Chapter 4

Learning mixtures of trees in the Bayesian framework

The ML paradigm assumes that the model is estimated using the data only and excluding other sources of knowledge about parameters or model structure. But, if prior knowledge exists and if it can be represented as a probability distribution over the space of mixtures of trees models, then one can use the Bayesian formulation of learning to combine the two sources of information.

In the Bayesian framework, the main object of interest is the posterior distribution over models (in our case mixtures of trees) given the observed data $Pr[Q|\mathcal{D}]$. By Bayes' formula, the posterior is proportional to

$$Pr[Q|\mathcal{D}] \;\propto\; Pr[Q,\,\mathcal{D}] \;=\; Pr[Q] \prod_{x\in\mathcal{D}} Q(x) \tag{4.1}$$

In the above $Pr[Q]$ represents the prior distribution over the class of mixture of trees models; the second factor is $Q(\mathcal{D})$, the likelihood of the data given the model Q.

The probability of an observation x is obtained by *model averaging*

$$Pr[x] \;=\; \int Q(x) Pr[Q|\mathcal{D}]\, dQ \tag{4.2}$$

However, except for a few special cases neither $Pr[x]$ nor the posterior $Pr[Q|\mathcal{D}]$ are representable in closed form. A common approach is to approximate the posterior distribution around its mode(s) for example by the Laplace approximation []. Another approach is to replace the integration in equation (4.2) by a finite sum over a set $\mathcal{Q}$ of models with high posterior probability.

$$Pr[x] \;=\; \sum_{Q\in\mathcal{Q}} Q(x) Pr[Q|\mathcal{D}] \tag{4.3}$$

This approximation is equivalent to setting $Pr[Q|\mathcal{D}]$ to 0 for all the mixtures not in $\mathcal{Q}$. Consequently, the normalization constant in the above formula is computed over $\mathcal{Q}$ only.

Finally, if we are to choose one model only to summarize the posterior distribution $Pr[Q|\mathcal{D}]$ then a natural choice is the mean of the distribution. As it will be shown the mean can sometimes be expressed as the MAP estimate under a certain parameterization.

Finding the modes of the posterior distribution $Pr[Q|\mathcal{D}]$ is a necessary step in all the three approaches; therefore, although *Maximum A-Posteriori (MAP)* estimation of Q is not seen as a purpose per se, this chapter will be concerned with maximizing the posterior distribution $Pr[Q|\mathcal{D}] \propto Pr[Q,\mathcal{D}]$.

4.1 MAP estimation by the EM algorithm

According to equation (4.1) in the previous section, maximizing the posterior log-likelihood of a model is equivalent to maximizing

$$\log Q(\mathcal{D}) \;+\; \log Pr[Q] \tag{4.4}$$

which differs from $\log Pr[Q|\mathcal{D}]$ only by an additive constant. The EM algorithm previously used to find Maximum Likelihood estimates can be adapted to maximize the above expression (which represents the expression of the log-likelihood plus the term $\log Pr[Q]$) [52]. The quantity to be iteratively maximized by EM is now

$$E[\log Pr[Q|x^{1,\ldots N}, z^{1,\ldots N}]] \;=\; \log Pr[Q] \;+\; E[l_c(x^{1,\ldots N}, z^{1,\ldots N}|Q)] \tag{4.5}$$

It is easy to see that the added term does not have any influence on the E step of the EM algorithm, which will proceed exactly as before. However, in the M step, we must be able to successfully maximize the r.h.s. of (4.5). We have seen that in the previous chapter exact maximization was enabled by the fact that we obtained a separate set of equations for λ and for each of the mixture components T^k. Therefore, we will look for priors over the mixtures of trees models that are amenable to this decomposition.

