T-122.102 ANALYSIS OF BINARY DATA

Beyond Independence: Probabilistic Models for Query Approximation on Binary Transaction Data

Pavlov, Mannila, Smyth

presented by Tapani Raiko

Mar 11, 2003

Example Problem

- Web log data
- $D = 300$ different pages on the site
- $N = 100000$ site visitors
- A binary $(N \times D)$ matrix $\mathbf{X}$
- X_{nd} tells whether visitor n loaded page d or not
- "How many visitors loaded pages A and B but not C?"
- Assume that going through the whole data set online is unreasonably slow

Query Approximation

- Queries assumed conjunctive (extension to arbitrary Boolean expressions straigthforward)
- Important for:
  - optimization of database management systems
  - interactive data mining
  - prediction (!?)
- Tradeoff between:
  - accuracy
  - online time
  - offline time
  - memory

Methods

- Whole data scan
- Random data scan
- Independence model
- Chow-Liu tree model
- Mixtures of independence models
- Inclusion-exclusion model
- Maximum entropy model

Data Scan

- Whole data scan
  - Given a query, simply go through the data to find the answer
  - Accuracy is perfect
  - Online time complexity is $O(NQ)$ where N is the number of data samples and Q is the size of the query
  - No free parameters
- Random data scan
  - Use a subset of the data as above (the number of samples is a free parameter)
  - Easy to use as an anytime algorithm

Independence Model

- Assume D attributes independent of each other:
  $P(\mathbf{x}) = \prod_{d=1}^{D} P(x_d)$
- Collect statistics $\theta_d = \sum_{n=1}^{N} X_{nd}/N$ offline
- No free parameters
- Online time complexity is $O(Q)$ where Q is the size of the query

Independence Example

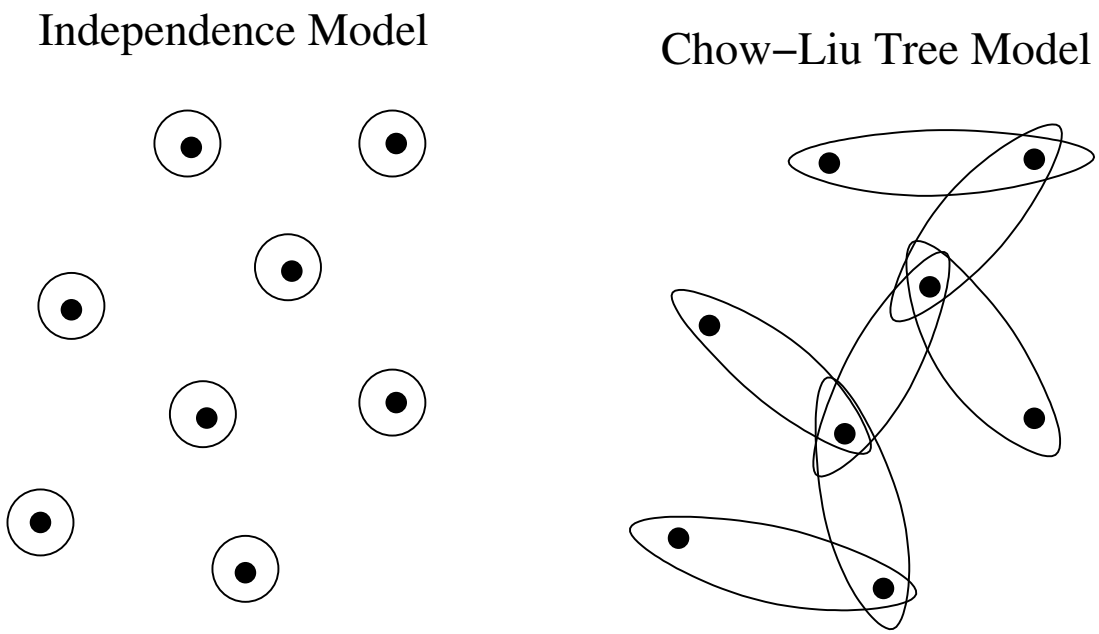
1 0 0
1 0 0
0 1 0
0 1 0
0 1 1
0 0 1

- Data dimensions: $D = 3$, $N = 6$
- Probabilistic model: $P(x_1 = 1) = 1/3$, $P(x_2 = 1) = 1/2$, $P(x_3 = 1) = 1/3$
- Query: "How often $x_1 = 0$ and $x_2 = 1$?"
- Answer: $(1 - P(x_1 = 1))P(x_2 = 1) = 1/3$
- Correct answer: $P(x_1 = 0, x_2 = 1) = 1/2$

