Today – Bayesian networks

• One of the most exciting advancements in statistical AI and machine learning in the last 10-15 years
• Generalizes naïve Bayes and logistic regression classifiers
• Compact representations of exponentially-large probability distributions
• Exploit conditional independences

Judea Pearl: Turing award in 2011 for his contributions to Bayesian networks
CAUSAL STRUCTURE

• Draw a directed acyclic (causal) graph for the following
  • Direct arrows from cause to effect

• Story:
  • There is a Burglar alarm that rings when we have Burglary
  • However, sometimes it may ring because of winds that exceed 60mph
  • When the alarm rings your neighbor Mary Calls
  • When the alarm rings your neighbor John Calls
There is a Burglar alarm that rings when we have Burglary.

However, sometimes it may ring because of winds that exceed 60mph.

When the alarm rings your neighbor Mary Calls.

When the alarm rings your neighbor John Calls.
Representation of Joint Distribution

\[ P(W, B, A, J, M) = P(W)P(B)P(A|B, W)P(J|A)P(M|A) \]

In general:

\[ P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | pa(x_i)) \]
Possible Queries

- Inference
  - $P(W=?|J=True)$
- Most probable explanation
  - Assignment of values to all other variables that has the highest probability given that $J=True$ and $M=False$
- Maximum Aposteriori Hypothesis.
Number of Parameters

- Assume Binary variables
- How many parameters with each CPT $P(x_i | x_1, ..., x_j)$?
  - $(\text{Domain-size of } x_i - 1)^* (\text{All possible combinations of } x_1, ..., x_j)$.

```
```
Can any distribution be represented as a Bayesian network

Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be a set of discrete random variables such that each variable $X_i$ takes values from a finite domain $D(X_i)$. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be an assignment to all variables $\mathbf{X}$ such that each variable $X_i$ is assigned the value $x_i$ from its domain $D(X_i)$. Let $(X_1, \ldots, X_n)$ be any ordering of the variables, then:

$$
\Pr(\mathbf{x}) = P(x_1) \prod_{i=2}^{n} P(x_i|x_1, \ldots, x_{i-1})
$$

Let $pa(X_i)$ be a subset of variables of the set $\{X_1, \ldots, X_{i-1}\}$ such that $X_i$ is conditionally independent of all other variables ordered before it given $pa(X_i)$. Then, we can rewrite the joint probability distribution as:

$$
\Pr(\mathbf{x}) = \prod_{i=1}^{n} P(x_i|\pi(\mathbf{x}, pa(X_i))) \tag{1}
$$

where $\pi(\mathbf{x}, pa(X_i))$ is the projection of $\mathbf{x}$ on the parents of $X_i$. (For instance, the projection of the assignment $(X_1 = 0, X_2 = 0, X_3 = 1)$ on $\{X_1, X_3\}$ is the assignment $(X_1 = 0, X_3 = 1)$).
Graph represents conditional independence assumptions!!

- Each variable is conditionally independent of its non-descendants given its parents, called **Markov assumptions**.
  - \( I(X, \text{parents}(X), \text{non-descendants}(X)) \)

\[
\begin{align*}
\text{Winds} & \quad P(W) \quad P(B) \\
\text{Burglary} & \\
\text{Alarm} & \quad P(A|B,W) \\
\text{John Calls} & \quad P(J|A) \\
\text{Mary Calls} & \quad P(M|A)
\end{align*}
\]
Properties of conditional Independence

- **Symmetry** \( I(X, Z, Y) \Rightarrow I(Y, Z, X) \)
- **Decomposition** \( I(X, Z, Y \cup W) \Rightarrow I(X, Z, Y) \)
- **Weak Union** \( I(X, Z, Y \cup W) \Rightarrow I(X, Z \cup W, Y) \)
- **Contraction** \( I(X, Z \cup Y, W) \& I(X, Z, Y) \Rightarrow I(X, Z, Y \cup W) \)
- **Intersection** For any positive distribution:
  \( I(X, Z \cup W, Y) \& I(X, Z \cup Y, W) \Rightarrow I(X, Z, Y \cup W) \)
Proof of Symmetry

Assume that $I(X, Z, Y)$ holds. This implies that:

$$\Pr(X, Y|Z) = \Pr(X|Z) \times \Pr(Y|Z)$$

(1)

i.e. $I(Y, Z, X)$ holds too (exchanging the positions of $X$ and $Y$).
Proof of Decomposition

