Introduction to Machine Learning

Lecture 13
Unsupervised Learning III
Association Rule Mining

Goker Erdogan
26 – 30 November 2018
Pontificia Universidad Javeriana

Association rule mining

- Popular data mining technique
  - Rule-based machine learning

- Given a dataset X
  - Find joint values of features that appear frequently

- Often applied to binary data
  - Market basket analysis

Example database with 5 transactions and 5 items

transaction ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

[2]

Density estimation perspective

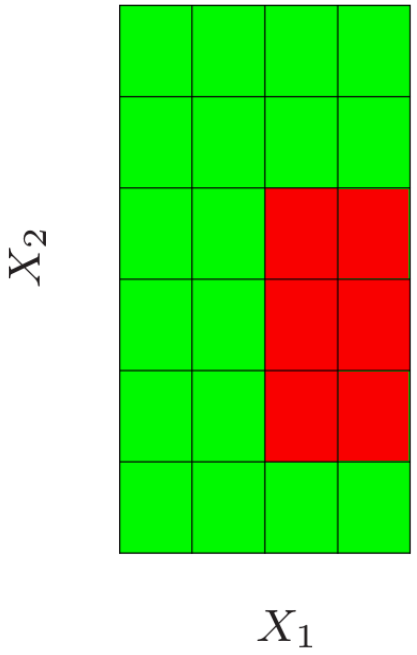
- More generally,
  - Find x such that P(x) is large

- If x is high dimensional
  - Curse of dimensionality
  - Little data to estimate P(x)

- Find regions of high probability
  - Region R with large P(x in R)

$$\Pr\left[\bigcap_{j=1}^{p}(X_j \in s_j)\right],$$

  - s_j: subset of all possible values for X_j


X_2
X_1

Market basket analysis

- Constrain the form of the rule (region)
  - s_j is a single value

$$\Pr\left[\bigcap_{j\in\mathcal{J}}(X_j = v_{0j})\right]$$

- Use dummy variables (one-hot) to turn all features to binary

$$\Pr\left[\bigcap_{k\in\mathcal{K}}(Z_k = 1)\right] = \Pr\left[\prod_{k\in\mathcal{K}} Z_k = 1\right]$$

