juusto',
6. ’Maito’, ’Leipä’, ’Hedelmät',
7. 'Maito', 'Hedelmät'

Niiden tapahtumien osuus **kaikista tapahtumista**, joissa kaikki säännön tuotteet esiintyvät

Jos luku on matala, voi kyse olla sattumasta tai ilmiö on harvinainen, jos korkea, ilmiö on yleinen.

Esimerkiksi

- {hedelmät, leipä} -> {Juusto}
- Tuki 2/7 eli 0,29
- Sääntö esiintyy kahdessa tapauksessa seitsemästä

Luottamus (Confidence) 1

Ostotapahtumat:

1. Maito', 'Leipä', 'Juusto', 'Hedelmät’
2. ’Leipä', ’Juusto', 'Hedelmät',
3. 'Kahvi’, ’Leipä’, ’Juusto’
4. ’Kahvi, ’Maito', ’Leipä', ’Hedelmät',
5. 'Maito', 'Leipä', 'Juusto',
6. ’Maito’, ’Leipä’, ’Hedelmät',
7. 'Maito', 'Hedelmät'

Kuinka todennäköisesti säännön oikean puoleinen joukko havaitaan, jos vasemman puolen joukko on havaittu.

$$\text{confidence}(X \rightarrow Y) = P(Y \mid X)$$

Todennäköisyys, että Y tapahtuu, kun X on tapahtunut eli jos asiakas ostaa X:n kuinka todennäköisesti ostaa myös Y:n.

Luottamus (Confidence) 2

Ostotapahtumat:

1. Maito', 'Leipä', 'Juusto', 'Hedelmät’
2. ’Leipä', ’Juusto', 'Hedelmät',
3. 'Kahvi’, ’Leipä’, ’Juusto’
4. ’Kahvi, ’Maito', ’Leipä', ’Hedelmät',
5. 'Maito', 'Leipä', 'Juusto',
6. ’Maito’, ’Leipä’, ’Hedelmät',
7. 'Maito', 'Hedelmät'

Esimerkiksi

- {hedelmät, leipä} -> {Juusto}
- Luottamus 2/4 eli 0,5
- Sääntö esiintyy kahdessa tapauksessa, {hedelmät, leipä} lisäksi kahdessa muussa tapauksessa eli yhteensä neljässä

Lift (noste) 1

Ostotapahtumat:

1. Maito', 'Leipä', 'Juusto', 'Hedelmät’
2. ’Leipä', ’Juusto', 'Hedelmät',
3. 'Kahvi’, ’Leipä’, ’Juusto’
4. ’Kahvi, ’Maito', ’Leipä', ’Hedelmät',
5. 'Maito', 'Leipä', 'Juusto',
6. ’Maito’, ’Leipä’, ’Hedelmät',
7. 'Maito', 'Hedelmät'

Kuinka paljon sääntö on parempi verrattuna siihen, jos a ja b ovat täysin riippumattomia

Ostettaisiinko tuote muutenkin vai onko tuotteilla syy-yhteys

Lift = confidence(X → Y) / support(Y)

Lift (noste) 2

Ostotapahtumat:

1. Maito', 'Leipä', 'Juusto', 'Hedelmät’
2. ’Leipä', ’Juusto', 'Hedelmät',
3. 'Kahvi’, ’Leipä’, ’Juusto’
4. ’Kahvi, ’Maito', ’Leipä', ’Hedelmät',
5. 'Maito', 'Leipä', 'Juusto',
6. ’Maito’, ’Leipä’, ’Hedelmät',
7. 'Maito', 'Hedelmät'

Esimerkiksi

- {hedelmät, leipä} -> {Juusto}
- Luottamus 2/4 eli 0,5
- Tuki 4/7 eli 0,57 (juusto)
- Lift 0,5/0,57 = 0,86

