Bayesian Networks
Hidden Markov Models

Description and applications

Rafael Pérez Torres
Selected Topics on Pattern Recognition
Profesor: Dr. Wilfrido Gómez Flores
LTI Cinvestav

Jun 19th, 2015

Agenda

Introduction to Bayesian Networks (BN)

- Some classification techniques suffer the *curse of dimensionality* because their focus on all dependencies between attributes.
- Other techniques (like Naïve Bayes) remove all dependencies making the model really simply.
- Sometimes an intermediante approach is best.

Introduction to Bayesian Networks (BN) II

- Such approach can be obtained through the *chain rule*:

$$p(x_1, x_2, \ldots, x_l) = p(xl|x_{l-1}, \ldots, x_1) \ldots p(x_2|x_1)p(x_1) \qquad (1)$$

- This rule states that the joint PDF can be expressed in terms of a product of several conditional PDFs and a marginal one.
- Then it is possible to select the features that will hold a conditional dependency with an attribute x_i, re-written as:

$$p(\mathbf{x}) = p(x_1) \prod_{i=2}^{l} p(x_i|A_i) \qquad (2)$$

where donde $A_i \subseteq \{x_{i-1}, x_{i-2}, \ldots, x_1\}$

INTRODUCTION TO BAYESIAN NETWORKS (BN) III

- For example, let $l = 6$ and

$$
\begin{aligned}
p(x_6|x_5,\ldots,x_1) &= p(x_6|x_5,x_4) && (3)\\
p(x_5|x_4,\ldots,x_1) &= p(x_5|x_4) && (4)\\
p(x_4|x_3,x_2,x_1) &= p(x_4|x_2,x_1) && (5)\\
p(x_3|x_2,x_1) &= p(x_3|x_2) && (6)\\
p(x_2|x_1) &= p(x_2) && (7)
\end{aligned}
$$

INTRODUCTION TO BAYESIAN NETWORKS (BN) III

- The assumptions made would be $A_6 = \{x_5, x_4\}, A_5 = \{x_4\}, A_4 = \{x_2, x_1\}, A_3 = \{x_2\}, A_2 = \emptyset$.
- Such assumptions can be represented like on the figure.
- x_i is conditionally independent of any combination of its *non-descendents* given its parents.
- Then Naïve Bayes can be considered a special case where $A_i = \emptyset, i = 2, \ldots, l$

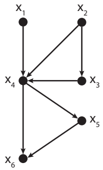

Figure: Conditional dependencies

BN Definition

Definition

- A BN is an directed acyclic graph (DAG) where nodes correspond to random variables (the features).
- Each node is associated with a set of conditional probabilities $p(x_1|A_i)$, where x_i is the variable associated with such node and A_i is the set of its parents in the DAG.
- The full specification of a BN requires:
  1. The probabilities of root nodes.
  2. The conditional probabilities of non-root nodes.
- Joint probability is obtained multiplying such probabilities.
- The required work is a *topological sorting* of the random variables, making that each appears before its descendents in the DAG.

BN Applications

- BN have been widely employed in medical diagnosis, but is useful in any area where probability inference is needed.

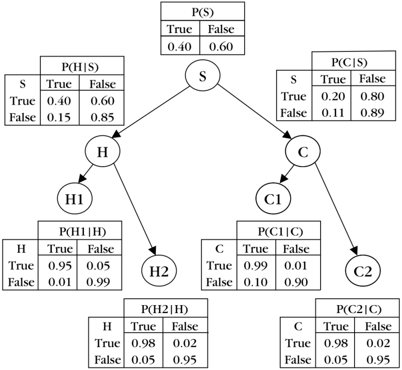

Figure: BN modeling conditional dependencies for smokers (S) and their tendencies for developing cancer (C) and heart diseases (H), along associated variables for heart (H1, H2) and cancer tests (C1, C2)

BN Training

- Structural training:
    - Trees
    - Poli-trees
    - Multiconnected networks
- Parametric training:
    - Given enough data: frequency analysis.
    - Not enough data: EM algorithm.

Introduction to Hidden Markov Models (HMM)

- Useful when data to be analyzed come from systems where time is important.
- Then, the patterns are heavily linked to time.
- Commands given to a computer, sequence of phonems in spoken words, any event-based situation.

Introduction to Hidden Markov Models (HMM) II

Pattern types:

- **Deterministic**: Generated by deterministic systems; each state depend solely on the previous state, and can be represented as a SM.

