1 Graphical models

1.1 Causal models

In the previously discussed regression example we mainly considered a model for one random variable or at most two. We now consider more complex models with many more factors described by random variables. Probability theory nicely generalizes to multiple random variables, and such multivariate cases are described by a joint probability as outlined in the review of probability theory. An example from Sebastian Thrun of a model to diagnose when a car is not starting is shown in Figure 1.1.

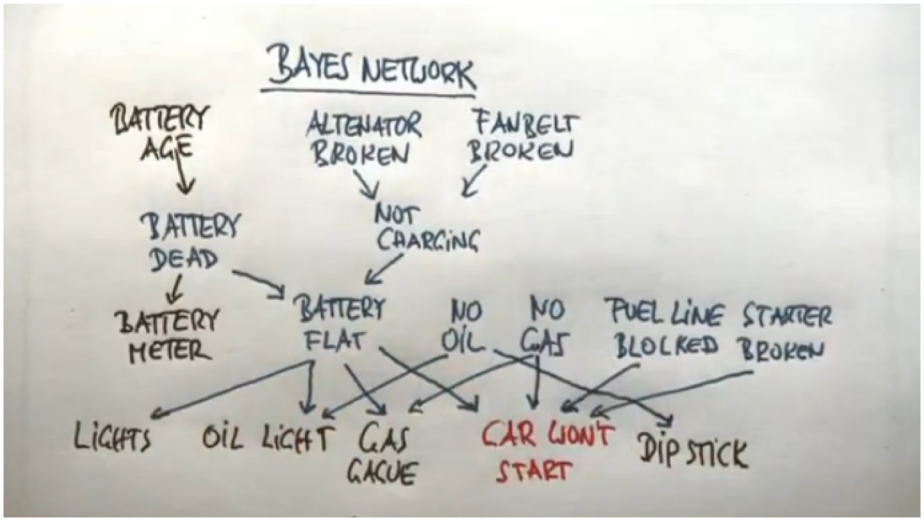

Fig. 1.1 Example of causal model.

This example is reasonably sized although real world problems would be likely be even larger than this. This model considered 16 random numbers to determine possible causes if the car does not start (the variable "car does not start is not a random number as we are using this model when we know already that the car is not starting). The random numbers themselves could have two possible outcomes (like if there is gas in the tank) or even multiple possible values (like the age of the battery). At this time let us simplify the model with only considering binary values. That is, the age of the battery would only be specified as new or old. The joined probability table for the 16 variables would then have $2^16 - 1 = 65535$ entries. These parameters have to be estimated (learned) from examples using MAP or MLE.

In addition to the shear explosion of parameters with increasing model complexity,

there is another reason why the joined probability function is also not exactly what we need to know. The joint density functions of multiple variables describe the co-occurrence of specific values of the random variables. Indeed, the joint probability function $p(X, Y)$ is symmetric in its arguments,

$$p(X, Y) = p(Y, X), \tag{1.1}$$

What we really want to do is to a model to reason about the world, or specifically, to reason about possible events. For this we want to add knowledge or hypotheses about **causal relations**. For example, a fire alarm should be triggered by a fire, although there is some small chance that the alarm will not sound when the unit is defect. However, it is (hopefully) unlikely that the sound of a fire alarm will trigger a fire. It is useful to illustrate such casual relations with graphs such as

In such **graphical models**, the nodes represent random variables, and the links between them represent causal relations with conditional probabilities, $p(A|F)$. Since we use arrows on the links we are discussing here **directed graphs**, and we are also restricting our discussions here to graphs that have no loops, so called **acyclic graphs**. **Directed acyclic graphs** are also called **DAG**s.