Tulkinta

> 1 , X ja Y esiintyvät useammin kuin sattumalta, X lisää Y:n todennäköisyyttä

= 1, esiintyvät riippumattomasti, ei vaikutusta

< 1, X vähentää Y:n todennäköisyyttä

Laske tuki, luottamus ja noste

	Tuki (Support	**Luottamus (Confidence)**	**Noste (Lift)**
Leipä -> Juusto			
Juusto -> Leipä			
Kahvi -> Maito, Hedelmät			
Kahvi, Juusto -> Leipä			
Kahvi -> Juusto, Hedelmät			

Mittarit

Mittareiden avulla voidaan tarkastella kuinka yleinen, luotettava tai hyödyllinen sääntö

Kertauksena:

- Tuki (*support*), kuinka usein LHS ja RHS esiintyvät yhdessä aineistossa
- Luottamus (*confidence*), todennäköisyys, että sääntö toteutuu, kuinka usein alkuehto ja seuraus esiintyvät yhdessä ja kuinka usein taas alkuehto esiintyy yksin
- Noste (*lift*), kuinka paljon parempi sääntö on verrattuna siihen, että valinta olisi täysin riippumaton. Suurempi arvo kuin 1 tarkoittaa, että sääntö on hyödyllinen.

Haasteita

- Työläs tapa laskea yhteyksiä
- Parametrien määritys yrityksen ja erehdyksen kautta
- Tuloksen tulkinta, onko tuloksella merkitystä vai onko yhteys vain sattumaa
- Assosiaation automatisoituun etsintään on kehitetty algoritmeja
- Tässä esimerkkinä apriori

Apriori-algoritmi (vuodesta 1994)

- Sovelletaan erityisesti ostoskorianalyysiin
- Ensin etsitään kaikki tuotejoukot, joissa yksi tuote ja lasketaan niiden tuki (support)
- Sitten etsitään joukot, joissa kaksi tuotetta, sitten kolme, jne., jotka esiintyvät usein yhdessä
- Kun joukot on löydetty, rakennetaan niiden välille assosiaatiosäännöt ja suodatetaan pois ne joukot, joiden tuki ei ole riittävä
- Säännöistä valitaan ne, joiden luottamus on riittävä

- https://en.wikipedia.org/wiki/Apriori_algorithm

Esimerkki apriori-algoritmista

Ostettu Tuote’:

1. Maito', 'Leipä', 'Juusto', 'Hedelmät’
2. ’Leipä', ’Juusto', 'Hedelmät',
3. 'Kahvi’, ’Leipä’, ’Juusto’
4. ’Kahvi, ’Maito', ’Leipä', ’Hedelmät',
5. 'Maito', 'Leipä', 'Juusto',
6. ’Maito’, ’Leipä’, ’Hedelmät',
7. 'Maito', 'Hedelmät'

Säädetään miminitueksi 3/7 (=0,429) eli etsitään yleisiä joukkoja, jotka esiintyvät vähintään kolme kertaa

Säädetään minimiluottamukseksi 0,75

Lasketaan frekvenssit:

Frekvenssit, yksi tuote (tuki)

Tuotejoukko / frekvenssi

~~Kahvi 2~~

Juusto 4

Maito 5

Hedelmät 5

Leipä 6

Karsitaan tuotteet, joiden frekvenssi on alle minimituen (3), jonka jälkeen jokaisen tuoteryhmän frekvenssi on vähintään minimituen verran.

Yhdistetään tuotteet kahden tuotteen kokoisiksi joukoiksi.

Frekvenssi, kaksi tuotetta (tuki)

Tuotejoukko / frekvenssi

Maito, Leipä 4

~~**Maito, Juusto 2**~~

Maito, Hedelmät 4

Leipä, Juusto 4

Leipä, Hedelmät 4

~~**Juusto, Hedelmät 2**~~

Yhdistelmiä tulee kuusi.

Karsitaan tuotteet, joiden frekvenssi on alle minimituen (3), jonka jälkeen jokaisen tuoteryhmän frekvenssi on vähintään minimituen verran.

Yhdistetään tuotteet kolmen kokoisiksi joukoiksi.

Kustakin kolmen tuotteen joukosta tarkistetaan, että kaikki sen kahden suuruisilla osajoukoilla on frekvenssi yli minimituen.