  - *K*: item-set
- Find item-sets with high probability

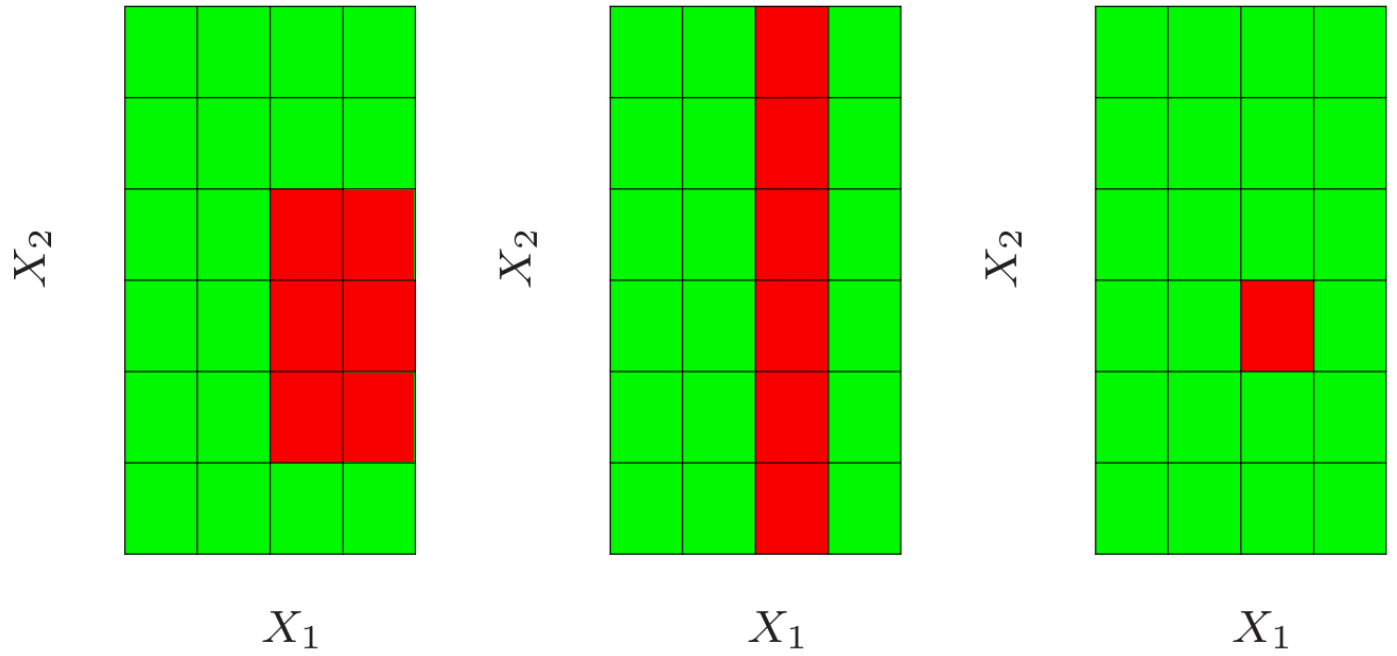


FIGURE 14.1. *Simplifications for association rules. Here there are two inputs X_1 and X_2, taking four and six distinct values, respectively. The red squares indicate areas of high density. To simplify the computations, we assume that the derived subset corresponds to either a single value of an input or all values. With this assumption we could find either the middle or right pattern, but not the left one.*

Market basket analysis

- T(*K*) = Support of item-set *K*: Probability of an item set *K*

$$\widehat{\Pr}\left[\prod_{k\in\mathcal{K}}(Z_k=1)\right]=\frac{1}{N}\sum_{i=1}^{N}\prod_{k\in\mathcal{K}}z_{ik}.$$

- Find all rules with support > t
  - Not feasible to look at all possible subsets of K

- Two observations
  - If t is large, there are only a few item-sets with support > t
  - If $\mathcal{L}\subseteq\mathcal{K}\Rightarrow T(\mathcal{L})\geq T(\mathcal{K}).$

A-priori algorithm

A-priori algorithm

- For k=1,2,…,|*K*|
  - Calculate support of all k item-sets: K_1, K_2, ...
  - Discard all item-sets K_i with $T(K_i) < t$

Example database with 5 transactions and 5 items

transaction ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

Association rules

- For an item-set, split it into two sets A and B
  - A → B is an association rule
  - A: antecedent, B: consequent

Analogous to

- Support of A → B $T(A \Rightarrow B) = T(A \cup B)$ P(A,B)
- Confidence of A → B $C(A \Rightarrow B) = \frac{T(A \Rightarrow B)}{T(A)},$ P(B|A)
- Lift of A → B $L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{T(B)}.$ P(A,B)/P(A)P(B)

What does lift mean?

- $L(A \rightarrow B) = 1$
  - A and B are independent
  - An increase/decrease in A has little effect on B
- $L(A \rightarrow B) > 1$
  - A has a positive effect on B
  - Potentially useful for predicting B from A
- $L(A \rightarrow B) < 1$
  - A has a negative effect on B
  - A and B are substitutes

- Ideally a good rule has high support, confidence, and lift.

Association rule mining

- We want rules with high confidence

$$\{A \Rightarrow B \,|\, C(A \Rightarrow B) > c\}$$

- Given the item-sets with high support
  - Rules A → B can be found with a variant of Apriori algorithm

- Then we have a collection of association rules

$$T(A \Rightarrow B) > t \quad \text{and} \quad C(A \Rightarrow B) > c.$$

TABLE 14.1. *Inputs for the demographic data.*

Feature	Demographic	# Values	Type
1	Sex	2	Categorical
2	Marital status	5	Categorical
3	Age	7	Ordinal
4	Education	6	Ordinal
5	Occupation	9	Categorical
6	Income	9	Ordinal
7	Years in Bay Area	5	Ordinal
8	Dual incomes	3	Categorical
9	Number in household	9	Ordinal
10	Number of children	9	Ordinal
11	Householder status	3	Categorical
12	Type of home	5	Categorical
13	Ethnic classification	8	Categorical
14	Language in home	3	Categorical