Chow-Liu Tree Model

- Assume only pairwise dependencies between the D attributes
- Collect statistics θ_d as before and additionally:
  $\theta_{ij} = \sum_{n=1}^{N} X_{ni} X_{nj} / N$
- Compute mutual information between the attributes
- Find the minimum spanning tree (Kruskal's algorithm)
- Transform it to a simple Bayesian network where each node has just one parent
- Answer queries by doing standard belief propagation
- No free parameters
- Online time complexity is $O(QD)$

Independence and Chow-Liu Tree Models

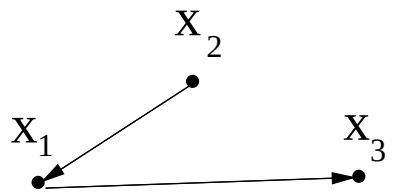


- $D = 8$ dots represent attributes and ellipses represent explicit distributions

Chow-Liu Example

1 0 0
1 0 0
0 1 0
0 1 0
0 1 1
0 0 1

- Mutual information:
  $I(X;Y) = H(X) + H(Y) - H(X,Y)$
- Entropy: $H(X) = -\sum P(X) \log P(X)$
- x_2 and x_3 do not contain any mutual information
- Bayesian network:



- $P(x_2 = 1) = 1/2$
- $P(x_1 = 1 \mid x_2 = 0) = 2/3$, $P(x_1 = 1 \mid x_2 = 1) = 0$
- $P(x_3 = 1 \mid x_1 = 0) = 1/2$, $P(x_3 = 1 \mid x_1 = 1) = 0$

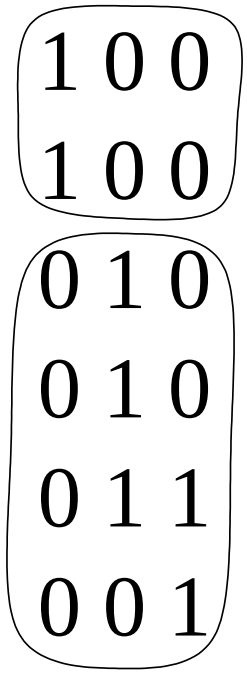
Mixtures of Independence Models

- Assume that data can be divided into C clusters
- Each cluster has an independence model

$$P(\mathbf{x}) = \sum_{c=1}^{C} P(c)P(\mathbf{x} \mid c) = \sum_{c=1}^{C} P(c) \prod_{d=1}^{D} P(x_d \mid c) \qquad (1)$$

- Parameters are estimated using the EM algorithm (update $P(c)$ and $P(\mathbf{x} \mid c)$ alternately while assuming the other one fixed)
- The number of clusters C is a free parameter
- Online time complexity is $O(QC)$

Mixtures of Independence Models Example



- $C = 2$ clusters
- $P(c = 1) = 1/3$, $P(c = 2) = 2/3$
- $P(x_1 = 1 \mid c = 1) = 1$, $P(x_2 = 1 \mid c = 1) = 0$, $P(x_3 = 1 \mid c = 1) = 0$
- $P(x_1 = 1 \mid c = 2) = 0$, $P(x_2 = 1 \mid c = 2) = 3/4$, $P(x_3 = 1 \mid c = 2) = 1/2$
- Note: In general, the division into clusters is fuzzy

Frequent Itemsets and Queries (Theory)

- Itemset: a conjunction of positively initialized attributes
- T-frequent itemset: itemset whose count in the data is at least T
- Theorem 1: Any subset of a T-frequent itemset is T-frequent as well
- Assume that we know the frequencies of all T-frequent itemsets
- Theorem 2: A query involving only variables in a particular T-frequent itemset can be answered exactly (without looking at the data)
  Idea of a proof: Itemset of size D has $2^D - 1$ nonempty subsets and the full probability distribution has also $2^D - 1$ degrees of freedom

Frequent Itemsets and Queries (Practice)

- There exist well-known efficient algorithms to find frequent itemsets
- Can be done offline
- Step from frequencies to probabilistic models
- All itemsets on a single attribute are included
- The information that certain itemsets are not T-frequent, is ignored
- The parameter T is left for user to adjust model complexity
  (the smaller the T the more itemsets are frequent)

Inclusion Exclusion Method

- Inclusion-exclusion principle:
  $P(x_1 = 0, x_2 \dots) = P(x_2 \dots) - P(x_1 = 1, x_2 \dots)$
- Transform the query into a sum of terms that use only positively initialized attributes ($\rightarrow$ itemsets)
- Assume that all not T-frequent itemsets have frequency 0
- Use the frequencies of the T-frequent itemsets found offline
- Note: does not always correspond to any distribution $P(\mathbf{x})$
- The frequency treshold T is a free parameter
- Online time complexity is $O(2^Q)$ (worst case: query with all 0's)

