CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Tree Models

Lecturer: Michael I. Jordan *Scribes: Luca Schenato, Bruno Sinopoli*

1 Introduction

A tree model is an acyclic graph, with probability model:

$$p(x) = p(x_r) \prod_{i \neq r} p(x_i | x_{\pi_i}),$$

where r is the root node, and where π_i indexes the parent of node i. We have chosen a directed graphical representation, but we could equally well work with an undirected representation.

Consider the plate in figure 1, which contains N IID replicas of a basic graphical model on six nodes.

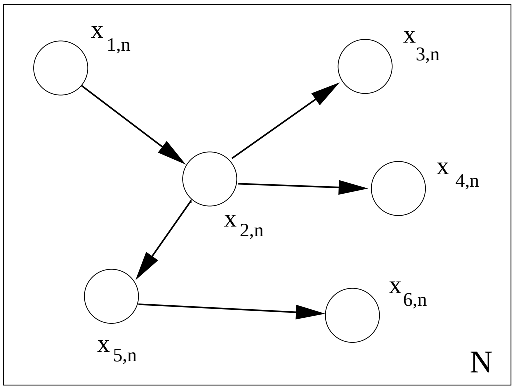

Figure 1: A Tree Model

Let M denote the number of nodes in a replica, N the number of replicas, $X_{i,n}$ denotes node i in replica n, with $i = 1 \dots M$ and $n = 1 \dots N$. Let $U = \{1, \dots, M\}$ be the set of indices of nodes in each replica, $X_{U,n} = \{X_{1,n}, \dots, X_{M,n}\}$ the set of nodes in the n^{th} replica. Given $C \subseteq U$, $X_{C,n}$ will denote the subset of nodes in the n^{th} replica whose indices are in C. Finally let $\mathcal{D} = \{x_{U,1}, \dots, x_{U,N}\}$ denote the observed data set.

We wish to find the best tree to fit our data. This requires us to solve a model selection problem—to find the best set of edges—and a parameter estimation problem—to find the best set of numeric parameters associated with the edges.

Given that every tree has the same number of edges and thus the same number of parameters, model selection methods based on the number of parameters are of little use. In fact this is a situation where maximum likelihood provides guidance in solving both the model selection aspect of the problem as well as the parameter estimation aspect of the problem.

The likelihood for the tree is $\prod_{n=1}^{N} p(x_{U,n}|\theta)$ and the log-likelihood is $\ell(\theta|D) = \sum_{n=1}^{N} \log p(x_{U,n}|\theta)$.

Let x_i be a discrete variable. In this context it is more convenient to work with sums over configurations rather than working with sums over observations.

Let's define $m(x_U) \triangleq \sum_{n=1}^{N} \delta(x_U, x_{U,n})$ where x_U is a configuration and $x_{U,n}$ is the n^{th} observation. Thus, $m(x_U)$ counts the number of times a configuration x_U appears in the data.

The log likelihood can be written as follows:

$$\begin{aligned}
\ell(\theta|D) &= \sum_{n=1}^{N} \log \left(\prod_{x_U} p(x_U|\theta)^{\delta(x_U, x_{U,n})} \right) \\
&= \sum_{n=1}^{N} \sum_{x_U} \delta(x_U, x_{U,n}) \log p(x_U|\theta) \\
&= \sum_{x_U} \left(\sum_{n=1}^{N} \delta(x_U, x_{U,n}) \right) \log p(x_U|\theta) \\
&= \sum_{x_U} m(x_U) \log p(x_U|\theta)
\end{aligned}$$

From this formula it can be easily seen that tree models belong to the exponential family, where $m(x_U)$ is the sufficient statistic and $\log p(x_U|\theta)$ is the natural parameter.

1.1 Maximum Spanning Trees

Let $G(U, E)$ be an undirected graph, where U is the set of nodes and E is the set of edges. Assign to each edge (u, v) a weight $w_{(u,v)}$. A Maximum Spanning Tree (MST) is a tree that includes every node and has maximal total weight. The following greedy algorithm, known as Kruskal's algorithm, can be used to compute the MST on a graph.

```
KRUSKAL (G,w)
    A ⟵ ∅
    sort E by nonincreasing weight w for each (u,v) ∈ E in order
    if adding (u,v) creates a cycle
        continue
    else
        A ⟵ A ∪ {(u, v)}
    end
```

An example run of Kruskal's algorithm is given in Figure 2.

