

Data Mining Techniques

Chapter 9: Market Basket Analysis and Association Rules

Market basket analysis	2
Market basket data I	3
Market basket data II	4
Association rules	5
How good is an association rule?	6
Building association rules	7
Pizza example	8
Calculations	9
Refinements	10
Extensions	11

Market basket analysis

- Undirected data mining technique (no “target” or “response” variable).
- Kind of problem: what merchandise are customers buying and when?
- Which products tend to be purchased together and which are amenable to promotion?
 - suggest new store layouts;
 - determine which products to put on special;
 - indicate when to issue coupons;
 - tie data to individual customers through a loyalty card or website registration.
- Other potential applications—see p. 288.
- Pitfalls:
 - rules can end up merely describing previous marketing promotions;
 - product differentiation may not be apparent until long history.

© Iain Pardoe, 2006

2 / 11

Market basket data I

- Point-of-sale transaction data.
- Three levels of market basket data:
 - customers;
 - orders (purchases, baskets, item sets);
 - items.
- Track customers over time:
 - average # orders per customer;
 - average # (unique) items per order;
 - proportion of customers purchasing a particular product;
 - average # orders per customer including a particular product;
 - average quantity ordered per customer including a particular product.

© Iain Pardoe, 2006

3 / 11

Market basket data II

- Order characteristics.
- Item popularity:
 - in a one-item order;
 - in a multi-item order;
 - amongst customers who are repeat purchasers;
 - over time;
 - over regions.
- Tracking marketing interventions.
- Clustering products by usage—which product(s) in a purchase suggest the purchase of other particular products at the same time:
 - association rules (handfuls of items);
 - cluster analysis (larger sets).

© Iain Pardoe, 2006

4 / 11

Association rules

- If a customer buys item A, we expect he/she will also buy item B.
- **Actionable**: useful rules with understandable, high-quality info, e.g., if Barbie then chocolate;
 - might suggest more prominent product placement (e.g., beer and diapers—see p. 298), product tie-ins and promotions, or particular ways to advertise products.
- **Trivial**: already known by managers, e.g., if maintenance contract then large appliance;
 - may reflect past marketing or product bundling;
 - exceptions may signal failures in business operations, data collection, or processing.
- **Inexplicable**: seem to have no explanation and do not suggest a course of action, e.g., if new hardware store then toilet bowl cleaners;
 - may reflect over-fitting.

© Iain Pardoe, 2006

5 / 11

How good is an association rule?

- If I_1 (condition/antecedent) then I_2 (result/consequent) e.g., if OJ then milk (p. 299, beware mistakes).
- Support = proportion of transactions with I_1 & I_2 , e.g., $1/5 = 20.0\%$.
- Confidence = $(\# \text{ transactions with } I_1 \& I_2) / (\# \text{ transactions with } I_1)$, e.g., $1/4 = 25.0\%$.
- Lift = confidence / proportion of transactions with I_2 , e.g., $(1/4) / (1/5) = 1.25$:
 - or, lift = $(\text{actual } \# \text{ transactions with } I_1 \& I_2) / (\text{expected } \# \text{ trans with } I_1 \& I_2 \text{ if no relationship}) = \#(I_1 \& I_2) / (\#I_1 \times \#I_2 / N) = 1 / (4 \times 1 / 5)$;
 - excess = $\#(I_1 \& I_2) - (\#I_1 \times \#I_2 / N) = 1 - (4 \times 1 / 5) = 0.20$.

© Iain Pardoe, 2006

6 / 11

Building association rules

- Determine the item set:
 - select right level of detail (start broad, repeat with finer detail to hone in);
 - product hierarchies help to generalize items, e.g., frozen food → dessert, vegetable, dinner;
 - hybrid approach depending on price or frequency (analysis easier when roughly same number of transactions for each item);
 - “virtual” items go beyond product hierarchy, e.g., designer labels, low-fat products, energy-saving options, payment method, day, demographic info, etc.
- Calculate counts/probabilities of items and combinations of items.
- Analyze support/confidence/lift to find actionable rules.