Inclusion Exclusion Example

1 0 0
1 0 0
0 1 0
0 1 0
0 1 1
0 0 1

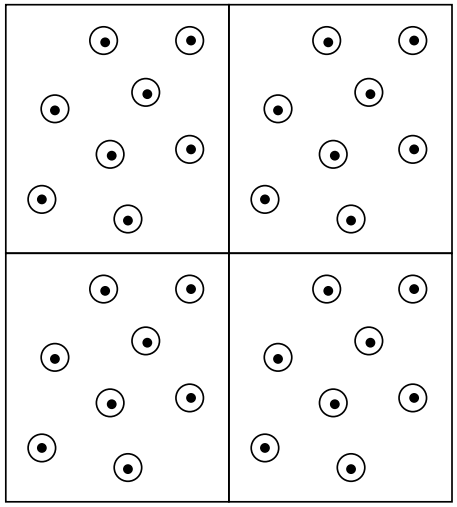
- Select $T = 1$
- Frequent itemsets are $\{x_1\}$, $\{x_2\}$, $\{x_3\}$ and $\{x_2, x_3\}$
- Corresponding frequencies are $1/3$, $1/2$, $1/3$ and $1/6$
- Query: "How often $x_1 = 0$ and $x_2 = 1$?"
- Answer: $P(x_1 = 0, x_2 = 1) =$
  $P(x_2 = 1) - P(x_1 = 1, x_2 = 1) = 1/2 - 0$ (correct)
- Note: $T = 1$ is low enough to always give correct answers - in a real problem it would lead to too many frequent itemsets

Maximum Entropy Model

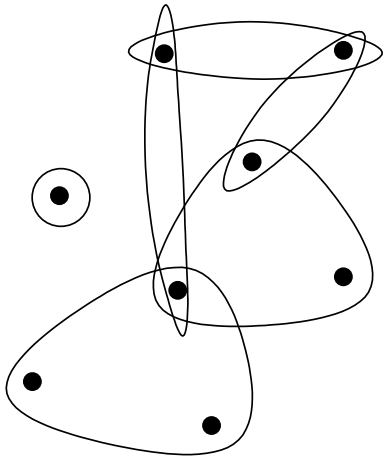
- Assume T-frequent itemsets known (T is a free parameter)
- Itemset frequencies are constraints to the distribution $P(\mathbf{x})$
- Select the most uninformed of those (maximum entropy)
- Corresponds to a Markov random field (MRF) with maximal frequent itemsets as cliques
- Online time complexity is $O(F^2 2^Q)$ where F is the number of frequent itemsets

Mixture and Maximum Entropy Models

Mixture of Independence Models



Maximum Entropy Model



- $D = 8$ dots represent attributes and ellipses represent explicit distributions

Algorithm: Iterative Scaling

- $$P(\mathbf{x}) = \mu_0 \prod_j \mu_j^{I(\mathbf{x} \text{ satisfies } V_j)},$$

  where V_j are the itemsets and I is an indicator function

- μ_j are estimated by iteratively enforcing each constraint:

$$\begin{aligned} \mu_0 &\leftarrow \mu_0 \frac{1 - f_j}{1 - S_j} \\ \mu_j &\leftarrow \mu_j \frac{f_j(1 - S_j)}{S_j(1 - f_j)}, \end{aligned}$$