Frekvenssi, kolme tuotetta (tuki)

Tuotejoukko / frekvenssi

Maito, Leipä, hedelmät 3

~~**Leipä, Juusto, Hedelmät 2**~~

~~**Maito, Leipä, juusto 2**~~

Jäljelle jää yksi tuotejoukko, jossa on kolme tuotetta.

Jäljellä olevan joukon frekvenssi on kolme eli se on riittävän yleinen

Neljän tuotteen joukkoja ei voi enää muodostaa

Yleiset joukot ja frekvenssit

Tuotejoukko / frekvenssi

Juusto 4

Maito 5

Hedelmät 5

Leipä 6

Maito, Leipä 4

Maito, Hedelmät 4

Leipä, Juusto 4

Leipä, Hedelmät 4

Maito, Leipä, hedelmät 3

Jäljelle jää yksi tuotejoukko, jossa on kolme tuotetta.

Jäljellä olevan joukon frekvenssi on kolme eli se on riittävän yleinen

Neljän tuotteen joukkoja ei voi enää muodostaa.

Yhdistetään tuotejoukot, joiden frekvenssi on kolme tai yli niin, että jokainen tuotejoukko otetaan vuorollaan vasemmaksi puoleksi.

Assosiaatiosääntöjen laskenta

- Maito, leipä
  - Maito -> leipä 4/5
  - Leipä -> maito 4/6
- Maito, hedelmät
  - Maito->hedelmät 4/5
  - Hedelmät->maito 4/5
- Leipä-juusto
  - Leipä->juusto 4/6
  - Juusto->leipä 4/4
- Leipä-hedelmät
  - Leipä>hedelmät 4/5
  - Hedelmät->leipä 4/5

- Maito, leipä, hedelmät
  - Maito, leipä -> hedelmät 3/4
  - Maito, hedelmät -> leipä 3/4
  - Leipä, hedelmät- > maito 3/4
  - Maito -> leipä, hedelmät- 3/4
  - ~~Leipä -> hedelmät-, maito 3/6~~
  - ~~hedelmät-, maito -> leipä 3/5~~

- Valitaan ne joiden luottamus on vähintään 0,75

Valmis sääntö

Assosiaatiosääntö	Luottamus	Tuki	Noste
• Maito -> leipä	4/5 (0,8)	4/7 (0,57)	0,93
• Maito -> Hedelmät	4/5	4/7	1,12
• Hedelmät -> Maito	4/5	4/7	1,12
• Juusto -> Leipä	4/4 (1,0)	4/7	1,67
• Hedelmät - > Leipä	4/5	4/7	0,93
• Maito, leipä -> Hedelmät	3/4 (0,75)	3/7 (0,43)	1,05
• Maito, Hedelmät -> Leipä	3/4	3/7	0,875
• Leipä, hedelmät- > maito	3/4	3/7	1,05

Apriori-mallin hyvät ja huonot puolet

Hyvää:

- suhteellisen yksinkertainen ja helppo ymmärtää
- tunnistaa tehokkaasti **usein esiintyvät assosiaatiosäännöt** suurissa tietojoukoissa
- karsii tehokkaasti harvinaisempi tuotejoukkoja, mikä vähentää algoritmin laskennallista monimutkaisuutta
- vakiintunut ja suhteellisen luotettava algoritmi

Huonot puolet:

- ei sovellu todella suurille aineistoille, koska algoritmin laskennallinen monimutkaisuus kasvaa eksponentiaalisesti tietojoukon koon myötä
- Ei analysoi harvinaisempia yhteyksiä
- vähimmäistuen ja vähimmäisluottamuksen raja vaikuttavat tulosten laatuun
- Voi olla altis tuottamaan suuren määrän assosiaatiosääntöjä, mikä voi vaikeuttaa tulosten tulkintaa

Association Rule Mining in Python Tutorial | DataCamp

Seuraavaksi python ja apriori

- Malli antaa aina jonkin vastauksen
- Säädä parametrejä oleellisen tuloksen löytämiseksi