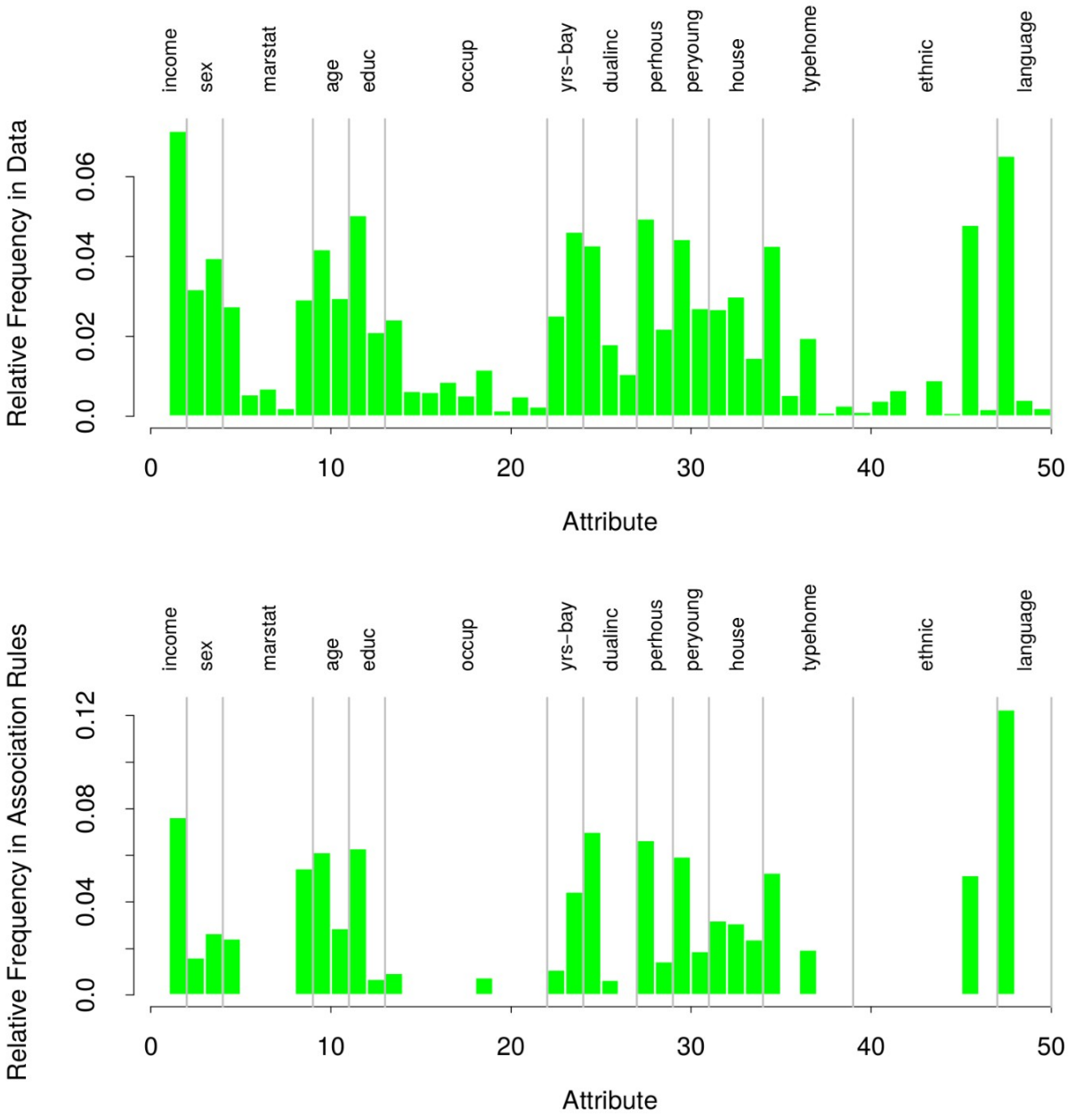


FIGURE 14.2. *Market basket analysis: relative frequency of each dummy variable (coding an input category) in the data (top), and the association rules found by the Apriori algorithm (bottom).*

Association rule 1: Support 25%, confidence 99.7% and lift 1.03.

$$\left[\begin{array}{rcl} \text{number in household} & = & \textit{1} \\ \text{number of children} & = & \textit{0} \end{array}\right] \Downarrow \text{language in home} = \textit{English}$$

Association rule 2: Support 13.4%, confidence 80.8%, and lift 2.13.

$$\left[\begin{array}{rcl} \text{language in home} & = & \textit{English} \\ \text{householder status} & = & \textit{own} \\ \text{occupation} & = & \{\textit{professional/managerial}\} \end{array}\right] \Downarrow \text{income} \geq \textit{\$40,000}$$

