school of social and behavioral sciences

STATISTICAL PROGRAMMING - PYTHON

Machine Learning: association rules

Pablo Monfort

Instituto de Empresa

Summary

- Machine Learning overview
- Unsupervised learning
    - clustering
        - KMeans
        - hierarchical clustering
    - dimensionality reduction
        - feature selection
        - feature extraction
            - PCA: principal component analysis
            - LDA: linear discriminant analysis
    - association rules
- Supervised learning
    - classification
    - regression

Machine learning: algorithms

school of social and behavioral sciences

Supervised learning:

- Classification: classification trees…
- Regression: linear regression, logistic regression, stepwise regression, regression trees…

Unsupervised learning:

- Clustering: k-means, k-medians, hierarchical clustering…
- Dimensionality reduction: principal component analysis, discriminant analysis…
- Association rule: a priori agorithm…

Machine Learning: use cases

Supervised learning:

- Classification: customer retention, fraud detection, image classification…
- Regression: market forecasting, population growth prediction…

Unsupervised learning:

- clustering: customer segmentation, recommender systems…
- dimensionality reduction: structure discovery, big data visualization…
- association rule: targetted marketing…

Summary

- Machine Learning overview
- Unsupervised learning
  - clustering
    - KMeans
    - hierarchical clustering
  - dimensionality reduction
    - feature selection
    - feature extraction
      - PCA: principal component analysis
      - LDA: linear discriminant analysis
  - association rules
- Supervised learning
  - classification
  - regression

Association rules: overview

Goal:
Find association relationships in large data sets.

How does it work?
When we look for association rules, some parameters have to be defined:

- **support** of an itemset: percentage of the data set which contains this itemset
- **confidence** of a rule A => B: P(A and B)/P(A). It means, the rule is true for at least that confidence value.
- **lift of a rule** (A => B): support(A U B)/(support(A) * support(B)). Lift > 1 means both occurrences are dependent on one other

Example: rules with support 0.40 and confidence 0.80
{peanuts, coke} => {chips} means that:

- at least, a 40% of the cases included peanuts and coke
- in the cases where peanuts and coke were sold, chips were sold as well at least in a 80% of these transactions.

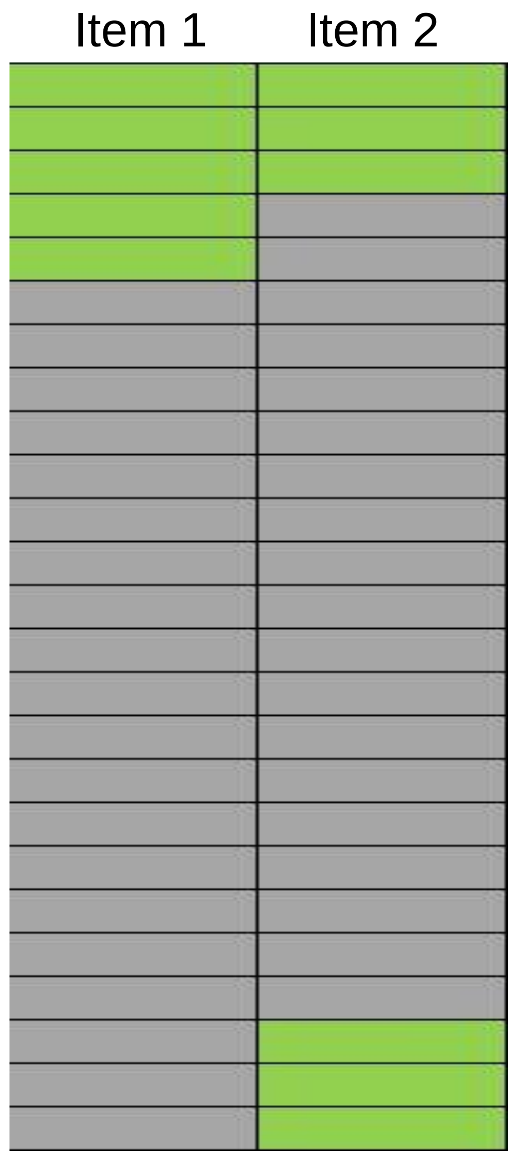
Usages:
Market analysis
Click stream tracking
Online recommendation engines

