An Explanation of the Solution for the Mixture of Trees Problem

Jens Lagergren

November 16, 2020

I Mixture of trees with observable variables

The purpose of this note is to explain why the Mixture of Trees problem in Assignment 2 can be solved the way described in the assignment. That is, you can successfully solve the problem without understanding this solution; however, in case you would like to understand, please go ahead and read this explanation. This note has three subsections: Subsection 1.1 – a repetition of the problem formulation; Subsection 1.2 – an ML solution for the less complex problem concerned with a single tree, which subsequently is used for the general problem; and Subsection 1.3 – the solution for the actual mixture problem.

1.1 Problem formulation

Consider the mixture model $\mathcal{M} = (\pi, \tau)$, where π is a categorical distribution on $[K]$ and $\tau = \{(T_k, \theta_k) : k \in [K]\}$ is a set of K graphical models that each is a tree with vertices V and root r. All variables are are binary and observable. There is an EM algorithm that, for given data $\mathcal{D} = \{x^n : n \in [N]\}$, estimates $\mathcal{M}$ by iteratively performing the following steps w.r.t. to a current $\mathcal{M} = (\pi, \tau)$, which gives a new $\mathcal{M}' = (\pi', \tau')$

1. For each n, k, compute the responsibilities
$$r_{n,k} = \pi_k p(x^n | T_k, \theta_k)/p(x^n).$$
2. Set $\pi'_k = \sum_{n=1}^{N} r_{n,k}/N$.
3. For each k, let G_k be a directed graph with edge weights defined by $w(st) = I_q(X_s, X_t)$, where $I_{q^k}(X_s, X_t)$ is the mutual information between X_s and X_t under the distribution q^k defined by
$$q^k(X_s = a, X_t = b) = \frac{\sum_{n \in [N]: X_s^n = a, X_t^n = b} r_{n,k}}{\sum_{n \in [N]} r_{n,k}},$$
i.e.,
$$I_{q^k}(X_s, X_t) = \sum_{a,b \in \{0,1\}} q^k(X_s = a, X_t = b) \log \frac{q^k(X_s = a, X_t = b)}{q^k(X_s = a) q^k(X_t = b)}.$$
Moreover, any term in $I_{q^k}(X_s, X_t)$ for which $q^k(X_s = a, X_t = b) = 0$ is considered to be 0.
4. Let T'_k be a maximum spanning tree in G_k.
5. Let $\theta'_k(X_r) = q^k(X_r)$ and $\theta'_k(X_s = a | X_t = b) = q^k(X_s = a | X_t = b)$.

The root stays the same; it is facilitating our computations, but any root would give the same result. Initialize the EM algorithm randomly, independently of the data, and use sieving.

1.2 MLE for single tree

Consider the less complex problem of given $\mathcal{D} = \{x^n : n \in [N]\}$ obtaining a single tree GM (T, θ) maximizing the likelihood $p(\mathcal{D}|T, \theta)$.
Strategy: (1) express the log-likelihood for one fix, but arbitrary, tree T as the sum of edges weight for a specific edge weight assignment, (2) notice that the weight is the same across all trees, and (3) conclude that the ML tree is the maximum spanning tree (MST) in the complete graph on V with these edge weights.

Let T be an arbitrary, but fixed, tree with vertex set V and root r. We consider the edges of T to be directed away from the root and write $st \in A(T)$ if s is the parent of t. Let

$$N_a^s = \sum_{n=1}^{N} r_n I(x_s^n = a)$$

and

$$N_{a,b}^{st} = \sum_{n=1}^{N} r_n I(x_s^n = a, x_t^n = b),$$

here $I(\cdot)$ is the indicator function, not to be confused with the mutual information, and r_n is simply 1, but later on it will be useful to have performed this derivation for an arbitrary $r_n \geq 0$. At the first reading, consider r_n to be 1, which yields $\sum_{a,b} N_{ab}^{st} = N$; however, in general $\sum_{a,b} N_{ab}^{st} = \sum_{n=1}^{N} r_n$.