Association rule 3: Support 26.5%, confidence 82.8% and lift 2.15.

$$\left[\begin{array}{rcl} \text{language in home} & = & \textit{English} \\ \text{income} & < & \textit{\$40,000} \\ \text{marital status} & = & \textit{not married} \\ \text{number of children} & = & \textit{0} \end{array}\right] \Downarrow$$

$$\text{education} \notin \{\textit{college graduate, graduate study}\}$$

Market basket analysis

- Quite popular technique

- Limitations
  - Considers simple rules
    - Only a single value for each feature
  - Considers only rules with support > t
    - Misses high confidence, low-support rules
    - e.g., `vodka → caviar`
    - T(`caviar`) is small

- Generalized association rules

Generalized association rules

- Convert unsupervised density estimation problem to a supervised problem
  - Given a dataset X
    - Pick a reference distribution $g_0(x)$ [e.g., uniform]
      - Randomly draw a set of samples X_0 from $g_0(x)$
    - Create a classification problem
      - For all x_n in X, y = 1
      - For all x_n in X_0, y = 0
  - Fit a classifier to {X U X_0, Y}
    - Predict p(y=1|x)
    - This estimates the density g(x) of data

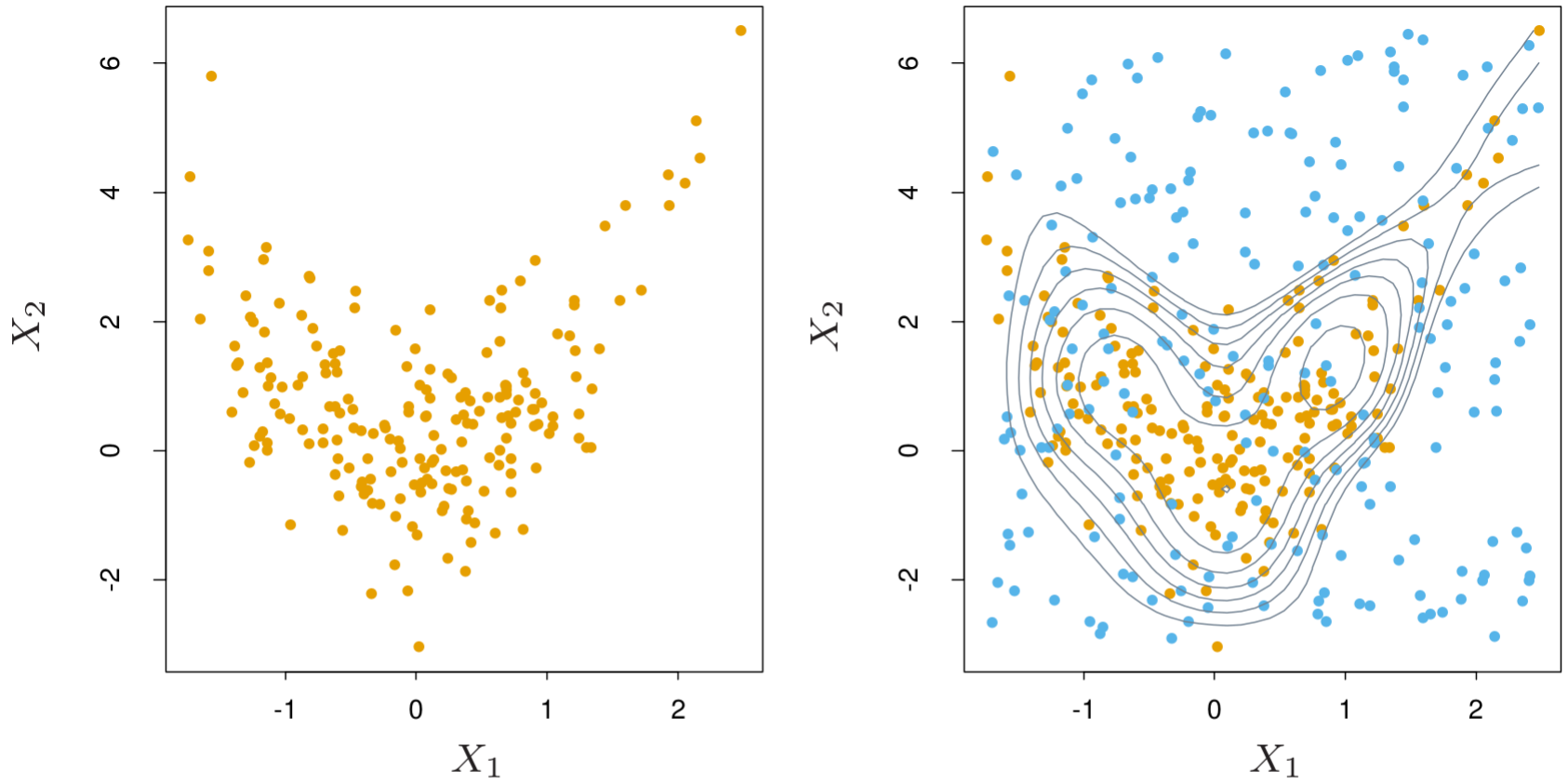


FIGURE 14.3. *Density estimation via classification. (Left panel:) Training set of* 200 *data points. (Right panel:) Training set plus* 200 *reference data points, generated uniformly over the rectangle containing the training data. The training sample was labeled as class* 1*, and the reference sample class* 0*, and a semiparametric logistic regression model was fit to the data. Some contours for* $\hat{g}(x)$ *are shown.*

Why does this work?

- Classifier learns to predict

$$\begin{aligned} \mu(x) = E(Y \mid x) &= \frac{g(x)}{g(x) + g_0(x)} \\ &= \frac{g(x)/g_0(x)}{1 + g(x)/g_0(x)} \end{aligned}$$

- Using the prediction from classifier, we can get

$$\hat{g}(x) = g_0(x) \frac{\hat{\mu}(x)}{1 - \hat{\mu}(x)}.$$

- An estimate of the true density

How to apply to learning association rules?

- Use a decision tree to classify {X U X_0}
  - Each leaf is an association rule of the general form