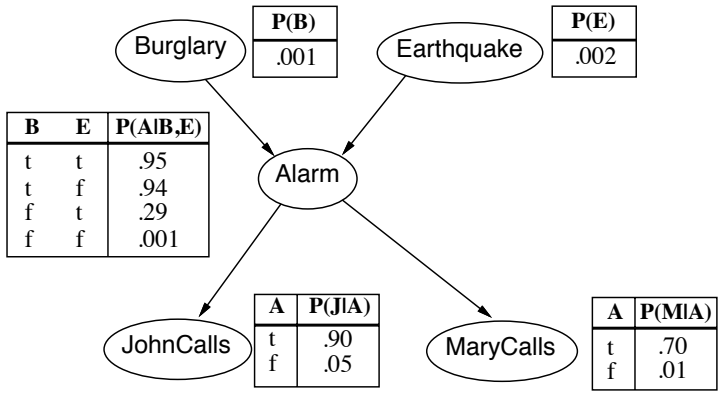

Fig. 1.2 Example of causal model.

Graphical causal models have been advanced largely by Judea Pearl, and the following example is taken from his book[1]. The model is shown in Figure 1.2. Each of the five nodes stands for a random binary variable (Burglary B={yes,no}, Earthquake E={yes,no}, Alarm A={yes,no}, JohnCalls J={yes,no}, MaryCalls M={yes,no}) The figure also include **conditional probability tables (CPTs)** that specify the conditional probabilities represented by the links between the nodes.

The joined distribution of the five variables can be factories in various ways following the chain rule mentioned before (equations **??**), for example as

[1] Judea Pearl, 'Causality: Models, Reasoning and Inference', Cambridge University Press 2000, 2009'.

$$p(B, E, A, J, M) = P(B|E, A, J, M)P(E|A, J, M)P(A|J, M)P(J|M)P(M) \tag{1.2}$$

However, the the causal model represents a specific factorization of the joint probability functions, namely

$$p(B, E, A, J, M) = P(B)P(E)P(A|B, E)P(J|A)P(M|A), \tag{1.3}$$

which is much easier to handle. For example, if we do not know the conditional probability functions, we need to run many more experiments to estimate the various conditions ($2^4 + 2^3 + 2^2 + 2^1 + 2^0 = 31$) instead of the reduced conditions in the causal model ($1 + 1 + 2^2 + 2 + 2 = 10$). It is also easy to use the casual model to do inference (drawing conclusions), for specific questions. For example, say we want to know the probability that there was no earthquake or burglary when the alarm rings and both John and Mary call. This is given by

$$\begin{aligned}
&P(B = f, E = f, A = t, J = t, M = t) = \\
&= P(B = f)P(E = f,)P(A = t|B = f, E = f)P(J = t|A = t)P(M = t|A = t) \\
&= 0.998 * 0.999 * 0.001 * 0.7 * 0.9 \\
&= 0.00062
\end{aligned}$$

Although we have a casual model where parents variables influence the outcome of child variables, we can also use a child evidence to infer some possible values of parent variables. For example, let us calculate the probability that the alarm rings given that John calls, $P(A = t|J = t)$. For this we should first calculate the probability that the alarm rings as we need this later. This is given by

$$\begin{aligned}
P(A = t) = &P(A = t|B = t, E = t)P(B = t)P(E = t) + ... \\
&P(A = t|B = t, E = f)P(B = t)P(E = f) + ... \\
&P(A = t|B = f, E = t)P(B = f)P(E = t) + ... \\
&P(A = t|B = f, E = f)P(B = f)P(E = f) \\
= &0.95 * 0.001 * 0.002 + 0.94 * 0.001 * 0.998 + ... \\
&0.29 * 0.999 * 0.002 + 0.001 * 0.999 * 0.998 \\
= &0.0025
\end{aligned}$$

We can then use Bayes' rule to calculate the required probability,

$$\begin{aligned}
P(A = t|J = t) &= \frac{P(J = t|A = t)P(A = t)}{P(J = t|A = t)P(A = t) + P(J = t|A = f)P(A = f)} \\
&= \frac{0.90.0025}{0.90.0025 + 0.050.9975} \\
&= 0.0434
\end{aligned}$$

We can similarly apply the rules of probability theory to calculate other quantities, but these calculations can get cumbersome with larger graphs. It is therefore useful to use numerical tools to perform such inference. A Matlab toolbox for Bayesian networks is introduced in the next section.

