Graphical Models

Key Idea:
- Conditional independence assumptions useful
  - but Naïve Bayes is extreme!
- Graphical models express sets of conditional independence assumptions via graph structure
- Graph structure plus associated parameters define joint probability distribution over set of variables

Two types of graphical models:
- Directed graphs (aka Bayesian Networks)
- Undirected graphs (aka Markov Random Fields)
Graphical Models – Why Care?

• Among most important ML developments of the decade

• Graphical models allow combining:
  – Prior knowledge in form of dependencies/independencies
  – Prior knowledge in form of priors over parameters
  – Observed training data

• Principled and ~general methods for
  – Probabilistic inference
  – Learning

• Useful in practice
  – Diagnosis, help systems, text analysis, time series models, ...

Conditional Independence

*Definition:* X is conditionally independent of Y given Z, if the
probability distribution governing X is independent of the value
of Y, given the value of Z

\[(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)\]

Which we often write \( P(X|Y, Z) = P(X|Z) \)

E.g. \( P(\text{Thunder}|\text{Rain}, \text{Lightning}) = P(\text{Thunder}|\text{Lightning}) \)
Marginal Independence

Definition: X is marginally independent of Y if

\[(\forall i, j)P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)\]

Equivalently, if

\[(\forall i, j)P(X = x_i|Y = y_j) = P(X = x_i)\]

Equivalently, if

\[(\forall i, j)P(Y = y_i|X = x_j) = P(Y = y_i)\]

Represent Joint Probability Distribution over Variables

- Visit to Asia \(X_1\)
- Smoking \(X_2\)
- Tuberculosis \(X_3\)
- Lung Cancer \(X_4\)
- Bronchitis \(X_5\)
- Tuberculosis or Cancer \(X_6\)
- XRay Result \(X_7\)
- Dyspnea \(X_8\)
Describe network of dependencies

Bayes Nets define Joint Probability Distribution in terms of this graph, plus parameters

Benefits of Bayes Nets:
• Represent the full joint distribution in fewer parameters, using prior knowledge about dependencies
• Algorithms for inference and learning
Bayesian Networks Definition

A Bayes network represents the joint probability distribution over a collection of random variables.

A Bayes network is a directed acyclic graph and a set of conditional probability distributions (CPD's):
• Each node denotes a random variable
• Edges denote dependencies
• For each node $X_i$ its CPD defines $P(X_i | Pa(X_i))$
• The joint distribution over all variables is defined to be
  
  $$P(X_1 \ldots X_n) = \prod_{i} P(X_i | Pa(X_i))$$

$Pa(X) = \text{immediate parents of } X \text{ in the graph}$

Bayesian Network

Nodes = random variables

A conditional probability distribution (CPD) is associated with each node $N$, defining $P(N | Parents(N))$

| Parents | $P(W|Pa)$ | $P(\neg W|Pa)$ |
|---------|-----------|----------------|
| L, R    | 0         | 1.0            |
| L, $\neg$R | 0   | 1.0            |
| $\neg$L, R | 0.2 | 0.8            |
| $\neg$L, $\neg$R | 0.9 | 0.1            |

The joint distribution over all variables:

$$P(X_1 \ldots X_n) = \prod_{i} P(X_i | Pa(X_i))$$
What can we say about conditional independencies in a Bayes Net?

One thing is this:
Each node is conditionally independent of its non-descendents, given only its immediate parents.

| Parents | P(W|Pa) | P(¬W|Pa) |
|---------|--------|----------|
| L, R    | 0      | 1.0      |
| L, ¬R   | 0      | 1.0      |
| ¬L, R   | 0.2    | 0.8      |
| ¬L, ¬R  | 0.9    | 0.1      |

Some helpful terminology

- **Parents** = Pa(X) = immediate parents
- **Antecedents** = parents, parents of parents, ...
- **Children** = immediate children
- **Descendents** = children, children of children, ...

| Parents | P(W|Pa) | P(¬W|Pa) |
|---------|--------|----------|
| L, R    | 0      | 1.0      |
| L, ¬R   | 0      | 1.0      |
| ¬L, R   | 0.2    | 0.8      |
| ¬L, ¬R  | 0.9    | 0.1      |
Bayesian Networks

- CPD for each node $X_i$ describes $P(X_i | Pa(X_i))$

Chain rule of probability says that in general:

But in a Bayes net:
$$P(X_1 \ldots X_n) = \prod_i P(X_i|Pa(X_i))$$

---

How Many Parameters?

