Bayesian Network

1st April 2016

Outline

1. Probability Fundamentals
2. Bayesian Network
3. Training of BN
4. Construction of BN
5. Bayesian Classifier

Probability

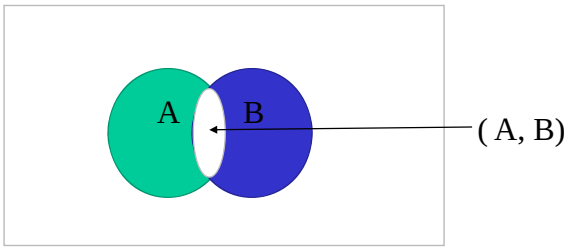
- Probability – the chance or likelihood that an uncertain (particular) event will occur
  - Probability is always between 0 and 1, inclusive
- Take events A_i for I = 1 to k to be:
  - Mutually exclusive: $A_i \cap A_j = 0$ for all i,j
  - Exhaustive: $A_1 \cup \cdots \cup A_k = S$

Conditional Probability

- Chance" of an event given that something is true

Notation: P(a|b)

- Probability of event a, given b is true



Conditional Probability

- Diagnosis using a clinical test
- Sample Space = all patients tested
  Event A: Subject has disease
  Event B: Test is positive

Interpret:

- Probability patient has disease and positive test (correct!) $p(A \cap B)$
- Probability patient has disease BUT negative test (false negative) $p(A \cap B')$
- Probability patient has no disease BUT positive test (false positive) $p(A' \cap B)$
- Probability patient has disease given a positive test $p(A|B)$
- Probability patient has disease given a negative test $p(A|B')$

Basic Formulas for Probabilities

• Product Rule : probability P(AB) of a conjunction of two events A and B:

$$P(A,B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

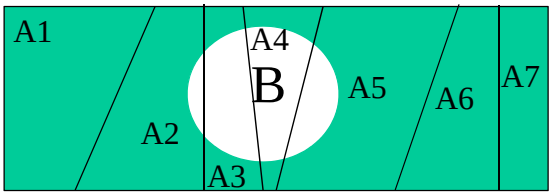
•Sum Rule: probability of a disjunction of two events A and B:

$$P(A+B) = P(A) + P(B) - P(AB)$$

Basic Formulas for Probabilities

•Theorem of Total Probability : if events A1, …., An are mutually exclusive with

$$P(B)=\sum_{i=1}^{n}P(B|A_i)P(A_i)$$



Bayes Theorem

Bayes Theorem:

$$p(A_j|B)=\frac{p(A_j\cap B)}{p(B)}=\frac{p(B|A_j)\cdot p(A)}{\sum_{i=1}^{k}p(B|A_i)\,p(A_i)}$$

- P(A) = prior probability of event A
- P(B) = prior probability of event B
- P(A|B) = probability of A given B (posterior or updated probability)
- P(B|A) = probability of B given A (likelihood of B given A)

Bayes Theorem

The Goal of Bayesian Learning: the most probable hypothesis given the training data (Maximum A Posteriori hypothesis)

$$
\begin{aligned}
h_{map} &= \max_{h \in H} P(h \mid D) \\
&= \max_{h \in H} \frac{P(D \mid h) P(h)}{P(D)} \\
&= \max_{h \in H} P(D \mid h) P(h)
\end{aligned}
$$

An Example

Does patient have cancer or not?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this cancer.

$$
\begin{aligned}
&P(cancer) = .008, P(\neg cancer) = .992 \\
&P(+ \mid cancer) = .98, P(- \mid cancer) = .02 \\
&P(+ \mid \neg cancer) = .03, P(- \mid \neg cancer) = .97 \\
&P(cancer \mid +) = \frac{P(+ \mid cancer) P(cancer)}{P(+)} \\
&P(\neg cancer \mid +) = \frac{P(+ \mid \neg cancer) P(\neg cancer)}{P(+)}
\end{aligned}
$$

Repetition of Dependent Events

Relies on conditional probability calculations.
If a sequence of outcomes is {*A,B,C*}

$$P(A \cap B \cap C) = P(C \mid A \cap B) \cdot P(A \cap B)$$
$$= P(C \mid A \cap B) \cdot P(A \mid B) \cdot P(A)$$

This is the basis of **Markov Chains.** Probability of the sequence is given by the product of the probability of the first event with the probabilities of all subsequent occurrences

Bayesian Inference

- Allows us to combine observed data and prior knowledge
- Provides practical learning algorithms
- It is a generative (model based) approach, which offers a useful conceptual framework
  - –This means that any kind of objects (e.g. time series, trees, etc.) can be classified, based on a probabilistic model specification

Bayesian Network

- A Bayesian network is a graphical model for probabilistic relationships among a set of variables.

