An Introduction to Probabilistic Decision-Support Systems

David Bellot

david.bellot@gmail.com

May 20th 2009
Outline

1. Basic Components for Probabilistic Reasoning Systems
2. Introduction to Graphical Models
3. Bayesian Inference and Machine Learning
4. Time Series and Applications
Building and using Probabilistic models

- Easy design of new and enhanced probabilistic models
- Making faster computations to deal with more data and/or bigger models
- Control, analysis, explanation, deduction...
- Probabilistic Graphical Models represent an evolution of legacy prob/stat models
- Applications in finance, control, vision, medicine, communications, robots, games, speech, ...
Rational agent, from logic to uncertainty

- An agent represents its knowledge about the world: it’s the view it has from the world, not an absolute description of the world
- **Logical agent**: everything is true, false or unknown
- **Agent under uncertainty**: its view of the world might be partial, imperfect, **uncertain**

### Rational decision

- **Rational decision in a perfect world**: take the best decisions to reach the goal using logical reasoning
- **Rational decision under uncertainty**: take the best decision that maximizes the expected utility
Probabilities

- Degree of belief that a sentence is true: *probability*.
- **Probabilities**: a fact do or do not hold in the world
- **Fuzzy logic**: a fact partly holds in the world.
- **Examples**:
  - probability of 0.8: that is a *strong* expectation the fact to be true
  - possibility of 80%: the fact is only true at 80%, 20% remains for something else
  - €/$ = 1.3126. Now, it’s true.
  - I believe with a strength of 76% that the next value in 10 minutes will be 1.3146
Initially, I have beliefs about the world (personal beliefs, statistics, ...): prior or unconditional probabilities,
I know a new fact about the world: how do my beliefs change?
Conditional or posterior probabilities: I updated my beliefs about the world
Example:
€/$ value is on average 1.3126. I believe it at 95% of probabilities
€/£ increases, what is my new belief?
What is $P(€/$ = 1.3126 | €/£ ↑) ?
Conditional Probabilities II

- **Probability**
  - $P(€/\$ = 1.3126) = 0.76$
  - $P(\text{head}) = 0.49$

- **Conditional probability**
  - $P(a|b) = \frac{P(a \land b)}{P(b)}$, $P(b) > 0$
  - $P(a \land b) = P(a|b).P(b) = P(b|a).P(a)$  **product rule**
  - $P(a|b)$ is not a logical implication with uncertainty added.
    - $P(a|b) = 0.8 \neq (b \text{ holds } \Rightarrow P(a) = 0.8)$

- **Semantic is the Kolmogorov’s axioms:**
  - $0 \leq P(a) \leq 1$
  - $P(\text{true}) = 1$ and $P(\text{false}) = 0$
  - $P(a \lor b) = P(a) + P(b) - P(a \land b)$
Inference on joint probability distribution

- A joint probability distribution is a distribution over several variables
- \( P(X_1 \ldots X_n), P(\varepsilon/\$, £/¥), P(\text{toothache, catch, cavity}), \ldots \)
- **Inference:**
  - \( P(X_{12} | X_5 = \text{true}, X_3 = \text{blue}) \)
  - \( P(\varepsilon/\$ | £/¥ = 147.23) \)
- **Query:**

\[
P(X | e) = \alpha P(X, e) = \alpha \sum_Y P(X, e, y)
\]
Marginalization and Conditioning

- **Marginalization**: summing out a subset of variables
  \[ P(X_j \ldots X_k) = \sum_{\{1\ldots n\setminus j\ldots k\}} P(X_1 \ldots X_n) \]

- **Conditioning**:
  \[ P(X) = \sum_z P(X|z)P(z) \]

- **BUT**: summing out takes an exponential time to compute in an exponentially sized **table of probabilities**
  - \( n \) Boolean variables: \( O(2^n) \) to sum out from a \( O(2^n) \) tables of probabilities!
  - Imagine when you have more than 2 possible values for each variables
Conditional Independence

- **Independence:**
  - assumptions about independence or conditional independence between variables?
  - Use it to reduce the size of the problem: **sparsity**
  - \( P(X | Y) = P(X) \) if \( X \perp Y \). And \( P(Y | X) = P(Y) \) by the way...
  - \( P(X, Y) = P(X).P(Y) \)
  - Flip \( n \) coins: \( P(C_1 \ldots C_n) = \prod P(C_i) \)

- **Conditional independence:**
  - \( P(\text{Toothache}, \text{Cavity}, \text{Catch}, \text{Weather}) = P(\text{Cavity}).P(\text{Toothache}|\text{Cavity}).P(\text{Catch}|\text{Cavity}).P(\text{Weather}) \)
  - comes from causal relationships between variables
  - **example:** Naive Bayes, ARMA, Kalman filter, etc...
Bayes’s rule

\[ P(a|b) = \frac{P(b|a) \cdot P(a)}{P(b)} \]

\[ P(Y|X, e) = \frac{P(X|Y, e) \cdot P(Y|e)}{P(X|e)} \]
Graphical Models: introduction

- Marriage between Statistics and Graph Theory
- Factorization of a joint probability distribution
- Equivalence with a graph representing causality
- **Tractable models, generalize other models**

**Figure:** A simple directed graphical model (also called **Bayesian Network**

```
Weather -> Cavity
   
   Toothache  

Cavity -> Catch
```
Semantic

- **Directed Graphical Models**
  (Bayesian Networks)
  \[
P(x_1 \ldots x_n) = \prod_{i=1}^{n} p(x_i | \text{parents}(x_i))
\]
  where \( x_i \) is directly dependant upon the subset \( \text{parents}(x_i) \)