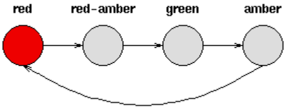

Figure: Traffic lights

- **Non-deterministic**: States do not exhibit a fixed sequence. However, it is still possible to make an approximation to the system.

Markov process

- A weather model defines sunny, rainy and cloudy states.
- We can assume that current state depends only on previous states.
- Previous idea is the **Markov property**.
- It is based on that *Given the present, past and future are independent*.

MARKOV PROCESS II

- A Markov process moves from one state to another depending on n previous states along ***discrete*** time.
- n states affecting choice? *n-order* Markov process.

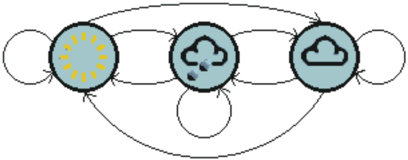

Figure: First order transitions on weather

MARKOV PROCESS III

- Each transition is dictated by a probability.
- All M^2 probabilities build the state transition matrix .

$$
\begin{array}{cc}
 & \textit{Today} \\
\textit{Yesterday} & \begin{array}{c} \\ \text{sun} \\ \text{cloud} \\ \text{rain} \end{array}
\begin{bmatrix} \text{sun} & \text{cloud} & \text{rain} \\ 0.50 & 0.375 & 0.125 \\ 0.25 & 0.125 & 0.625 \\ 0.25 & 0.375 & 0.375 \end{bmatrix}
\end{array}
$$

- Sum of each row equals 1.
- Probabilities in initial state are given by the π vector

$$
\begin{array}{ccc} \text{Sun} & \text{Cloud} & \text{Rain} \end{array}
$$
$$
\begin{pmatrix} 1.0 & 0.0 & 0.0 \end{pmatrix}
$$

Markov process definition

Definition

- A set of discrete states.
- π vector for states probability at t_0.
- State transition matrix. Fixed along system execution.

Any process that can be described on this way qualifies as a Markov process.

HIDDEN MARKOV MODELS

- A Markov process may not be enough to describe a system.
- Suppose you can not observe directly the system, but another set of input that is known to be related to the system states.
- For example, know the weather only from seaweed, identify a word by its speech.
- What is accessible is inside the *visible states*.
- What is not accessible is inside the *hidden states*.
- Usually, the hidden states hold the important information of the system.

Hidden Markov Models II

- Amount of observable and hidden states does not need to be the same.
- A HMM models the link between the hidden and visible states through an Markov process hidden.
- Because of that it is also called a *double embedded stocastic process*.
- An additional matrix (*confusion matrix*) is needed to hold probabilities to reach a visible state from a given hidden state.

Seaweed

weather		Dry	Dryish	Damp	Soggy
	Sun	0.60	0.20	0.15	0.05
	Cloud	0.25	0.25	0.25	0.25
	Rain	0.05	0.10	0.35	0.50

Hidden Markov Model definition

Definition

A HMM is a triple (π, A, B)

- $\Pi = (\pi_i)$ vector of probabilities in initial state.
- $A = (a_{ij})$ state transition matrix, holding probabilities $P(x_{i_t} | x_{j_{t-1}})$.
- $B = (b_{i_j})$ confusion matrix, holding probabilities $P(y_t | x_j)$

Applications

Once a system is described through an HMM, three applications can be given to it.

- Evaluation
- Decoding
- Learning

EVALUATION

Evaluation

Given a set of HMM (several triple (π, A, B)) describing different systems, and a sequence of observations we need to know which HMM is the most probable to have generated such sequence.

- A HMM for summer, another for winter.
- Recognize a word from a given speech.

The forward algorithm is widely employed to solve this problem.

DECODING

Decoding

Finding the most probable sequence of hidden states given some observations.
The example of seaweed and weather falls in this category.
The Viterbi algorithm is employed for solving this problem. A widely usage of Viterbi algorithm is on the NLP (Natural Language Processing) for tagging words according to its syntactic class (verb, noun, ...).

LEARNING

Learning

Generating a HMM from a sequence of observations.
Given a sequence of observations (from a known set), known to represent a set of hidden states, fit the most probable HMM; that is, determine the (π, A, B) triple that most probably describes what is seen.
The forward-backward algorithm is helpful to approximate a solution for this problem.

–

Thank you!