Assume that $I(X, Z, Y \cup W)$ holds. Then,

$$\Pr(X, Y, W|Z) = \Pr(X|Z) \times \Pr(Y, W|Z)$$

$$\Pr(X, Y|Z) = \sum_w \Pr(X, Y, w|Z) \quad (2)$$

$$= \sum_w \Pr(X|Z) \times \Pr(Y, w|Z) \quad (3)$$

$$= \Pr(X|Z) \sum_w \Pr(Y, w|Z) \quad (4)$$

$$= \Pr(X|Z) \Pr(Y|Z) \quad (5)$$

i.e. $I(X, Z, Y)$ holds too.
Implication of conditional Independence properties

- Start with conditional independence statements based on the Markov assumptions and derive new properties.

- For example
  - $I(X, \text{parents}(X), \text{non-descendants}(X))$ implies that
    - $I(\text{non-descendants}(X), \text{parents}(X), X)$
    - Symmetry
  - $I(X, \text{parents}(X), Y)$ where $Y$ is a subset of the non-descendants of $X$.
    - Decomposition

- Symmetry $I(X, Z, Y) \Rightarrow I(Y, Z, X)$
- Decomposition $I(X, Z, Y \cup W) \Rightarrow I(X, Z, Y)$
- Weak Union $I(X, Z, Y \cup W) \Rightarrow I(X, Z \cup W, Y)$
- Contraction $I(X, Z \cup Y, W) \& I(X, Z, Y) \Rightarrow I(X, Z, Y \cup W)$
- Intersection For any positive distribution:
  $I(X, Z \cup W, Y) \& I(X, Z \cup Y, W) \Rightarrow I(X, Z, Y \cup W)$
Implication of conditional Independence properties (continued)

- Given a Bayesian network if I ask you a question, is $W$ independent of $M$ given $A$?
- Derive all possible Conditional independence statements using the five properties and Markov assumptions and check if the above statement is in it.
Graphical Test of Conditional Independence Properties: D-separation

• D-sep($X,Y,Z$) only if $I(X,Z,Y)$

• To decide whether $D$-Sep($X,Z,Y$)
  • Check whether $X$ and $Y$ are disconnected in a new DAG $G'$ obtained from $G$ using the following steps:
    • We delete any leaf node $W$ from DAG $G$ as long as $W$ is not in $X$, or $Y$ or $Z$. This process is repeated until no more nodes can be deleted.
    • Delete all edges outgoing from nodes in $Z$

A more complicated algorithm for d-separation is given in your book.
D-Separation

Nodes in Z are shaded. Pruned nodes and edges are dotted.

Is $X = \{A, S\}$ d-separated from $Y = \{D, X\}$ by $Z = \{B, P\}$?
D-separation

Nodes in $Z$ are shaded. Pruned nodes and edges are dotted.

Is $X = \{T, C\}$ d-separated from $Y = \{B\}$ by $Z = \{S, X\}$?
Recap: What is a Bayesian network?

- A compact representation of the joint distribution
- A compact representation of a set of conditional independence assumption.

  - A directed acyclic graph (a Bayesian network) describes a set of conditional independence assumptions $I_G$.
  - It is an I-map of a distribution $I_{Pr}$ if $I_G \Rightarrow I_{Pr}$
  - It is a D-map if $I_{Pr} \Rightarrow I_G$
  - It is a P-map if it is both an I-map and a D-map.
  - I-maps and D-maps can be constructed trivially. Therefore, we enforce minimality.
Relationship to Naïve Bayes

- Distribution?

Number of parameters?
Real Bayesian networks applications

• Diagnosis of lymph node disease
• Speech recognition
• Microsoft office and Windows
  • http://www.research.microsoft.com/research/dtg/
• Study Human genome
• Robot mapping
• Robots to identify meteorites to study
• Modeling fMRI data
• Anomaly detection
• Fault diagnosis
• Modeling sensor network data
Now What?

• Given a Bayesian network, answer queries
  • Inference

• Learning Bayesian networks from Data
  • Structure and **weight learning**
  • Partially and fully observed data
  • **MLE** and Bayesian approach
  • Requires Inference, i.e., computing \( P(X|\text{evidence}) \) where evidence is an assignment to a subset of variables

• Unfortunately, Inference is NP-hard.
  • Tractable classes based on treewidth
  • Approximate Inference approaches