6

Association rules: overview

school of social and behavioral sciences

Support item 1 = 20% of the total invoices



Confidence rule = 60% of the total invoices (where item 1 was bought) include ítem 2

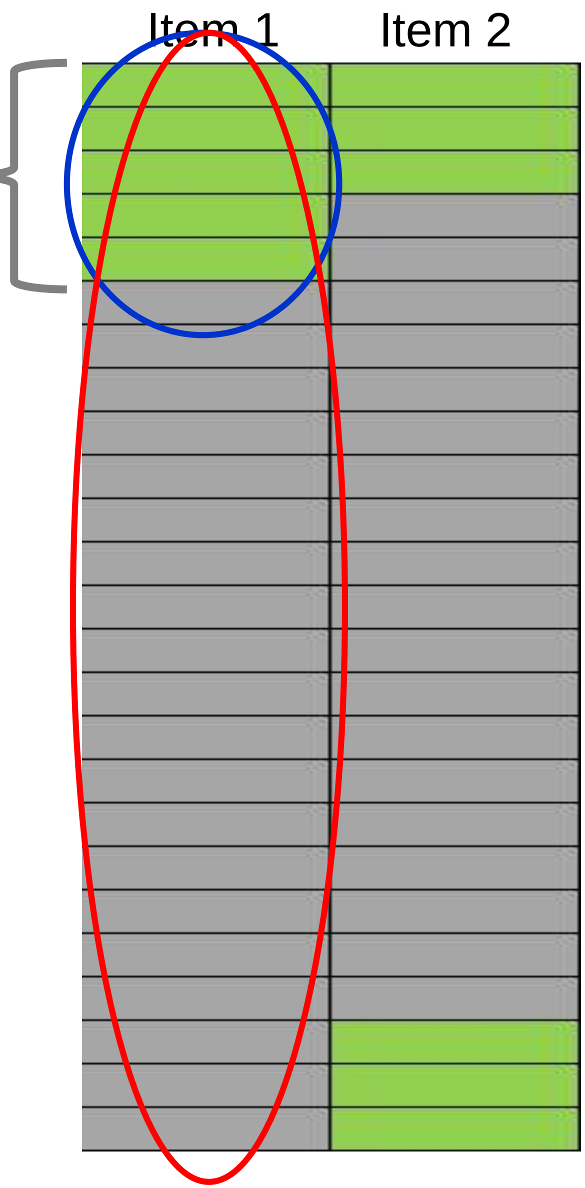
rule:
{item 1} => {item 2}

items
green = bought

7

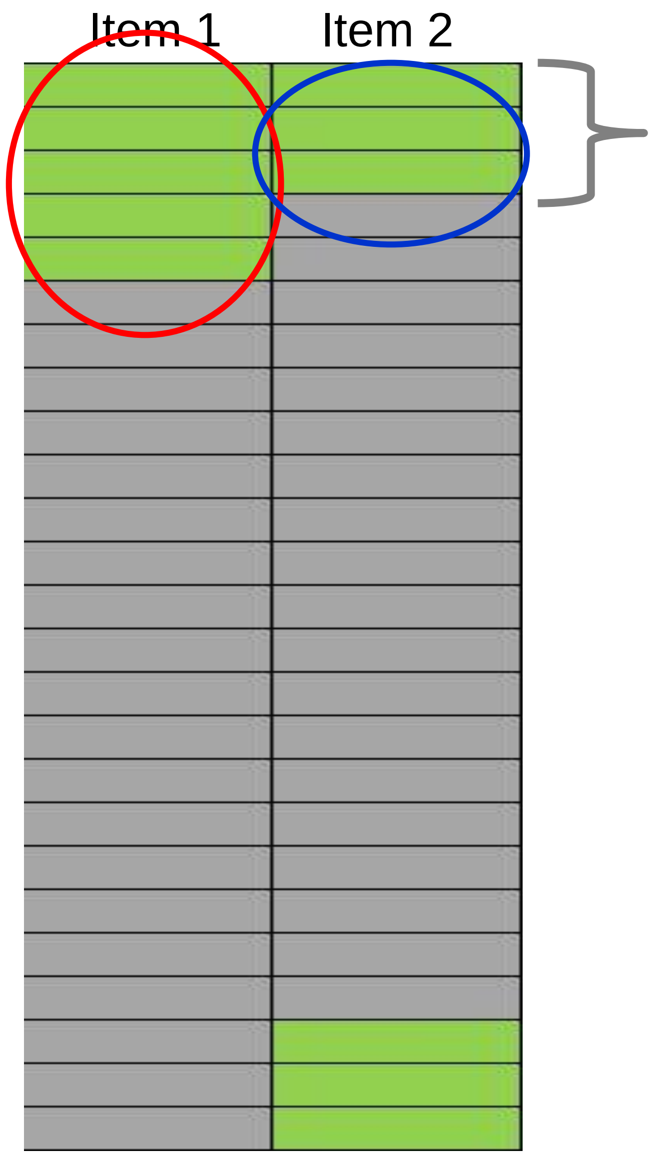
Association rules: overview

Support item 1 = 20% of the total invoices



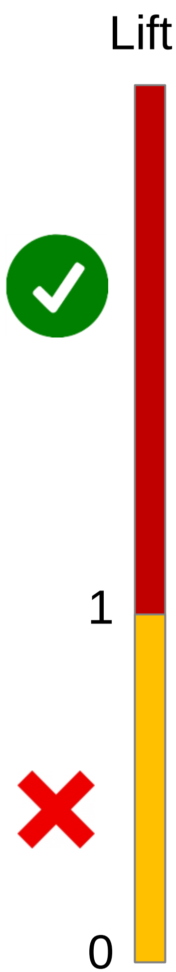
8

Association rules: overview



Confidence rule = 60% of the total invoices where item 1 was bought

Association rules: overview



high lift:
indicates it is not just a coincidence. They are dependent events

Lift 1:
independent events

Lift <1:
independent events

Association rules: data structure

Mode 1: Sparse vectors

1, 1, 19
2, 5, 35, 42, 44
3, 10, 17, 26, 27, 28, 31, 32
4, 2, 5, 18, 22, 27, 46
5, 23, 24, 40, 41, 43

Mode 3: Items tables

1,4,1
1,2,19
2,5,42
2,5,5
2,5,44
2,2,35
3,4,28
3,4,27
3,2,32
3,1,10
3,4,17
3,2,31
3,4,26
4,3,22
4,3,18
4,3,27
4,4,46
4,2,5
4,5,2
5,2,23
5,3,24
5,5,41
5,1,43
5,1,40

All of them are equivalent

Mode 2: Full binary vectors

1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
2,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,1,0,0,0,0,0
3,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,1,1,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
4,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0
5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,0,0,0

Association rules: data structure

school of social and behavioral sciences

Mode 1: Sparse vectors

1, 1, 19
2, 5, 35, 42, 44
3, 10, 17, 26, 27, 28, 31, 32
4, 2, 5, 18, 22, 27, 46
5, 23, 24, 40, 41, 43

Mode 3: Items tables

1,4,1
1,2,19
2,5,42
2,5,5
2,5,44
2,2,35
3,4,28
3,4,27
3,2,32
3,1,10
3,4,17
3,2,31
3,4,26
4,3,22
4,3,18
4,3,27
4,4,46
4,2,5
4,5,2
5,2,23
5,3,24
5,5,41
5,1,43
5,1,40

All of them are equivalent

Mode 2: Full binary vectors

1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
2,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,1,0,0,0,0,0,0
3,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,1,1,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
4,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0
5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,0,0,0

Challenge

- How many transactions are there?
- Which is the number of different items in transaction #4?
- Which is the number of sold items in transaction #2?