- Gives Causal Relationships between different variables
- Belief Network
- Components:

1. Directed Acyclic Graph (DAG)
2. Conditional Probability Tables

Usages

1. To predict outcomes or diagnose
2. To discover causal relationships when structures are unknown (Mostly in Biology)
3. Financial Analysis

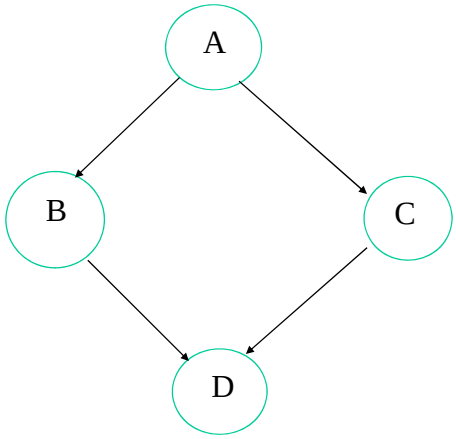
WHY BN ?

- Readily handle incomplete data sets
- Allows us to learn about causal relationships (which gives an understanding about problem domain)
- during exploratory data analysis. In addition, knowledge of causal relationships allows us to make predictions in the presence of interventions.

WHY BN ?

- For example, a marketing analyst may want to know whether or not it is worth while to increase exposure of a particular advertisement in order to increase the sales of a product
- combines **domain** knowledge and data.
- Strength of **causal** relationships can be encoded with probabilities.
- Offers an efficient approach for avoiding the over fitting of data (no requirement of holding data for testing)

Bayesian Network



DAG

P(A)

P(B|A)

P(C|A)

P(D|B,C)

Conditional Probability Table

Bayesian Network

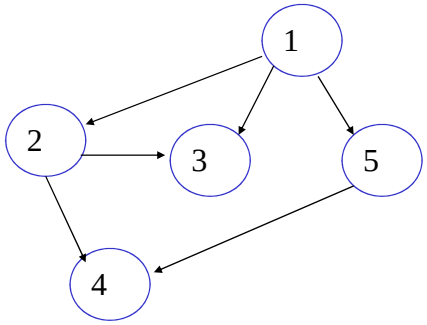
Input : fully or partially observable data cases

Output : parameters (probability tables) AND also structure

- Supervised Learning: BN and other approaches are similar. But when no cause/ effect relationships are known, BN are considered superior
- Unsupervised Learning: BN can be used.

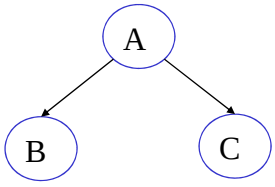
Directed Acyclic Graph (DAG)

- It is a graph with no direct cycles
- Need to ensure that there is at least one node with no child
- BN is not a Tree



Training

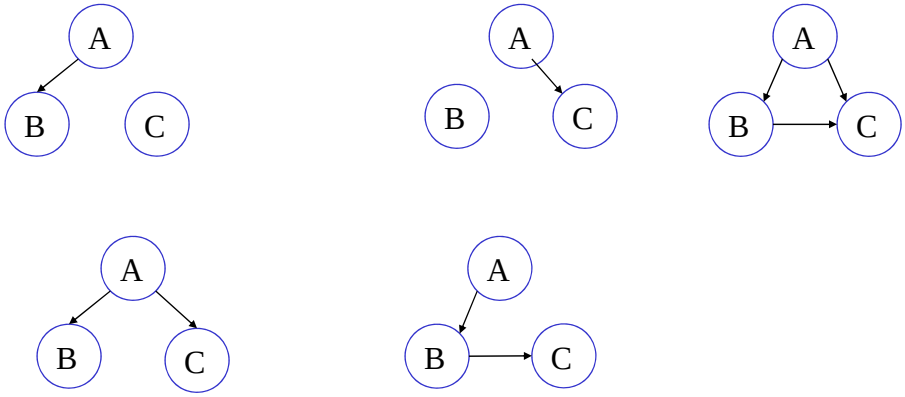
1. Construction of Probability tables



	A=a1	A=a2	A=a3
B=b1	0.5	0.2	0.3
B=b2	0.1	0.6	0.3
B=b3	0.3	0.3	0.4

Construction of BN

Possible Structures



D-Separation

- Nodes A and B are d-separated if on any (undirected) path between A and B there is some variable C such that is either C is in a serial or diverging connection and C is known, or C is in a converging connection and neither C nor any of C's descendants are known.
- If nodes A and B are d-separated by C, then A and B are conditionally independent given C.

