Scribe 3

Yaw Nti-Addae (yntiadda@andrew.cmu.edu)

01/21/2009

Topics:

1. Bayesian Network
2. Parameter estimation
3. Expectation maximization algorithm

Bayesian Networks

Bayesian network is a directed acyclic graph (Slide 10)

.

Example: Bayesian network for student example
I - intelligence, $val(I) \leftarrow \{i^0, i^1\}$
G - grades, $val(G) \leftarrow \{g^a, g^b, g^c\}$
S - SAT, $val(S) \leftarrow \{s^0, s^1\}$
L - letter, $val(L) \leftarrow \{l^0, l^1\}$

.

Refer to slide 11 of handout for a diagram of this network.

.

Conditional Independence Statements (Assumptions)

1. $(L \perp D, I, S \mid G)$
2. $(S \perp D, G, L \mid I)$
3. $(G \perp S \mid D, I)$
4. $(D \perp S)$
5. $(D \perp I)$

.

$(G \perp S, L \mid D, I)$ is false as $P(g^a \mid i^1, d^1, l^1) > P(g^a \mid i^1, d^1)$
Given parents of x_i, it is false to say that x_i is not dependent on its children.

.

Representation of Joint Distribution
$P(I, D, G, L, S) = P(I) * P(D \mid I) * P(G \mid I, D) * P(L \mid I, G, D) * P(S \mid I, D, G, L)$
This is too complex as we need more specification.
But we cas reduce its complexity using our assumptions. Hence,
$P(L \mid I, D, G) = P(L \mid G)$
$P(S \mid I, D, G, L) = P(S \mid I)$
$P(D \mid I) = P(D)$

.

Note that the Joint Distribution is based on order. It should be from parents to children.

.

.

Formal description of a BN structure
Refer to slide 14 of handout.

.

Local markov assumption: $(X_i \perp Nondependents \mid PaX_i)$

.
Going back to student example.
Defining the BN (slide 15).
$P(X_1 ... X_n)$ satisfies the condition.
If $P(X_i, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid PaX_i)$, then P factorizes according to G and $P(X_i, ..., X_n)$ is called the **Chain Rule** for BNs and $X_i, ..., X_n$ is called **Conditional Probability Distribution** or **Local Probability Models**

.
* BN is a pair of (G, P) where

1. P factorizes over G
2. $P(X_1, ..., X_n)$ should satisfy out joint distribution conditions.

.

Example of use of BN in Biological situation (Slide 16)
Determination of blood type.
$g_{blood} \in \{A, B, O\} \leftarrow alleles$
.
We will define our variables as:
$G - genotype \in \{AA, AB, ...\}$
$B - phenotype \in \{A, N, O\}$
.
Assumptions

1. B_i has only one parent G_i, and give that parent G_i, B_i is independent on all other nodes.
2. G_i has two parents G_{father} and G_{mother}, and giving these two parents, G_i is independent of all other nodes.

.

Refer to slide 16 for graphs.

Given this graph, we can give it any query such as
$P(B_{all}, G_{all}) = P(G_{Harry}) * P(G_{Jack}) * P(B_{Harry} \mid G_{Jack})...$
A more difficult one is
$P(B_{Lisa}, B_{parent}) = P(B_{parent}) * P(B_{Lisa} \mid B_{parent})$, this is know as **Inferencing**, and we have to go through variable elimination.

Note that:
$P(A, B, C) = P(A, B, C = c^0) + P(A, B, C = c^1) = \sum_c P(A, B, C = c)$

Parameter Estimation

Learning parameters in a BN could be done in 2 ways:

1. ask experts
2. learn from data

.

Data $D = \{d[1], ..., d[n]\}$, where $d[i] = \begin{bmatrix} d_{i,1} \\ . \\ . \\ . \\ d_{i,n} \end{bmatrix}$

Coin toss
X, $val(X) = \{head, tail\}$
$P(X = head) = \theta_h$
$P(X = tail) = \theta_t$
$\theta_h + \theta_t = 1$
.
Given data $d[1]...d[m]$, such as $< htthhtttttth >$, $\theta_h = \frac{M_h}{M}$, and $\theta_t = \frac{M_t}{M}$

But we dont need two parameters since,
$P(X = head) = \theta$, and $P(X = tail) = 1 - \theta$.
Hence, $P(D : \theta) = \theta^{M_h} * (1 - \theta)^{M_t}$
.
.
<u>Likelihood Function</u>
$L(\theta : D) = P(D : \theta)$
.
Learning θ is called **maximum likelihood estimation** (MLE)
$\theta^* = argmax_\theta L(\theta : D)$
and given the data above, $L(\theta : D) = \frac{M_h}{M_h + M_t}$
.
Proof:

$$\begin{aligned}
L(\theta : D) &= \theta^{M_h}(1-\theta)^{M_t} \\
l(\theta) &= log(\theta^{M_h}(1-\theta)^{M_t} \\
&= M_h log(\theta) + M_t log(1-\theta) \\
\frac{\delta}{\delta\theta}\mid_{\theta^*} &= M_h \frac{1}{\theta^*} + M_t \frac{1}{1-\theta^*} = 0 \\
\theta^* &= \frac{M_h}{M}
\end{aligned}$$