1.2 KL Divergence

Let p, q be discrete probability distributions on the same alphabet $\mathcal{A}$.

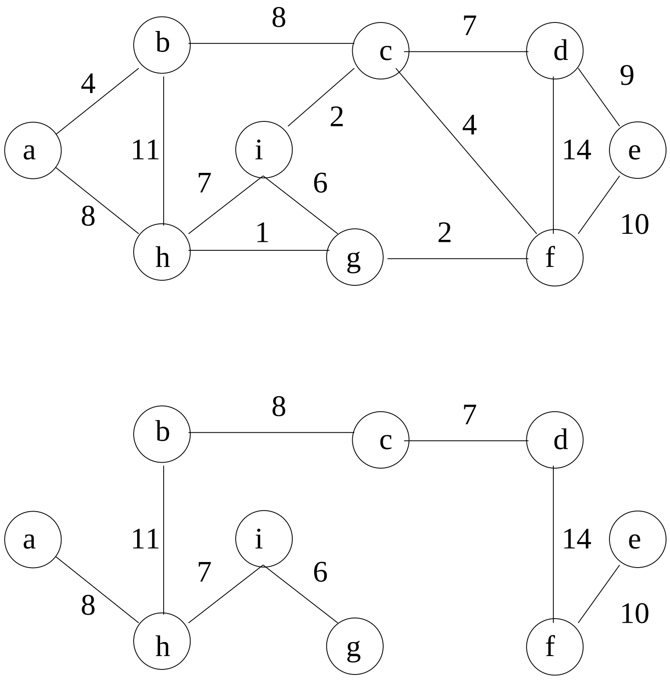

Figure 2: Maximum Spanning Tree on a graph

$$D(p||q) \triangleq \sum_{x \in \mathcal{A}} p(x) \log \frac{p(x)}{q(x)} \tag{1}$$

Theorem

$D(p||q) \geq 0$.

Proof:

$$\begin{aligned} 0 &= \log 1 \\ &= \log \sum_{x \in \mathcal{A}} q(x) \\ &= \log \sum_{x \in \mathcal{A}} p(x) \frac{q(x)}{p(x)} \\ &\geq \sum_{x \in \mathcal{A}} p(x) \log \frac{q(x)}{p(x)}, \end{aligned}$$

which implies:

$$\sum_{x \in \mathcal{A}} p(x) \log \frac{p(x)}{q(x)} \geq 0.$$

Note that $D(p||q) = 0$ if and only if $p = q$. Note also that KL divergence is not a real distance since it is not symmetric and does not satisfy the triangle inequality.

An immediate application of KL divergenge can be found in information theory. Given a communication channel where p represents the true underlying probability distribution, $D(p||q)$ represents the extra code length to pay if q is used instead of p.

1.3 Mutual information

The mutual information is a measure of the information that a random variable contains about another random variable.

Consider random variables X and Y with joint probability distribution $p(x, y)$ and marginal distributions $p(x)$ and $p(y)$. The mutual information $I(X, Y)$ is defined as:

$$
\begin{aligned}
I(X,Y) &\triangleq D\left(p(x,y)||p(x)q(y)\right) \\
&= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)q(x)}
\end{aligned}
$$

Note that the mutual information is symmetric: $I(X, Y) = I(Y, X)$.

1.4 Maximum Likelihood for tree models

Algorithm

1) Compute empirical $\tilde{I}(X_u, X_v)$ for all (u,v);

2) using $\tilde{I}(X_u, X_v)$ as the weight $w_{(u,v)}$ find the MST.

Derivation:

Define $\tilde{p}(x_U) \triangleq m(x_U)/N$; the empirical distribution.