Learning Structure (from Data)

- Heuristic Search Algorithms:

Complete data – local computations Incomplete data (score non decomposable):stochastic methods Local greedy search; **K2 algorithm**

- Constrained based methods

Data impose independence relations (constraints) on graph structure or network

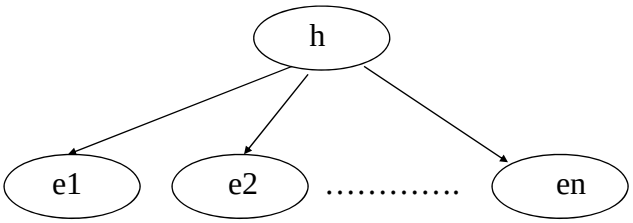
Model Selection

- search algorithm is used to find an equivalence class that receives a high score

1. Naïve Bayes (NB):
   Too simple (less parameters, but bad model)

2. TAN: Tree-augmented Naïve Bayes,
   Too complex (possible over fitting + complexity)

Special Case : Naïve Bayes



P(e1, e2, ……en, h) = P(h) P(e1 | h) …….P(en | h)

Missing Data (Information)

- Missing information can be simulated (should fit the distribution)
- Monte Carlo Simulations

Example

- **Detecting Credit Card Fraud**

- Possible choice of variables for this problem is :
- **Fraud (F)**, whether the current purchase is fraudulent or not

- **Gas (G),** whether or not there was a gas purchase in the last 24 hours
- **Jewelry (J),** whether or not there w as a jewelry purchase in the last 24 hours
- **Age (A)** and
- **Sex (S)**

Example

- **Approach**

1. Identify the goals of modeling (e.g., prediction versus explanation versus exploration)
2. Identify many possible observations that may be relevant to the problem
3. Determine what subset of those observations is worth while to model
4. Organize the observations in to variables having mutually exclusive and collectively exhaustive states.

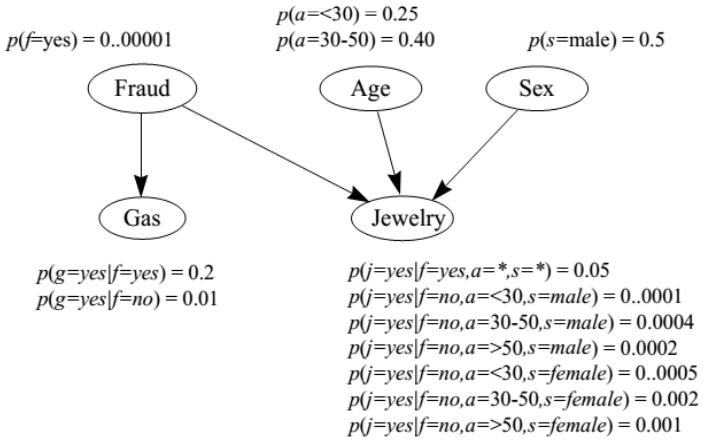
Example

- **Structure determination**

1. Order the variables (ex. J ; G; S; A; F)) , total 5! Ways are possible (serious problem)
2. Then Look for independency or dependency between these variables

Instead,
1. Establish causal relationships among variables
2. Look for conditional dependence

Example



Example

- **Inference:** The computation of a probability of interest given a model

$$p(f|a,s,g,j) = \frac{p(f,a,s,g,j)}{p(a,s,g,j)} = \frac{p(f,a,s,g,j)}{\sum_{f'} p(f',a,s,g,j)}$$

$$\begin{aligned} p(f|a,s,g,j) &= \frac{p(f)p(a)p(s)p(g|f)p(j|f,a,s)}{\sum_{f'} p(f')p(a)p(s)p(g|f')p(j|f',a,s)} \\ &= \frac{p(f)p(g|f)p(j|f,a,s)}{\sum_{f'} p(f')p(g|f')p(j|f',a,s)} \end{aligned}$$

Bayesian Classification: Why?

