An Introduction to Bayesian Networks

Torgeir Dingsoeyr

November 14, 1997

Bayesian Networks:

A statistical method for Data Mining, i.e. a statistical method for discovering valid, novel and potentially useful patterns in data.

Bayesian networks are used to represent essential information in databases in a network structure.

Figure 1: From data to knowledge in a Bayesian Network.

Why use Bayesian Networks?

- Easy to interpret for humans
- You can store probability relations
- Control inferences
- Produce explanations

Organization of material

- Theoretical basis: Bayes rule
- Bayesian Networks
- Bayesian inference
- Learning Bayesian Networks – Known Structure
- Learning Bayesian Networks – Unknown Structure
- How do we handle missing values?
- Examples of available tools
- Conclusion

Bayes rule

$P(A)$ denotes the probability that event A will occur. Belief measure obeys the three basic axioms of probability theory:

$$0 \leq P(A) \leq 1$$

$$P(sure\ proposition) = 1$$

$$P(A\ or\ B) = P(A) + P(B)$$

Bayes rule is the basis for Bayesian Networks:

Theorem. [Bayes] $P(H \mid e) = \frac{P(e|H)P(H)}{P(e)}$

Proof. The multiplication rule of probability theory is: $P(AB) = P(A \mid B)P(B)$

Then: $P(H \mid e) = \frac{P(He)}{P(e)}$

By swapping He and using the multiplication rule again:

$= \frac{P(e|H)P(H)}{P(e)}$

$\square$

Where H is a hypothesis, and e an event. $P(e \mid H)$ is the posterior probability, and $P(H)$ is the prior probability.

Excercise

There are three urns labeled one, two and three which contains balls with different colours:

Urn	White	Black
1	3	3
2	4	2
3	1	2

A) Find the probability of selecting urn 2 and drawing a black ball.

B) Find the probability of drawing a black ball.

C) Find the conditional probability that urn 2 was selected, given that a black ball was drawn.

Solution:

We first label the possible outcomes, and give the probabilities:

(1W)	(1B)	(2W)	(2B)	(3W)	(3B)
$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{9}$

Then:

A) $P(S2\ DB) = \frac{1}{9}$ (From the table)

B) $P(DB) = P(1B) + P(2B) + P(3B) = \frac{1}{6} + \frac{1}{9} + \frac{1}{9} = \frac{7}{18}$

C) $P(S2|DB) = \frac{P(S2\ DB)}{P(DB)} = \frac{\frac{1}{9}}{\frac{7}{18}} = \frac{2}{7}$

Bayesian Networks

A Bayesian Network consists of vertices which are events and edges which are relations between events.

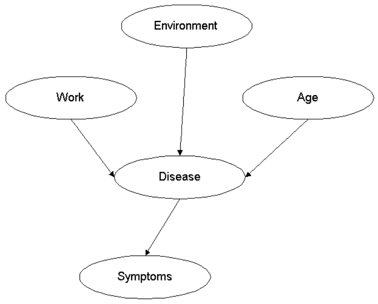

Figure 2: An Example Bayesian Network.

Some terminology:

If a subset of Z nodes in a graph G intercepts all paths between the nodes X and Y ($\langle X \mid Y \mid Z \rangle_G$), then this corresponds to conditional independence between X and Y given Z:

$\langle X \mid Y \mid Z \rangle_G \Rightarrow I(X, Z, Y)_M$

conversely:

$I(X, Z, Y)_M \Rightarrow \langle X \mid Y \mid Z \rangle_G$

with respect to some dependency model M.

Figure 3: X is independent of Y given Z.

Definition. [I-map] *A DAG (Directed, Acyclic Graph) D is said to be a I-map of a dependency model M if for every three disjoint sets of vertices, X, Y and Z we have:*

$\langle X \mid Y \mid Z \rangle_G \Rightarrow I(X, Z, Y)_M$

A DAG is a minimal I-map of M if none of its arrows can be deleted without destroying its I-mapness.

Definition. [Bayesian Network] *Given a probability distribution P on a set of variables U, a DAG $D = (U, \vec{E})$ is called a Bayesian network of P iff D is a minimal I-map of P.*

A Bayesian Network is shown in Fig 4, representing the probability distribition P:

$P(x_6 \mid x_5)P(x_5 \mid x_2, x_3)P(x_4 \mid x_2, x_1)P(x_3 \mid x_1)P(x_2 \mid x_1)P(x_1)$

Figure 4: A Bayesien Network representing the distribution P

What can you do with a Bayesian Network?

Bayesian Inference

To infer knowledge from a Bayesian Network, we use Bayesian Inference. First we introduce some terminology:

Definition. [Prior Odds] *Predictive support with only background knowledge:* $O(H) = \frac{P(H)}{P(\neg H)} = \frac{P(H)}{1-P(H)}$

Definition. [Likelihood Ratio] *Diagnostic support given to H by observed evidence:* $L(e \mid H) = \frac{P(e|H)}{P(e|\neg H)}$

Definition. [Posterior Odds] *Predictive support given observed evidence, e:* $O(H \mid e) = \frac{P(H|e)}{P(\neg H|e)} = L(e \mid H)O(H)$

By using these definitions, we can infer knowledge like in probability theory.