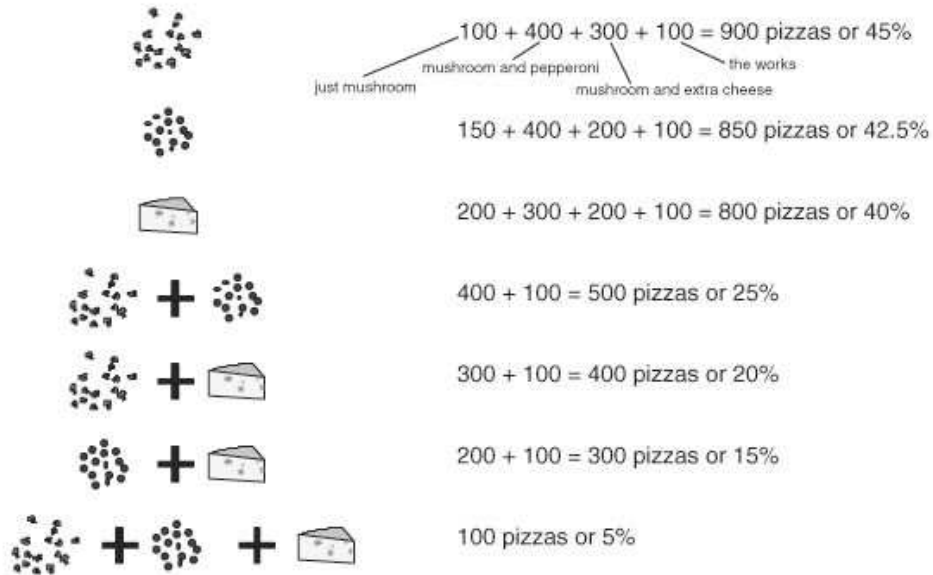
© Iain Pardoe, 2006

7 / 11

Pizza example

A pizza restaurant has sold 2000 pizzas, of which:
 100 are mushroom only, 150 are pepperoni, 200 are extra cheese
 400 are mushroom and pepperoni, 300 are mushroom and extra cheese, 200 are pepperoni and extra cheese
 100 are mushroom, pepperoni, and extra cheese.
 550 have no extra toppings.

We need to calculate the probabilities for all possible combinations of items.



© Iain Pardoe, 2006

8 / 11

Calculations

Rule	$\Pr(I_1 \& I_2)$ Support	$\Pr(I_1)$	$\frac{\Pr(I_1 \& I_2)}{\Pr(I_1)}$ Confidence	$\Pr(I_2)$	$\frac{\Pr(I_1 \& I_2)}{\Pr(I_1)\Pr(I_2)}$ Lift
If M then P	0.25	0.450	0.556	0.425	1.31
If P then M	0.25	0.425	0.588	0.450	1.31
If M then C	0.20	0.450	0.444	0.400	1.11
If C then M	0.20	0.400	0.500	0.450	1.11
If P then C	0.15	0.425	0.353	0.400	0.88
If C then P	0.15	0.400	0.375	0.425	0.88
If (M,P) then C	0.05	0.250	0.200	0.400	0.50
If (M,C) then P	0.05	0.200	0.250	0.425	0.59
If (P,C) then M	0.05	0.150	0.333	0.450	0.74

- Best rule: if Pepperoni then Mushroom.
- Support 25% and confidence 58.8%.
- Lift 1.31 means when Pepperoni requested then Mushroom is 31% more likely to be requested than if Pepperoni & Mushroom were unrelated.

© Iain Pardoe, 2006

9 / 11

Refinements

- Negative rules: when lift < 1 then negating the result produces a better rule, e.g.:

Rule	$\Pr(I_1 \& I_2)$ Support	$\Pr(I_1)$	$\frac{\Pr(I_1 \& I_2)}{\Pr(I_1)}$ Conf	$\Pr(I_2)$	$\frac{\Pr(I_1 \& I_2)}{\Pr(I_1)\Pr(I_2)}$ Lift
If (P,C) then not M	0.10	0.150	0.667	0.550	1.21
If (M,P) then not C	0.20	0.250	0.800	0.600	1.33

- Overcoming practical limits: find rules with 2 items, 3 items, etc.
- Pruning: reduce # of items and combinations of items considered at each step (e.g., minimum support threshold).
- Problems of large datasets: very computer intensive.

© Iain Pardoe, 2006

10 / 11

Extensions

- Using association rules to compare stores, promotions at various times, geographic areas, urban vs. suburban, seasons, etc. (use virtual items).
- Dissociation rules, e.g., if I_1 and not I_2 then I_3 (use sparingly for only the most frequent items).
- Sequential analysis using association rules: requires identifying customers and tracking them over time.

© Iain Pardoe, 2006

11 / 11