NAVAL POSTGRADUATE SCHOOL

OS4118

Statistical and Machine Learning

Association Rules

Prof. Sam Buttrey

Fall FY 2020

Association Rules

- Another unsupervised technique
- Given categorical data, construct **rules**
- A rule usually has the form

If variable A has value a **and** variable B has value b, then variable D will have value d

- Here we have two **antecedents** (conditions) and one **consequent**
- Usually rules have antecedents joined by “and” and exactly one consequent

Rules

- Rules generally concern categorical variables (sets)
- Continuous variables must be discretized
- Binary variables should often be treated asymmetrically
- No response variable: any variable can be “predictor” or “response”
- Example: market basket
  - E.g. “If ‘hot dog buns’ are in the basket, then ‘hot dogs’ are in the basket

Market Basket Examples

- Items are items, baskets are baskets, look for shoppers’ similarities
- Items are words, baskets are documents, look for linked concepts
- Items are sentences, baskets are documents, look for plagiarism
- Items are genes, baskets are people…

Scale

- Humans have about 30,000 genes and there are six-seven billion of us
- Wal-Mart sells about 100,000 items and can store millions of baskets
- The Web has 100,000,000 words and billions of pages
- Data off-line, need few passes

Support and Confidence

- Object: find rules with high support (frequency, coverage) and high confidence (accuracy)
- Support: proportion of sample meeting the conditions of the antecedents
- Confidence: proportion of supported sample meeting the consequent

Example

- I go to work by tunnel, by Route 68, or through the Presidio of Monterey
- Each rule carries a probability
- "If I take Rte 68, I'll be late with probability 90%" compared to "Today I will be late with probability 20%"
- The difference between the 90% and the 20% **confidences** is one measure of the rule's usefulness – unless I never take Rte 68 (which is measured by **support**)

Example

Route	Rte 68	Tunnel	Presidio
Not Late	4%	20%	24%
Late	21%	21%	10%
Marginal	**25%**	**41%**	**34%**

The rule "If Rte 68, then Late" has **25%** support. The confidence is the usual estimate of the conditional probability

Pr (L | 68) = Pr (L & 68) / Pr (68) = .21/.25
= **84%**

Rules vs. Classification Trees

- In a tree, there is one response variable
- The “rules” are mutually exclusive and exhaustive (no overlap)
- Association rules use any variable as a response
- Many observations are covered by more than one rule (much overlap)

Finding Rules

- The hard part is finding “frequent itemset patterns,” sets of conditions whose support exceeds a threshold *s*
- In one pass we can find frequent items defined by one condition
- “A and B” is frequent only if A and B are both frequent: Pr (A & B) < Pr (A)
- In general a set is frequent only if all its subsets are themselves frequent

Finding Frequent Sets

- On pass two we can examine all pairs of frequent sets to compute their frequency
- On pass three, examine all triplets
  - The number of “sets of sets” grows quickly, so *s* needs to be chosen wisely
  - Number of conditions can be limited
- One more pass generates rules
  - For a set *A*, enumerate all possible rules like “if *A*, then *B*”, evaluate accuracy
- Lots of rules, lots of overlap

More on Rules

- Evaluating "If *A* then *B*" :

1. Confidence: Pr (B | A)
2. Conf. difference: Pr ($B|A$) – Pr (B)
3. Lift: Pr ($B|A$) / Pr (B)
   – Or Pr(A&B)/Pr(A)Pr(B) in our package
4. Conf. ratio: [Pr ($B|A$) / Pr (B)] – 1
5. Information Difference: Gain_A – Gain
6. Normalized χ^2 (Cramer's coefficient)

These seem like the two most commonly used

"A Priori" Notes

- R implementation in `library (arules)`
- Tabular vs. transactional data
  - Transactional data: only the items present in the basket are passed
- Items that are rarer than the minimum support level can never appear in a rule
  - Some algorithms let you specify a different minimum support for each item
  - We need to keep very common items out of rules for rare items
  - Note: our "support" is Pr(A), but `apriori()` uses Pr (A and B)

Fake-ier Example

- AdultUCI/Adult data set from `arules`
- Goal: Characterize relationships among descriptors of survey responders
- Code mostly taken from the help pages
- Problem of “very common” items dominating rule sets

Real-ier Example

- Database of tunnels under the U.S. border
- Goal 1: characterize tunnels by entrance, length, sophistication, etc.
- Goal 2: identify locations where tunnels are more likely to appear
- Let’s do this thing!