For a given m a prior over the space of mixtures of trees with m mixture components comprises a prior over the distribution $\lambda_{1,\ldots m}$ of the hidden variable and priors over each of the m trees' structures and parameters. We require that the prior is expressed in the product form

$$Pr[Q] \;=\; Pr[\lambda_{1,\ldots m}] \prod_{k=1}^{m} \underbrace{Pr[E_k] Pr[\text{parameters} T^k|E_k]}_{Pr[T^k]}. \tag{4.6}$$

Moreover, remember that the key to the efficient maximization in equation (3.21) was the fact that its r.h.s could be further decomposed into a sum of independent terms, each of them corresponding to an edge of T^k. Therefore, a similar property is desirable for the prior $Pr[T^k]$. A prior probability for a tree has two components: the prior for the tree structure (represented by the edge set E_k) and the prior for the tree parameters, given the structure. We shall require that both components decompose in a way that matches the factoring of the likelihood in equation (3.21):

$$Pr[T^k] \;=\; Pr[E_k] \prod_{(u,v)\in E_k} Pr[T^k_{uv}] \tag{4.7}$$

$$\propto \prod_{(u,v)\in E_k} \beta^k_{uv} Pr[T^k_{uv}] \tag{4.8}$$

A prior over the class of mixtures of trees having all the above properties is called a *decomposable prior*.

In the context of decomposable priors, at the M step of the EM algorithm the new model Q^{new} is obtained by solving the following set of decoupled maximization problems:

- for the λ parameters

$$\lambda^{new}_{1,\ldots m} \;=\; \operatorname*{argmax}_{\Sigma_k \lambda_k = 1} \sum_{k=1}^{m} \Gamma_k \log \lambda_k \;+\; \log P(\lambda_{1,\ldots m}). \tag{4.9}$$

- for each tree T^k equation (3.21) in the previous chapter is replaced by

$$T^{k\,new} \;=\; \underset{T^k}{\text{argmax}} \sum_{i=1}^{N} P^k(x^i) \log T^k(x^i) \;+\; \log Pr[T^k] \tag{4.10}$$

Requiring that the prior be decomposable is equivalent to making strong independence assumptions: for example, it means that the prior probability of each tree in the mixture is independent of all the others as well as of the probability distribution of the mixture variable. In the following sections, the implications of some of these independence assumptions will be made more explicit. It will also be shown that, although the independence assumptions made are strong ones, they are not too restrictive. The class of decomposable priors is rich enough to contain members that are interesting and of practical importance.

4.2 Decomposable priors for tree distributions

4.2.1 Decomposable priors over tree structures

The general form of a decomposable prior for the tree structure E is one where each edge contributes a constant factor independent of the presence or absence of other edges in E.

$$Pr[E] \;\propto\; \prod_{(u,v)\in E} \exp(-\beta_{uv}) \tag{4.11}$$

With this prior, the expression to be maximized in the M step of the EM algorithm (former equation (3.19))becomes

$$E[l_c \,|\, x^{1,\ldots N}, \mathcal{M}] \;=\; \sum_{k=1}^{m} \Gamma_k \log \lambda_k + \sum_{k=1}^{m} \Gamma_k \left(\sum_{i=1}^{N} P^k(x^i) \log T^k(x^i) - \sum_{(u,v)\in E_k} \frac{\beta^k_{uv}}{\Gamma_k} \right) \tag{4.12}$$

Consequently, each edge weight in tree T^k is penalized by its corresponding β divided by the total number of points that tree k is responsible for:

$$W^k_{uv} \;=\; I^k_{uv} - \frac{\beta^k_{uv}}{\Gamma_k} \tag{4.13}$$

A negative β_{uv} increases the probability of uv being present in the final solution. On the contrary, if β_{uv} is positive, it acts like a penalty on the presence of edge uv in the tree. If β_{uv} is sufficiently large the weight W_{uv} becomes negative and the edge is not added to the tree. Hence, by introducing edge penalties, one can obtain trees T^k having fewer than $n-1$ edges and thus being disconnected. Notice that the strength of the prior decreases with Γ_k the total number of points assigned to mixture component k. Thus, for equal priors for all trees T^k, trees accounting for fewer data points will be penalized stronger and therefore will be likely to have fewer edges.

If one chooses the edge penalties to be proportional to the increase in the number of parameters caused by the addition of edge uv to the tree,

$$\beta_{uv} \;=\; \beta(r_u - 1)(r_v - 1) \tag{4.14}$$

then a *Minimum Description Length* (MDL) [57] type of prior is implemented. Note that the effective penalty for each parameter is inversely proportional to the number of data points Γ_k the tree that the parameter belongs to is responsible for.