$$\widehat{\Pr}\left(\bigcap_{j\in\mathcal{J}}(X_j \in s_j)\right)$$

  - Called a generalized item-set

- We want rules with high support (i.e., probability)
  - Find terminal nodes t with high average y values

$$\bar{y}_t = \text{ave}(y_i \mid x_i \in t)$$

  - Support of rule R

$$T(R) = \bar{y}_t \cdot \frac{N_t}{N + N_0},$$

Association rule 1: Support 25%, confidence 99.7% and lift 1.35.

$$\left[\begin{array}{rcl} \text{marital status} & = & \textit{married} \\ \text{householder status} & = & \textit{own} \end{array}\right] \Downarrow \text{type of home} \neq \textit{apartment}$$

Association rule 2: Support 25%, confidence 98.7% and lift 1.97.

$$\left[\begin{array}{rcl} \text{age} & \leq & 24 \\ \text{occupation} & \notin & \{\textit{professional, homemaker, retired}\} \\ \text{householder status} & \in & \{\textit{rent, live with family}\} \end{array}\right]$$

$$\Downarrow$$

$$\text{marital status} \in \{\textit{single, living together-not married}\}$$

Association rule 3: Support 25%, confidence 95.9% and lift 2.61.

$$\left[\begin{array}{rcl} \text{householder status} & = & \textit{own} \\ \text{type of home} & \neq & \textit{apartment} \end{array}\right] \Downarrow \text{marital status} = \textit{married}$$

Association rule 4: Support 15%, confidence 95.4% and lift 1.50.

$$\left[\begin{array}{rcl} \text{householder status} & = & \textit{rent} \\ \text{type of home} & \neq & \textit{house} \\ \text{number in household} & \leq & 2 \\ \text{occupation} & \notin & \{\textit{homemaker}, \textit{student}, \textit{unemployed}\} \\ \text{income} & \in & [\$20{,}000, \$150{,}000] \end{array}\right]$$

$$\Downarrow$$

$$\text{number of children} = 0$$

Summary

- Association rule mining
- Market basket analysis
  - Apriori algorithm
  - Support, confidence, lift
  - Limitations
- Generalized association rules
  - Sample application to demographics data

- No exercises!

References

[1] Hastie, Tibshirani, and Friedman. The Elements of Statistical Learning. Chapter 14.

[2] https://en.wikipedia.org/wiki/Association_rule_learning