CSC412/2506 Winter 2020: Probabilistic Learning and Reasoning

Tianyu Du

February 8, 2020

Contents

1 Introduction

2 Probabilistic Models

3 Directed Graphical Models

3.1 Decision Theory

4 Exact Inference

Notation 4.1. Let X denote the set of all random variables in the model, and

1. $X_E =$ The observed evidence;
2. $X_F =$ The unobserved variable we want to infer;
3. $X_R = X - \{X_F, X_E\} =$ Remaining variables, extraneous to query.

The model defines the joint distribution of all random variables:

$$p(X_E, X_F, X_R) \tag{4.1}$$

Definition 4.1. The joint distribution over evidence and subject of inference is

$$p\left(X_F, X_E\right) = \sum_{X_R} p\left(X_F, X_E, X_R\right) \tag{4.2}$$

Definition 4.2. The conditional probability distribution for inference given evidence is

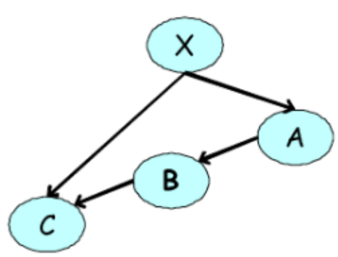
$$p\left(X_F|X_E\right) = \frac{p\left(X_F, X_E\right)}{p\left(X_E\right)} = \frac{p\left(X_F, X_E\right)}{\sum_{X_F} p\left(X_F, X_E\right)} \tag{4.3}$$

Definition 4.3. The distribution of evidence can be computed as

$$p\left(X_E\right) = \sum_{X_F, X_R} p\left(X_F, X_E, X_R\right) \tag{4.4}$$

4.1 Variable Elimination

4.2 Intermediate Factors



$$
\begin{aligned}
p(A,B,C) &= \sum_X p(X)p(A|X)p(B|A)p(C|B,X) && (4.5)\\
&= p(B|A)\underbrace{\sum_X p(X)p(A|X)p(C|B,X)}_{\text{unnormalized}} && (4.6)
\end{aligned}
$$

Definition 4.4. A **factor** ϕ describes the local relation between random variables, meanwhile, $\int d\phi$ is not necessarily one.

Remark 4.1. Let $X_\ell \subseteq X$ be a group of local random variables, then $p(X_\ell)$ is automatically a factor $\phi(X_\ell)$.

$$
\begin{aligned}
p(A,B,C) &= \sum_X \underbrace{p(X)p(A|X)p(B|A)p(C|B,X)}_{\text{from graphical representation}} && (4.7)\\
&= \sum_X \underbrace{\phi(X)\phi(A,X)\phi(A,B)\phi(X,B,C)}_{\text{factor representation}} && (4.8)\\
&= \phi(A,B)\sum_X \phi(X)\phi(A,X)\phi(X,B,C) && (4.9)\\
&= \phi(A,B)\ \underbrace{\tau(A,B,C)}_{\text{another factor}} && (4.10)
\end{aligned}
$$

4.3 Sum-Product Inference

Theorem 4.1. Consider a graphical model with random variables $X = Y \cup Z$. For an random variable Y in a directed or undirected model, $P(Y)$ can be computed using the **sum-product**

$$\tau(Y) = \sum_z \prod_{\phi\in\Phi} \phi(Scope[\phi]\cap Z, Scope[\phi]\cap Y) \tag{4.11}$$

where Φ is a set of factors.

Remark 4.2. For directed models,

$$\Phi = \{\phi_{x_i}\}_{i=1}^N = \{p\left(x_i | \text{ parents } (x_i)\right)\}_{i=1}^N \tag{4.12}$$

4.4 Complexity of Variable Elimination Ordering

Theorem 4.2. The complexity of the variable elimination algorithm is

$$\mathcal{O}(mk^{N_{max}}) \tag{4.13}$$

where

(i) m is the number of initial factors $|\Phi|$;

(ii) k is the number of states each random variable takes, assumed to be equal;

(iii) N_i is the number of random variables within each summation;

(iv) $N_{max} = \max_i N_i$.

5 Message passing, Hidden Markov Models, and Sampling

5.1 Message Passing (Computing All Marginals)

Notation 5.1. Let T denote the set of edges in a tree. For a node i, let $N(i)$ denote the set of its neighbours.

The factor of all random variables can be computed following

$$P\left(X_{1:n}\right)=\frac{1}{Z}\underbrace{\left[\prod_{i=1}^{n}\phi\left(x_i\right)\right]}_{\text{prior factors}}\underbrace{\prod_{(i,j)\in T}\phi_{i,j}\left(x_i,x_j\right)}_{\text{local factors}} \tag{5.1}$$

Definition 5.1. The **message** sent from variable j to $i\in N(j)$ is

$$m_{j\to i}\left(x_i\right)=\sum_{x_j}\left[\phi_j\left(x_j\right)\phi_{ij}\left(x_i,x_j\right)\prod_{k\in N(j)\neq i}m_{k\to j}\left(x_j\right)\right] \tag{5.2}$$

Algorithm 5.1 (Belief Propagation Algorithm)**.** Given a tree, inference on an arbitrary node $p(x_i)$ can be computed following:

1. Choose root r arbitrarily;
2. Pass messages from leaves to r;
3. Pass messages from r to leaves;
4. Compute inference

$$p\left(x_i\right)\propto\phi_i\left(x_i\right)\prod_{j\in N(i)}m_{j\to i}\left(x_i\right) \tag{5.3}$$

5.2 Markov Chains

Using chain rule of probability:

$$p\left(x_{1:T}\right)=\prod_{t=1}^{T}p\left(x_t|x_{t-1},\ldots,x_1\right) \tag{5.4}$$

Definition 5.2. A Markov chain is said to be **first-order** if

$$p\left(x_t|x_{1:t-1}\right)=p\left(x_t|x_{t-1}\right) \tag{5.5}$$

Simplification Therefore, for all first-order Markov chains, the full joint distribution can be reduced to

$$p\left(x_{1:T}\right)=\prod_{t=1}^{T}p\left(x_t|x_{t-1}\right) \tag{5.6}$$

Definition 5.3. A Markov chain is at m-order if

$$p\left(x_t|x_{1:t-1}\right)=p\left(x_t|x_{t-m:t-1}\right) \tag{5.7}$$

Definition 5.4. A Markov chain is said to be **homogenous** (i.e., stationary) if

$$p\left(x_t|x_{t-1}\right) = p\left(x_{t+k}|x_{t-1+k}\right) \quad \forall t, k \tag{5.8}$$

Parameterization Assume the random variable X_t takes k states, further suppose the chain is time homogenous. Then characterizing the transition probability

$$p(x_t|x_{t-1}, x_{t-2}, \cdots, x_{t-m}) \tag{5.9}$$

requires $(k-1)k^m$ parameters.

5.3 Hidden Markov Models

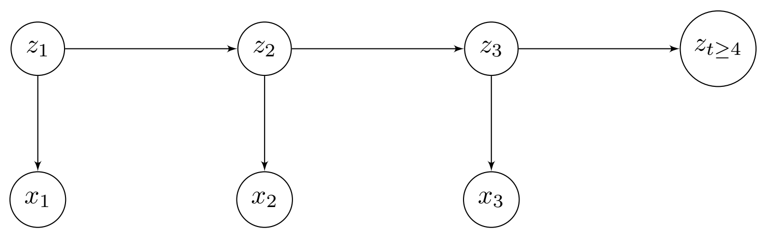


Figure 5.1: Hidden Markov Model

Parameterization Assuming the HMM is homogenous, then the set of parameters Φ consists of

(i)