Association rule learning

Paweł Teisseyre

- Introduction.
- Market Basket Analysis.
- Generalized Association rules.







Introduction

- example of unsupervised learning,
- the main goal is to find joint values of variable $x = (x_1, \dots, x_p)$ that appear most frequently in the data base.
- General task: find s_1, \ldots, s_p such that:

$$P\left[\bigcap_{j=1}^{p}(x_{j}\in s_{j})\right] \tag{1}$$

is large.

In other words: find regions of x where probability is high.









Introduction

- General approaches to solving (1) are not feasible in commercial applications ($p \approx 10^4$, $n \approx 10^8$).
- Some simplifications lead to MBA (all variables are binary).
- One can apply the technique of dummy variables to turn (1) into a problem involving only binary-valued variables.







• Transactions:

| Transaction Id | Milk | Bread | Butter | Beer |
|----------------|------|-------|--------|------|
| 1 | 1 | 1 | 0 | 0 |
| 2 | 0 | 0 | 1 | 0 |
| 3 | 0 | 0 | 0 | 1 |
| 4 | 1 | 1 | 1 | 0 |
| 4 | 1 | 1 | 1 | 0 |
| 5 | U | 1 | U | U |







Notation:

- *I* set of all products.
- $X, Y \in \mathcal{I}$ -itemsets.
- $\operatorname{supp}(X)$ -% of transactions containing X. Example: $\operatorname{supp}(\{Milk, Bread\}) = \frac{2}{5}$.

Goal of MBA:

• Find all itemsets X such that $supp(X) \ge t$, where t is user-specified threshold.







Apriori algorithm:

- Find itemsets X containing only one product such that $supp(X) \ge t$.
- ② Among itemsets found in step 1 find itemsets X with two elements such that $\operatorname{supp}(X) \geq t$.
- **3** Among itemsets found in step 2 find itemsets X with three elements such that $supp(X) \ge t$.
- Continue until all candidate itemsets from the previous pass have support less than the specified threshold.







Rules:

- $X, Y \in \mathcal{I}$. We write $X \implies Y$ to denote $X \cup Y$.
- We define $supp(X \implies Y) := supp(X \cup Y)$.
- Example: $\{Butter, Bread\} \implies \{Milk\}$. We have:

$$supp(\{Butter, Bread\} \implies \{Milk\}) = \frac{1}{5}$$







Useful quantities:

• Confidence (predict ability):

$$\operatorname{conf}(X \implies Y) := \frac{\operatorname{supp}(X \cup Y)}{\operatorname{supp}(X)}$$

- Confidence is an estimate of P(Y|X).
- Example:

$$conf(\{Butter, Bread\} \implies \{Milk\}) = \frac{1}{1} = 1.$$









Useful quantities:

Lift (association measure):

$$\operatorname{lift}(X \implies Y) := \frac{\operatorname{supp}(X \cup Y)}{\operatorname{supp}(X) \cdot \operatorname{supp}(Y)}$$

- Lift is an estimate of $\frac{P(X,Y)}{P(X)P(Y)}$.
- Example:

lift
$$(\{Butter, Bread\} \implies \{Milk\}) = \frac{1/5}{1/5 \cdot 2/5} = 2.5.$$







Problem:

- Rules with high confidence or lift, but low support, will not be discovered.
- For example, a high confidence rule such as vodka

 caviar

 will not be uncovered owing to the low sales volume of the
 consequent caviar.







Unsupervised as supervised learning

• g- unknown density function. Our goal : estimate g based on observed independent data points $x_1, \ldots, x_n \sim g$.







Unsupervised as supervised learning

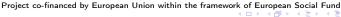
Procedure:

- g_0 reference, known distribution, e.g. uniform over a range of x.
- ② Generate artifficial sample from g_0 (N_0 points).
- **3** Assign weights $N_0/(N+N_0)$ to artifficial sample points and $N/(N+N_0)$ to original points.
- **4** Assign Y = 1 to original data points and Y = 0 to artifficial ones.
- Estimate $\mu = E(Y|x) = \frac{g(x)}{g(x) + g_0(x)}$ using supervised methods (e.g. logistic regression or decission tree).
- $\bullet \quad \text{Estimate } \hat{g}(x) := g_0(x) \frac{\hat{\mu}(x)}{1 \hat{\mu}(x)}.$









Paweł Teissevre



Unsupervised as supervised learning

- Choice of g₀ is important.
- Sometimes g_0 is chosen to represent departures of g from g_0 . For example:
 - if departures from uniformity are of interest, $g_0(x)$ might be the uniform density over the range of the variables.
 - if departures from joint normality are of interest, a good choice for $g_0(x)$ would be a Gaussian distribution with the same mean vector and covariance matrix as the data.
 - departures from independence could be investigated by using

$$g_0(x) = \prod_{j=1}^p g_j(x_j),$$

where $g_j(x_j)$ is the marginal data density of x_j , the jth coordinate of x.







Generalized Association Rules

- The goal is to find regions of high probability.
- We can use the idea of 'unsupervised as supervised learning'.
- As reference distribution we take

$$g_0(x) = \prod_{j=1}^p g_j(x_j),$$

where $g_j(x_j)$ is the marginal data density of x_j , the jth coordinate of x. A sample from this independent density is easily generated from the data itself by applying a different random permutation to the data values of each of the variables.







Generalized Association Rules

- Assign Y = 1 to original data points and Y = 0 to artifficial ones.
- Find regions

$$R=\bigcap_{j}(x_{j}\in s_{j})$$

for which P(Y = 1|x) is large with additional requirement that support is not too small.

 The regions are defined by conjunctive rules. Hence supervised methods that learn such rules would be most appropriate in this context (e.g. ???)