- Probabilistic learning: Calculate explicit probabilities for hypothesis, among the most practical approaches to certain types of learning problems
- Incremental: Each training example can incrementally increase/decrease the probability that a hypothesis is correct. Prior knowledge can be combined with observed data.
- Probabilistic prediction: Predict multiple hypotheses, weighted by their probabilities
- Standard: Even when Bayesian methods are computationally intractable, they can provide a standard of optimal decision making against which other methods can be measured

MAP Learner

For each hypothesis h in H, calculate the posterior probability

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

Output the hypothesis h_{map} with the highest posterior probability

$$h_{map} = \max_{h \in H} P(h \mid D)$$

Comments:

- Computational intensive
- Providing a standard for judging the performance of learning algorithms
- Choosing P(h) and P(D|h) reflects our prior knowledge about the learning task

Naïve Bayes Classifier

- What can we do if our data *d* has several attributes?
- Naïve Bayes assumption: Attributes that describe data instances are conditionally independent given the classification hypothesis

$$P(\mathbf{d} \mid h) = P(a_1, ..., a_T \mid h) = \prod_t P(a_t \mid h)$$

  - it is a simplifying assumption, obviously it may be violated in reality
  - in spite of that, it works well in practice

Naïve Bayes Classifier

- The Bayesian classifier that uses the Naïve Bayes assumption and computes the MAP hypothesis is called Naïve Bayes classifier
- One of the most practical learning methods
- Successful applications:
  - Medical Diagnosis
  - Text classification

Example. 'Play Tennis' data

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Based on the examples in the table, classify the following datum ***x***:
x=(Outl=Sunny, Temp=Cool, Hum=High, Wind=strong)

- That means: Play tennis or not?

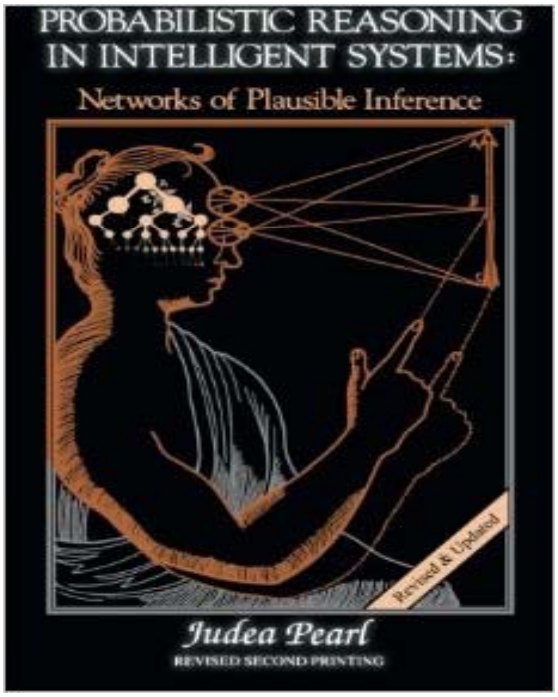
$$h_{NB} = \arg\max_{h \in [yes,no]} P(h)P(\mathbf{x} \mid h) = \arg\max_{h \in [yes,no]} P(h)\prod_t P(a_t \mid h)$$
$$= \arg\max_{h \in [yes,no]} P(h)P(Outlook = sunny \mid h)P(Temp = cool \mid h)P(Humidity = high \mid h)P(Wind = strong \mid h)$$

- Working:

$$P(PlayTennis = yes) = 9/14 = 0.64$$
$$P(PlayTennis = no) = 5/14 = 0.36$$
$$P(Wind = strong \mid PlayTennis = yes) = 3/9 = 0.33$$
$$P(Wind = strong \mid PlayTennis = no) = 3/5 = 0.60$$
etc.
$$P(yes)P(sunny \mid yes)P(cool \mid yes)P(high \mid yes)P(strong \mid yes) = 0.0053$$
$$P(no)P(sunny \mid no)P(cool \mid no)P(high \mid no)P(strong \mid no) = \mathbf{0.0206}$$
$$\Rightarrow answer : PlayTennis(x) = no$$

Learning to classify text

- Learn from examples which articles are of interest
- The **attributes** are the **words**
- Observe the Naïve Bayes assumption just means that we have a random sequence model within each class!
- NB classifiers are one of the most effective



To Download GeNIe:
http://download.bayesfusion.com/files.html?category=Academia


Thank You