In [33], in the context of learning Bayesian networks, the following prior is suggested:

$$Pr[E] \;=\propto \kappa^{\Delta(E,E^*)} \tag{4.15}$$

where $\Delta(\cdot)$ is a distance metric and E^* is the prior network structure. Thus, this prior penalizes deviations from the prior network. For trees, the distance metric becomes the symmetric difference:

$$\Delta(E,E^*) \;=\; |(E\setminus E^*)\bigcup(E^*\setminus E)| \tag{4.16}$$

This prior is also factorable, entailing

$$\beta_{uv} \;=\; \begin{cases} -\ln\kappa & (uv)\in E^* \\ \ln\kappa & (uv)\notin E^* \end{cases} \tag{4.17}$$

What happens in the case of mixtures of trees with shared structure? Recall that in this case each edge uv had a weight proportional to the mutual information between u and v conditioned on z: $I_{uv|z}$ (3.31). The effect of the decomposable prior is to penalize this weight by $-\frac{\beta_{uv}}{N}$.

4.2.2 Priors for tree parameters: the Dirichlet prior

We will introduce now an important subclass of decomposable priors for parameters called *Dirichlet* priors. The Dirichlet prior is the conjugate prior of the *multinomial* distribution. Since the distribution of the mixture variable is a multinomial as well, the facts shown in this section will cover this case too.

Let z be a discrete random variable taking r values and let $\theta_j = P_z(j),\ j=1,\dots r$. Then, the probability distribution of an i.i.d. sample of size N from P_z is given by

$$P(z^{1,\dots N}) \;=\; \prod_{j=1}^{r}\theta_j^{N_j} \tag{4.18}$$

where N_j, $j=1,\dots r$ represent the number of times the value j is observed in and are called the *sufficient statistics* of the data. The sample itself is said to obey a *multinomial* distribution.

The Dirichlet distribution is defined over the domain of $\theta_{1,\dots r}$ and depends on r real parameters $N'_{1,\dots r}>0$ by

$$D(\theta_{1,\dots r};N'_{1,\dots r}) \;=\; \frac{\Gamma(\sum_j N'_j)}{\prod_j\Gamma(N'_j)}\prod_j\theta_j^{N'_j-1} \tag{4.19}$$

In the above $\Gamma()$ represents the Gamma function defined by

$$\Gamma(p) \;=\; \int_0^\infty t^{p-1}e^t dt. \tag{4.20}$$

For any nonnegative integer n

$$\Gamma(n+1) \;=\; n! \tag{4.21}$$

The importance of the Dirichlet distribution in connection with a multinomially distributed variable resides in the following fact: If the parameters θ of the multinomial distribution have as prior a Dirichlet distribution with parameters $N'_j,\ j=1,\dots r$, then, after observing a sample with sufficient statistics $N_j,\ j=1,\dots r$, the posterior distribution of θ is a Dirichlet distribution with parameters $N'_j+N_j,\ j=1,\dots r$. This justifies denoting the distribution's parameters by N'. One popular alternative parameterization for the Dirichlet distribution is given by:

$$N' = \sum_{j=1}^{r} N'_j \tag{4.22}$$

$$P'_j = \frac{N'_j}{N'} \quad j=1,\dots r \tag{4.23}$$

$$\tag{4.24}$$

Note that in this parameterization the means of the parameters θ_j are equal to P'_j. We say that the Dirichlet distribution is a conjugate prior for the class of multinomial distributions. The property of having conjugate priors is characteristic for the exponential family of distributions.