To define joint distribution in general?

To define joint distribution for this Bayes Net?
Inference in Bayes Nets

\[
P(S=1, L=0, R=1, T=0, W=1) =
\]

Learning a Bayes Net

\[
\text{Consider learning when graph structure is given, and data } = \{ <s,l,r,t,w> \}
\]

What is the MLE solution? MAP?
**Algorithm for Constructing Bayes Network**

- Choose an ordering over variables, e.g., $X_1, X_2, \ldots, X_n$
- For $i=1$ to $n$
  - Add $X_i$ to the network
  - Select parents $Pa(X_i)$ as minimal subset of $X_1, \ldots, X_{i-1}$ such that
    \[
P(X_i|Pa(X_i)) = P(X_i|X_1, \ldots, X_{i-1})\]

Notice this choice of parents assures
\[
P(X_1 \ldots X_n) = \prod_{i} P(X_i|X_1 \ldots X_{i-1}) \quad \text{(by chain rule)}
\]
\[
= \prod_{i} P(X_i|Pa(X_i)) \quad \text{(by construction)}
\]

**Example**

- Bird flu and Allergies both cause Nasal problems
- Nasal problems cause Sneeze and Headaches
What is the Bayes Network for X1,...,X4 with no assumed conditional independencies?

What is the Bayes Network for Naïve Bayes?
What do we do if variables are mix of discrete and real valued?

Bayes Network for a Hidden Markov Model

Implies the future is conditionally independent of the past, given the present

$P(S_{t-2}, O_{t-2}, S_{t-1}, \ldots, O_{t+2}) =$
What You Should Know

• Bayes nets are convenient representation for encoding dependencies / conditional independence
• BN = Graph plus parameters of CPD’s
  – Defines joint distribution over variables
  – Can calculate everything else from that
  – Though inference may be intractable
• Reading conditional independence relations from the graph
  – Each node is cond indep of non-descendents, given only its parents
  – ‘Explaining away’

See Bayes Net applet: http://www.cs.cmu.edu/~javabayes/Home/applet.html

Inference in Bayes Nets

• In general, intractable (NP-complete)
• For certain cases, tractable
  – Assigning probability to fully observed set of variables
  – Or if just one variable unobserved
  – Or for singly connected graphs (ie., no undirected loops)
    • Belief propagation
• For multiply connected graphs
  • Junction tree
• Sometimes use Monte Carlo methods
  – Generate many samples according to the Bayes Net distribution, then count up the results
• Variational methods for tractable approximate solutions
Example

- Bird flu and Allegies both cause Sinus problems
- Sinus problems cause Headaches and runny Nose

Prob. of joint assignment: easy

- Suppose we are interested in joint assignment $<F=f,A=a,S=s,H=h,N=n>$

What is $P(f,a,s,h,n)$?

let's use $p(a,b)$ as shorthand for $p(A=a, B=b)$
Prob. of marginals: not so easy

- How do we calculate $P(N=n)$?

Let's use $p(a,b)$ as shorthand for $p(A=a, B=b)$.

Generating a sample from joint distribution: easy

How can we generate random samples drawn according to $P(F, A, S, H, N)$?

Let's use $p(a,b)$ as shorthand for $p(A=a, B=b)$. 
Generating a sample from joint distribution: easy

Note we can estimate marginals like \( P(N=n) \) by generating many samples from joint distribution, then count the fraction of samples for which \( N=n \).

Similarly, for anything else we care about \( P(F=1|H=1, N=0) \)

\( \overset{\rightarrow}{\rightarrow} \) weak but general method for estimating any probability term…

let's use \( p(a,b) \) as shorthand for \( p(A=a, B=b) \)

---

Prob. of marginals: not so easy

But sometimes the structure of the network allows us to be clever \( \overset{\rightarrow}{\rightarrow} \) avoid exponential work

eg., chain

A B C D E
Inference in Bayes Nets

- In general, intractable (NP-complete)
- For certain cases, tractable
  - Assigning probability to fully observed set of variables
  - Or if just one variable unobserved
  - Or for singly connected graphs (i.e., no undirected loops)
    - Variable elimination
    - Belief propagation
- For multiply connected graphs
  - Junction tree
- Sometimes use Monte Carlo methods
  - Generate many samples according to the Bayes Net distribution, then count up the results
- Variational methods for tractable approximate solutions