- **Undirected Graphical Models**
  (Markov Random Fields)
  \[
P(x_1 \ldots x_n) = \prod_{C \in \text{clique}(G)} \phi_C(x_C)
\]

*Figure: Graphical Models*
Construction of models

- Idea: \( P(x_1 \ldots x_n) = P(x_n|x_{n-1} \ldots x_1) \cdot P(x_{n-1}|x_{n-2} \ldots x_1) \cdots P(x_2|x_1)P(x_1) \) in general (chain rule)
- Use the chain rule with a specific order and with conditional independence assumptions
- Local structure: in general grows linearly in size
- Variables are discrete or continuous

Continuous:

- one of the main advantage of Graphical Models
- Define conditional distributions of discrete or continuous given discrete or continuous
- Examples: linear gaussian, mixtures, exponential families
- apply the same principles AND the same algorithms
Queries and Use: Inference

- **Vocabulary:**
  - **Inference** is computing a posterior probability (or distribution) given some evidences.
  - **Learning** is finding, fitting the parameters, or finding the decomposition or finding the probability forms, etc...

- **Many methods for inference:**
  - **Enumeration of the state space:** useful if you have one million years ahead.
  - **Variable elimination:** repetitive marginalization and conditioning. Basis for message-passing.
  - **Sampling:** Monte-Carlo methods, Gibbs, MCMC.
  - **Optimization:**
    - **Variational methods:** one of the fastest approach. Lower and upper bounds on the probability.
    - **Asymptotic methods,**...
Variable elimination

- Enumeration makes redundant computations.
- Variable elimination: do computations once and save the results
  \[ P(x_1 \ldots x_5) = P(x_1) \cdot P(x_2) \cdot P(x_3|x_1x_2) \cdot P(x_4|x_3) \cdot P(x_5|x_3) \]
  \[ P(x_1|x_4x_5) = \alpha P(x_1) \cdot \sum_{x_2} P(x_2) \cdot \sum_{x_3} P(x_3|x_1x_2) \cdot P(x_4|x_3) \cdot P(x_5|x_3) \]
- Eliminate variables: if not an ancestor of a query variable or evidence variables irrelevant to the query
- \( O(n) \) in the number of variables. Dominated by the largest intermediate factor.
Message-passing algorithms

- Message-passing algorithms: $O(n)$ for complex graphs and...
- ... generalize many known algorithms
- Based on Dynamic Programming algorithms: save and store intermediate results, by mainly ordering computations

**Examples:**
- a Kalman filter as a Bayesian network. MP does exactly the same computations
- Viterbi decoding (most probable sequence)
- FFT, CSP, Naive Bayes
Other methods

- Solve the same problems but are generally faster
- **Sampling:**
  - Rejection, Likelihood weighting, Gibbs, MCMC
  - Generate random events following down the structure of the Bayesian Networks
  - not always faster in fact!
- **Variational methods:**
  - reduced version of the original model
  - introduce a variational parameter $\lambda$
  - adjust it to minimize a distance $D$ between the original model and the reduced one (expl: $\partial D / \partial \lambda$)
  - compute lower and upper bounds on the posterior
  - convex optimization problem, most of the time
Machine Learning

- Learning the parameters
  - fitting the parameters
  - Maximum Likelihood or E.M. algorithm when data are missing

- Learning the graph structure
  - find causality
  - find relationships with a predefined structure

- Learning the probabilistic form
  - from a family
  - the Bayesian way: probability distributions over probability distributions with a Dirichlet process
Machine Learning II

- Aims at finding parameters, relationships, distributions
- Another important aspect:
  - explaining hidden structure from the data
  - automatically discovering of patterns and dependencies between variables
- Efficient construction of hybrid models
  - can join in an optimal way several time series models
  - can induce sparsity in legacy models like multivariate ARMA or state-space models
  - can find causality in multivariate time series
Proabilistic reasoning over (time series)

- Agent keeps track of the current state of the environment
- Hidden Markov Models, Kalman filters $\rightarrow$ Dynamic Bayesian Networks
- **Assumptions**: stationary process (at least weakly), Markov assumption
- A simple model:
  \[
  P(X_0X_1 \ldots X_nE_0 \ldots E_n) = P(X_0) \cdot \prod P(X_i|X_{i-1})P(E_i|X_i)
  \]
  - if $X_i$ and $E_i$ are discrete variables: it’s a Hidden Markov Model (without observations it’s a Markov chain)
  - if we assume linear Gaussian distribution, it’s a Kalman filter

![Diagram](image-url)
Dynamic Bayesian Networks

- **Discrete case:** HMM and DBN are the same but DBN takes advantage of sparsity
- **DBN** generalize other models, relaxing several assumptions:
  - Markov order at $n \geq 1$,
  - non-linear models and not always gaussian
- **Inference:**
  - DBN are Bayesian networks $\Rightarrow$ all algorithms are applicable
  - Exact inference solve many problems at once: *smoothing* (past), *filtering or monitoring* (present), *prediction* (future)
- **Approximate inference** can be used too, and are important
  - the prior is factorable but not the posterior for exact computations
Building new models I
Building new models II
Other examples: a simple one

- Posterior distribution of the mean of a Gaussian variable
- Example of i.i.d data seen from a graphical model's point of view
- $P(D|\mu) = \prod_{n=1}^{N} p(x_n|\mu)$: it means that $\mu$ generates the data
Other examples: a complex one
Conclusion

*The decomposition of large probabilistic domains into weakly connected subsets via conditional independence is one of the most important development in the recent history of Artificial Intelligence.* (S. Russell, P. Norvig)