While inference is an important application of causal models, inferring causality from data is another area where causal models revolutionize scientific investigations. Many traditional methods evaluate co-occurrences of events to determine dependencies, such as a correlation analysis. However, such a correlation analysis is usually not a good indication of causality. Consider the example above. When the alarm rings it is likely that John and Mary call, but the event that John calls is mutually independent of the event that Mary calls. Yet, when John calls it is also statistically more likely to observe the event that Mary calls. Sometimes we might just be interested in knowing about the likelihood of co-occurrence, for which a correlation analysis can be a good start, but if we are interested in describing the causes of the observations, then we need another approach. Some algorithms have been proposed for **structural learning**, such as an algorithm called **inferred causation (IC)**, which deduces the most likely causal structure behind given data is.

1.2 Bayes Net toolbox

An Matlab implementation of various algorithms for inference, parameters estimation and inferred causation is provided by Kevin Murphy in the Bayes Net toolbox[2]. We will demonstrate some of its features on the burglary/earthquake example above.

The first step is to create a graph structure for the DAG We have five nodes. The nodes are given numbers, but we also use variables with capital letter names to refer to them. The DAG is then a matrix with entries 1 where directed links exist.

```
N=5;% number of nodes
B=1; E=2; A=3; J=4; M=5;
dag = zeros(N,N);
dag(B,A)=1;
dag(E,A)=1;
dag(A,[J M])=1;
```

The nodes represent discrete random variables with two possible state. We only discuss here discrete random variables, although the toolbox contains methods for continuous random variables. For the discrete case we have to specify the number of possible states of each variable, and we can then create the corresponding Bayesian network,

```
% Make bayesian network
node_sizes=[2 2 2 2 2];  %binary nodes
bnet=mk_bnet(dag,node_sizes); %make bayesian net
```

The next step is to provide the numbers for the conditional probability distributions, which are the conditional probability tables for discrete variables. For this we provide the numbers in a vector according to the following convention. Say we specify the probabilities for node 3, which is conditionally dependent on nodes 1 and 2. We then provide the probabilities in the following order:

[2] The toolbox can be downloaded at http://code.google.com/p/bnt.

Node 1	Node 2	P(Node 3=X)
F	F	F
T	F	F
F	T	F
T	T	F
F	F	T
T	F	T
F	T	T
T	T	T

For our specific examples, the CPT are thus specified as

```
bnet.CPD{B} = tabular_CPD(bnet,B,[0.999 0.001]);
bnet.CPD{E} = tabular_CPD(bnet,E,[0.998 0.002]);
bnet.CPD{A} = tabular_CPD(bnet,A,[0.999 0.06 0.71 0.05 0.001 0.94 0.29 0.95]);
bnet.CPD{J} = tabular_CPD(bnet,J,[0.95 0.10 0.05 0.90]);
bnet.CPD{M} = tabular_CPD(bnet,M,[0.99 0.30 0.01 0.70]);
```

We are now ready to calculate some inference. For this we need to specify a specific inference engine. There are several algorithms implemented, a variety of exact algorithm as well as approximate procedures in case the complexity of the problem is too large. Here we use the basic exact inference engine, the **junction tree algorithm**, which is based on a message passing system.

```
engine=jtree_inf_engine(bnet);
```

While this is an exact inference engine, there are other engines, such as approximate engines, that might be employed for large graphs when other methods fail.

As an example of an inference we recalculate the example above, that of calculate the probability that the alarm rings given that John calls, $P(A = t|J = t)$. For this we have to enter some evidence, namely that $J = t$, into a cell array and add this to the inference engine,

```
evidence=cell(1,N);
evidence{J}=2;
[engine,loglik]=enter_evidence(engine,evidence)
```

We can then calculate the marginal distribution for a variable, given the evidence as,

```
marg=marginal_nodes(engine,A)
p=marg.T(2)
```

It is now very easy to calculate other probabilities.

The Bayesnet toolbox also includes routines to handle some continuous models such as models with Gaussian nodes. In addition, there are routines to do parameter estimation, including point estimates, such maximum likelihood estimation, and also full bayesian priors. Finally, he toolbox includes routines to inferred causation through structural learning.