We wish to maximize the following expression:

$$
\frac{\ell}{N} = \sum_{x_U} \tilde{p}(x_U) \log p(x_U|\theta) \tag{2}
$$

where we recall that:

$$
p(x_U|\theta) = p(x_r) \prod_{i \neq r} p(x_i|x_{\pi_i}).
$$

Note that the parameters θ are the probabilities $p(x_r)$ and $p(x_i|x_{\pi_i})$, which in the discrete case can be viewed as tables.

Recall also that the maximum likelihood estimates of the parameters of discrete distributions are sample proportions. Thus,

$$
\hat{p}_{ML}(x_r) = \tilde{p}(x_r)
$$

and

$$
\hat{p}_{ML}(x_i|x_{\pi_i}) = \tilde{p}(x_i|x_{\pi_i})
$$

where $\tilde{p}_(x_r)$ and $\tilde{p}(x_i|x_{\pi_i})$ are marginals and conditionals computed under the empirical distribution $\tilde{p}(x_U)$.

Therefore

$$
\frac{\ell}{N} = \sum_{x_U} \tilde{p}(x_U) \left(\log\, p(x_r) + \sum_{i \neq r} \log\, p(x_i|x_{\pi_i}) \right).
$$

Maximizing over the parameters for a fixed tree structure, and denoting the resulting maximum as ℓ^*, we have:

$$\begin{aligned}
\frac{\ell^*}{N} &= \sum_{x_U} \tilde{p}(x_U) \left(\log \tilde{p}(x_r) + \sum_{i \neq r} \log \tilde{p}(x_i | x_{\pi_i}) \right) \\
&= \sum_{x_U} \tilde{p}(x_U) \left(\log \tilde{p}(x_r) + \sum_{i \neq r} \log \frac{\tilde{p}(x_i, x_{\pi_i})}{\tilde{p}(x_{\pi_i})} \right) \\
&= \sum_{x_U} \tilde{p}(x_U) \sum_{i \neq r} \log \frac{\tilde{p}(x_i, x_{\pi_i})}{\tilde{p}(x_i)\tilde{p}(x_{\pi_i})} + \sum_{x_U} \tilde{p}(x_U) \sum_{i} \log \tilde{p}(x_i).
\end{aligned}$$

Note two facts about this intermediate result. First, the second term is independent of the tree structure—let us therefore refer to it as a constant "C." Second, there is a one-to-one correspondence between the nodes $i \neq r$ and the edges of the underlying undirected graph (U, E), where by we "underlying undirected graph" we mean that we simply ignore the directions of the edges. In particular, let i be associated with the edge (i, π_i) that links the node to its parent. Given this correspondence, the sum over $i \neq r$ is equivalent to a sum across the edges E of the underlying undirected graph.

We continue the derivation:

$$\begin{aligned}
\frac{\ell^*}{N} &= \sum_{i \neq r} \sum_{x_U} \tilde{p}(x_U) \log \frac{\tilde{p}(x_i, x_{\pi_i})}{\tilde{p}(x_i)\tilde{p}(x_{\pi_i})} + C \\
&= \sum_{i \neq r} \sum_{x_i, x_{\pi_i}} \tilde{p}(x_i, x_{\pi_i}) \log \frac{\tilde{p}(x_i, x_{\pi_i})}{\tilde{p}(x_i)\tilde{p}(x_{\pi_i})} + C \\
&= \sum_{(s,t) \in E} \sum_{x_s, x_t} \tilde{p}(x_s, x_t) \log \frac{\tilde{p}(x_s, x_t)}{\tilde{p}(x_s)\tilde{p}(x_t)} + C \\
&= \sum_{(s,t) \in E} \tilde{I}(X_s, X_t) + C.
\end{aligned}$$

where $\tilde{I}(X_s, X_t)$ is the mutual information of the random variables X_s and X_t under the empirical distribution $\tilde{p}(x_U)$.

Thus the likelihood of the graph is the sum of the mutual information values associated with the edges E of the graph. Letting these values be weights $w_{(s,t)}$, we see that the maximum likelihood tree is a maximal spanning tree.

Note finally that any directed tree consistent with the undirected tree (U, E) has the same likelihood. Thus our method picks out an undirected graphical structure; it does not orient the edges.