Exercise:

A salesman has installed an alarm in a shop, and it is known that the alarm obeys these probabilities:

$$P(Alarm \mid Burgular) = 0.95$$

$$P(Alarm \mid \neg Burgular) = 0.01$$

$$P(Burgulary) = 10^{-4}$$

What is the probability that it is a burgulary if the alarm goes?

Solution:

First we find the posterior oddsof ti being a burgular:

$$O(Burgular \mid Alarm) = L(Alarm \mid Burgular)$$

$$*O(Burgular) = \frac{0.95}{0.01}10^{-4} = 0.0095$$

To convert odds to probability, we use: (deducted from the prior odds)

$$P(A) = \frac{O(A)}{1 + O(A)}$$

This implies:

$$P(Burgular \mid Alarm) = \frac{0.0095}{1 + 0.0095} = 0.00941$$

Learning Bayesian Networks - Known Structure

Problem: How to learn the parameters for the graph from the database.

Example: Consider the flipping of a thumbtack. It can either land on the "head" or "tail". We have a database D with outcomes, and want to learn the distribution of the variable which we can call θ.

It can be shown that:

$$P(\theta \mid h\ heads, t\ tails, \epsilon) = c\theta^h(1-\theta)^t P(\theta \mid \epsilon)$$

Where ϵ is the background knowledge, h the number of heads, t the number of tails, c is a normalization constant. This result is only dependent on the data in the database D.

Learning Bayesian Networks - Unknown Structure

Problem: First learn the structure of the graph, then the parameters.

Figure 5: Some Bayesian Networks With Three Variables.

Learn the structure by:

- Constructing model space.
- Evaluate strengh with metric.

The Bayesian Information Criterion works like this:

$BIC(S_m|sample) = -\log p(sample|\hat{\theta_m}, S_m) + \frac{1}{2}dim(\theta_m)\log N.$

$\hat{\theta_m}$ is the maximum likelihood estimate of θ_m in structure S_m. N is the sample size and dim(θ_m) the dimensionality.

Several other metrics exist for evaluating the network:

- AIC – A Information Criterion
- MAP – Maximun a Posteriori

Problem: overfitting

How do we handle missing variables

When variables are missing, they can be filled in using the Missing Information Priciple:

Fill in the data using the available information

Methods exist such as Gibbs Sampling (A Markov Chain Monte Carlo method).

These methods are, however, time-consuming.

Example Application: AutoClass

An unsupervised Bayesian learner which learns classes from data. It uses the following model:

$$P(X_i \in C_j \mid \vec{V_c}, T_c, S, \mathcal{I})$$

Where X_i is an instance (object), C_j a class. T_c and $\vec{V}$ are the functional form and parameter values of the probability distribution for the class. S is the space of allowed probability distributions and $\mathcal{I}$ is implicit information not specifically represented.

Each class is then modelled with:

$$P(\vec{X_i} \mid X_i \in C_j, \vec{V_j}, T_j, S, \mathcal{I})$$

which is the probability of observing the instance attribute values $\vec{X_i}$ conditional on the assumption that instance X_i belongs in class C_j.

Example: AutoClass has been used on data from a infrared astronomical satellite. The data contained 5425 specra (objects) with 200 channels (attributes). AutoClass was able to make many subtle distinctions between spectra that look superficially similar, which were previously unknown. It also:

- tripled the number of suspected carbon stars
- revealed a large number of black body stars

which was earlier unknown to the experts in the field.

Example Application: Bayesian Knowledge Discoverer

Bayesian Knowledge Discoverer can be used to discover Bayesian Networks from incomplete data. It can predict variable values, explain observations, and visualize dependencies among variables.

It differs from AutoClass in that it can use Gibbs sampling to fill inn missing information, and it is both able to learn the structure and the parameters for a Bayesian Network.

Conclusions

Bayesian Networks is a solid method for Data Mining, that can:

- handle missing data, incomplete values
- combine domain knowledge and data
- produce easy-to understand graphs

Problems that remain:

- problem1

References

Wray Buntine. A guide to the literature on learning probabilistic networks from data. Transactions on Knowledge and Data Engineering.

David Heckerman. Bayesian Networks for Knowledge Discovery. In Advances in Knowledge Discovery and Data Mining. MIT/AAAI Press, 1996.

Judea Pearl. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, 1988.

Marco Ramoni Paola Sebastiani. Discovering Bayesian Networks in Incomplete Databases. Technical Report – The Knowledge Media Institute, 1997

B.D. Ripley. Pattern Recognition and Neural Networks. Cambridge University Press, 1996.

Handouts and more several links to software and papers can be found on:

`http:\\www.idi.ntnu.no\~dingsoyr`