Notice we have not specified θ, but we will start by identifying θ_{ML}. For any $x \in \mathcal{D}$ and θ,

$$\log\, p(x|T, \theta) = \log\, p(x_r|\theta) + \sum_{st \in A(T)} \log\, p(x_t|x_s, \theta)$$

and, moreover, for $r_n = 1$ the first equality below holds and in general (2) will be the appropriate starting point,

$$\begin{aligned}
&\log p(\mathcal{D}|T, \theta) && (1)\\
&= \sum_{n=1}^{N} r_n \log p(x_n|T_k, \theta_k) && (2)\\
&= \sum_{n=1}^{N} r_n \left[\log p(x_r^n|\theta) + \sum_{st \in A(T)} \log p(x_t^n|x_s^n, \theta)\right] && (3)\\
&= \sum_{a} N_a^r \log p(X_r = a|\theta) + \sum_{a,b} N_{a,b}^{st} \log p(X_t = b|X_s = a, \theta). && (4)
\end{aligned}$$

Let

$$q(X_s = a, X_t = b) \stackrel{\text{def}}{=} \frac{N_{a,b}^{st}}{\sum_{a,b} N_{a,b}^{st}}.$$

Notice, for $r_n = 1$, q is simply the empirical distribution, that is, the frequency with which a certain value, or combination of values, is observed. By the MLE for categorical distributions (4) is maximized by

$$\begin{aligned}
p(X_r = a|\theta) &= \tfrac{N_a^r}{\sum_a N_a^r} = q(X_r = a)\\
p(X_t = b|X_s = a, \theta) &= \tfrac{N_{a,b}^{st}}{\sum_b N_{a,b}^{st}} = q(X_t = b|X_s = a).
\end{aligned}$$

Noticing that $\sum_a N_a^r \log q(X_r = a)$ and $\sum_{a,b} N_{ab}^{st} \log q(X_t = b) = \sum_b N_b^t \log q(X_t = b)$ are independent of T, allows us to remove an additive term from (5) and to and to add one in (6), repectively. Moreover, recalling that $\sum_{a,b} N_{ab}^{st} = \sum_{n=1}^{N} r_n$, i.e., independent of st, it is evident when each term in

(7) is divided by this quantity, a proportional expression is obtained, i.e., (8).

$$
\begin{aligned}
&\log p(\mathcal{D}|T,\theta_{\mathrm{ML}}) \\
&= \sum_a N_a^r \log q(X_r=a) + \sum_{st\in A(T)} \sum_{a,b} N_{a,b}^{st} \log q(X_t=b|X_s=a) && (5)\\
&\stackrel{+}{=} \sum_{st\in A(T)} \sum_{a,b} N_{a,b}^{st} \log \frac{q(X_s=a,X_t=b)}{q(X_s=a)} - \sum_{st\in A(T)} \sum_{a,b} N_{a,b}^{st} \log q(X_t=b) && (6)\\
&= \sum_{st\in A(T)} \sum_{a,b} N_{a,b}^{st} \log \frac{q(X_s=a,X_t=b)}{q(X_s=a)q(X_t=b)} && (7)\\
&\propto \sum_{st\in A(T)} \sum_{a,b} \frac{N_{a,b}^{st}}{\sum_{a,b} N_{a,b}^{st}} \log \frac{q(X_s=a,X_t=b)}{q(X_s=a)q(X_t=b)} && (8)\\
&= \sum_{st\in A(T)} \sum_{a,b} q(X_s=a,X_t=b) \log \frac{q(X_s=a,X_t=b)}{q(X_s=a)q(X_t=b)} && (9)\\
&= \sum_{st\in A(T)} I_q(X_s,X_t) && (10)
\end{aligned}
$$

That is, the ML tree is the one maximizing (10). We let $I_q(X_s, X_t)$ be the weight of the edge st and notice that it is the same for both directions and, also, the same across all trees that st belongs to. Let G be the complete graph on V with edge weights w defined by $w(s,t) = I_q(X_s, X_t)$; an MST in G maximize (10) and is, hence, the ML tree.

1.3 An EM algorithm for mixtures of trees

Consider the mixture of trees model formulated in the problem in the assignment. Its complete likelihood can be expressed as follows.

$$p(\mathcal{D}|\mathcal{M}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\pi_k p(x_n|T_k,\theta_k)\right]^{I(Z_n=k)}.$$

Hence, the models ECLL is

$$\sum_{n=1}^{N} E_{Z_n|\mathcal{M},x_n}\left[\log\, p(x_n,Z_n|\mathcal{M}')\right] = \sum_{k=1}^{K}\left[\sum_{n=1}^{N} r_{n,k}\right]\log \pi_k' + \sum_{k=1}^{K}\sum_{n=1}^{N} r_{n,k}\log p(x_n|T_k',\theta_k').$$

Hence, the ECLL is maximized by

$$\pi_k' = \frac{\sum_{n=1}^{N} r_{n,k}}{\sum_{k=1}^{K}\sum_{n=1}^{N} r_{n,k}}$$

and, for each k, the pair T_k', θ_k' that maximize

$$\sum_{n=1}^{N} r_{n,k}\log p(x_n|T_k',\theta_k').$$

Moreover, by comparing this expression with (3), it is evident that the latter pair can be found by the algorithm described in Subsection 1.2 above.