Association rules: data structure

Mode 1: Sparse vectors

1, 1, 19
2, 5, 35, 42, 44
3, 10, 17, 26, 27, 28, 31, 32
4, 2, 5, 18, 22, 27, 46
5, 23, 24, 40, 41, 43

items

receipt

Mode 3: Items tables

1,4,1
1,2,19
2,5,42
2,5,5
2,5,44
2,2,35
3,4,28
3,4,27
3,2,32
3,1,10
3,4,17
3,2,31
3,4,26
4,3,22
4,3,18
4,3,27
4,4,46
4,2,5
4,5,2
5,2,23
5,3,24
5,5,41
5,1,43
5,1,40

Mode 2: Full binary vectors

1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
2,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,1,0,0,0,0,0
3,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,1,1,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
4,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0
5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,0,0

First receipt: item 1 (4 units) and ítem 19 (2 ítems)
Second receipt: ítem 5 (5 units), ítem 35 (2 units), ítem 42 (5 units) and item 44 (5 items)
…

Association rules: data structure

Mode 1: Sparse vectors

1, 1, 19
2, 5, 35, 42, 44
3, 10, 17, 26, 27, 28, 31, 32
4, 2, 5, 18, 22, 27, 46
5, 23, 24, 40, 41, 43

Mode 3: Items tables

1,4,1
1,2,19
2,5,42
2,5,5
2,5,44
2,2,35
3,4,28
3,4,27
3,2,32
3,1,10
3,4,17
3,2,31
3,4,26
4,3,22
4,3,18
4,3,27
4,4,46
4,2,5
4,5,2
5,2,23
5,3,24
5,5,41
5,1,43
5,1,40

<u>Required structure</u> for the association analysis algorithm:
index: transaction Ids
columns: products with 1s or 0s

Mode 2: Full binary vectors

1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
2,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,1,0,0,0,0,0,0
3,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,1,1,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
4,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0
5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,0,0,0

First receipt: item 1 (4 units) and ítem 19 (2 ítems)
Second receipt: ítem 5 (5 units), ítem 35 (2 units), ítem 42 (5 units) and item 44 (5 items)
…

14

Selecting the support and confidence parameters

15

Association rules

school of social and behavioral sciences

from mlxtend.frequent_patterns import apriori, association_rules

frequent_itemsets = apriori(…)

Arguments in apriori command:

- df: dataframe in "Full binary vectors" format # columns: products. index: transactions IDs
- min_support: number
- use_colnames: True or False (default) # False: column indices. True: column names

rules = association_rules(frequent_itemsets, …)

Arguments in association_rules command:

- df: frequent_itemsets
- metric: "lift" or "confidence" (default)
- min_threshold: number (default: 0.8)

Association rules: exercise

school of social and behavioral sciences

Programming challenge ARUL.1

File: association_rules_20000_XXX.csv

1. Give the correct structure to the data file in Python

Note: check the command set_index

2. Find a set of some rules

3. Try to quantify for some rule the number of non-sold units that we could try to sell doing recommendations

Association rules: exercise

Programming challenge ARUL.2

File: "Online Retail_Association rules_v2.csv"

1. Show the 3 items which have produced the highest income

18

Association rules: exercise

Programming challenge ARUL.2

File: "Online Retail_Association rules_v2.csv"

1. Show the 3 items which have produced the highest income
2. Give a calendar heatmap for the number of sold items per day in United Kingdom during December 2010 and January 2011

19

Association rules: exercise

school of social and behavioral sciences

Programming challenge ARUL.2

File: "Online Retail_Association rules_v2.csv"

1. Show the 3 items which have produced the highest income
2. Give a calendar heatmap for the number of sold items per day in United Kingdom during December 2010 and January 2011
3. Create the required structure to apply a association rules analysis for the invoices in France.

   After this, calculate all the rules with lift > 1 for products sold in a 7% of the transactions at least.
4. For all the previous rules, extract the rules with lift > 6 and confidence is, at least, 80%.

Association rules: exercise

Programming challenge ARUL.2

File: "Online Retail_Association rules_v2.csv"

1. Show the 3 items which have produced the highest income
2. Give a calendar heatmap for the number of sold items per day in United Kingdom during December 2010 and January 2011
3. Create the required structure to apply a association rules analysis for the invoices in France.

   After this, calculate all the rules with lift > 1 for products sold in a 7% of the transactions at least.
4. For all the previous rules, extract the rules with lift > 6 and confidence is, at least, 80%.
5. For the rule "ALARM CLOCK BAKELIKE GREEN" => "ALARM CLOCK BAKELIKE RED", how much extra money can I win if all the clients are moved to the second part as well?

21

Session Wrap-up

Glossary

apriori()
association_rules()

23