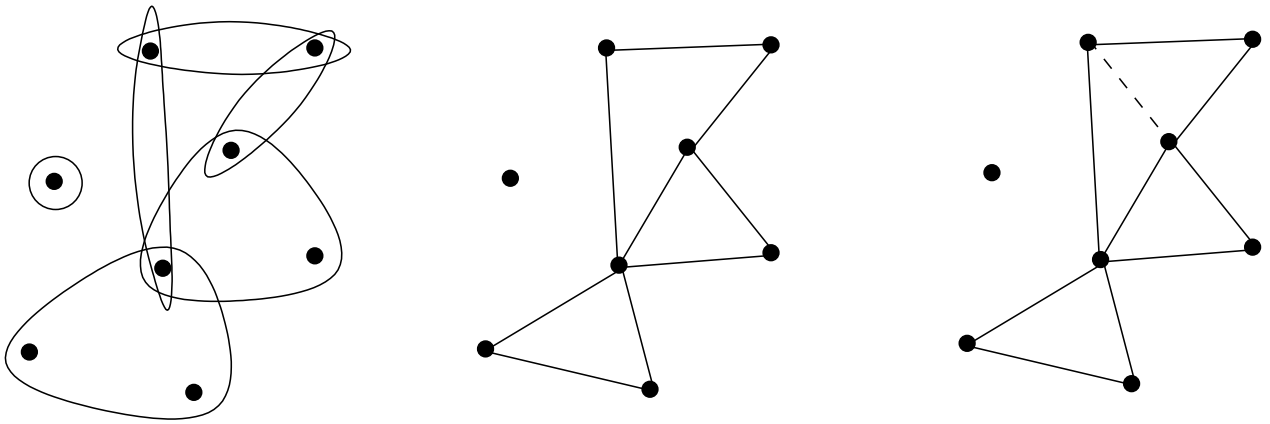
  where f_j is the ratio of the itemset V_j in the data and
  S_j is the probability of the itemset V_j given the current $P(\mathbf{x})$

- Converges to the unique maximum entropy solution

Speeding Up Iterative Scaling

- Size of $P(\mathbf{x})$ table is $2^Q \rightarrow$ Divide and conquer would help
- Find maximal cliques of the triangulated graph
- Place cliques into a join forest (like BN or MRF)
- Each smaller problem can be solved using iterative scaling

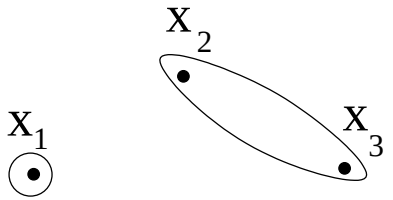
$$P(\mathbf{x}) = \frac{\prod P(\text{clique})}{\prod P(\text{clique intersection})}$$



Maximum Entropy Example

1 0 0
1 0 0
0 1 0
0 1 0
0 1 1
0 0 1

- Frequent itemsets as before
- Markov Random Field:



- Model: $P(x_1 = 1) = 1/3$, $P(x_2 = 1, x_3 = 1) = 1/6$, $P(x_2 = 1, x_3 = 0) = 1/3$, $P(x_2 = 0, x_3 = 1) = 1/6$
- In this case equivalent to the independence model, since the connection between x_2 and x_3 is useless

Connections of MaxEnt to others

- If $T = 0$, both inclusion exclusion method and maximum entropy method give perfect answers
- If $T = \infty$, the maximum entropy method becomes equivalent to the independence method
- If the frequent itemsets happen to match those of the Chow-Liu tree model, the methods are equivalent (both in efficiency and accuracy)
- The more sparse the data, the closer the frequency treshold approach is to the mutual information approach ("not occuring together" does not provide much information, but "occuring together" does)

Experiment Settings

- Microsoft Anonymous Web data set
  - 32711 records
  - 294 fields
  - 1.02% 1's (sparse)
- Consumer retail transactions
  - 54887 records
  - 52 fields
  - 7.86% 1's

Empirical Observations

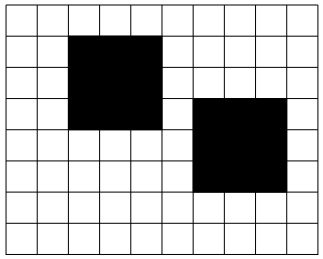
- Random samples cannot compete in online efficiency with the others
- Independence model is fastest and smallest, but least accurate
- Chow-Liu provides a modest improvement over the independence model
- Mixture models balance well memory, speed and accuracy
- Itemset inclusion-exclusion is very fast online, but requires a lot of memory. One of the best method with sparse data but not competitive with dense data.
- Maximum entropy is the most accurate with sparse data, but is very slow with large queries (and requires memory)

Discussion (presenters opinions) 1/2

- Frequency not the best measure of interestingness in general?
  - In sparse data, frequency of an itemset corresponds more closely to mutual information
  - MaxEnt works fine for sparse data but not with dense data
- Another formulation for a similar problem is:
  "What is the distribution of future data samples?"
  - Whole data scan is far from perfect (due to overfitting)
    $\longrightarrow$ Makes efficiency issues less important
  - Inclusion-exclusion does not work at all (does not define $P(\mathbf{x})$)
  - Would maximum entropy model show its true power?

Discussion (presenters opinions) 2/2

- Mixture model seems to perform well in all senses, but it might have difficulties with high dimensionality
- Let us think of binary images with some objects as data samples



- One cluster has to model all objects in an image
- In the maximum entropy model, each object corresponds to a frequent itemset and they are handled separately