The Dirichlet prior in natural coordinates

The multinomial distribution was represented before as defined by the parameters θ_j, $j=1,\dots r$. But there are infinitely many ways to parametrize the same distribution. For each set of parameters $y_{1,\dots r_y}$ the corresponding representation of the Dirichlet prior results from the well known change of variable formula

$$D(y_{1,\dots r_y};\, N'_{1,\dots r-1}) = D(\theta_{1,\dots r}(y_{1,\dots r_y}); N'_{1,\dots r}).|\frac{\partial\theta}{\partial y}| \tag{4.25}$$

with $|\frac{\partial y}{\partial \theta}|$ representing the absolute value of the determinant of the Jacobian of $\theta(y)$. Note that because of the presence of this factor, the maximum of $D(.; N'_{1,\dots r})$ has both a different value and a different position in each parametrization. This dependence of the parametrization is one fundamental drawback of MAP estimation which justifies the Bayesian and approximate Bayesian approaches mentioned above.

By contrast, the mean of $f(y)D(y;N')$ of any measurable function f over any measurable set is independent of the parametrization. In particular, the mean of the Dirichlet distribution is independent of the parametrization and equal to

$$E[\theta_j] = \frac{N'_j}{\sum_{j'=1}^{r} N'_{j'}} \quad j=1,\dots r \tag{4.26}$$

Of special interest is the so called *natural* parametrization of the multinomial, defined by $r-1$ unconstrained parameters ϕ:

$$\phi_i = \log\frac{\theta_i}{\theta_r} \quad i=1,\dots r-1. \tag{4.27}$$

The parameters ϕ take values in $(-\infty,\infty)$ when $\theta_{1,\dots r}>0$. The reverse transformation from ϕ coordinates to θ coordinates is defined by:

$$\theta_r = \frac{1}{1+\sum_{i=1}^{r-1} e^{\phi_i}} \tag{4.28}$$

$$\theta_j = \frac{e^{\phi_j}}{1+\sum_{i=1}^{r-1} e^{\phi_i}}, \quad j=1,\dots r-1 \tag{4.29}$$

$$\tag{4.30}$$

In the natural parametrization, the Dirichlet distribution is expressed as

$$D(\phi_{1,\ldots r-1};\ N'_{1,\ldots r}) \;=\; \frac{\Gamma(\sum_{j=1}^{r} N'_j)}{\prod_{j=1}^{r}\Gamma(N'_j)} \prod_{i=1}^{r-1}\left(\frac{e^{\phi_i}}{1+\sum_{i=1}^{r-1} e^{\phi_i}}\right)^{N'_i} \frac{1}{(1+\sum_{i=1}^{r-1} e^{\phi_i})^{N'_r}} \tag{4.31}$$

A remarkable property of the natural parametrization is that its mode coincides in position with the mean. To see this, it suffices to equate to 0 the partial derivatives of the Dirichlet distribution w.r.t the ϕ parameters. After some calculations, one obtains

$$\frac{e^{\phi_i}}{1+\sum_{i'=1}^{r-1} e^{\phi_{i'}}} \;=\; \frac{N'_i}{\sum_{j=1}^{r} N'_j} \quad i=1,\ldots r-1 \tag{4.32}$$

or equivalently

$$\theta_j \;=\; \frac{N'_i}{N'} \quad j=1,\ldots r \tag{4.33}$$

Dirichlet priors for trees and mixtures

[33] show that the assumptions of *likelihood equivalence* which says that data should not help discriminate between structures which represent the same probability distribution, *parameter modularity* which says that the parameters corresponding to an edge of the tree should have the same prior every time the edge is present in the tree and *parameter independence* which says that in any directed tree parametrization the parameters of each edge are independent of anything else. These combined with some weak technical assumptions[1] imply that the parameter prior is Dirichlet.

In the case of trees, likelihood equivalence is automatically assured by definition (all tree parametrizations, directed or not, represent the same distribution). Parameter modularity and parameter independence are implicitly assumed as a basis for the tree fitting procedure that we have used. Moreover, requiring a decomposable prior over tree structures preserves parameter modularity. Hence, requiring decomposable priors not only fits naturally in the framework already adopted, but introduces almost no additional constraints. [33] also show that the likelihood equivalence constrains the Dirichlet priors for all the parameter sets to share a common equivalent sample size N'. All these results hold for trees and mixtures as described by the present framework. Moreover, it is easy to see that when learning tree distributions the Dirichlet parameters collapse into the fictitious sufficient statistics represented by the pairwise "counts" N'_{uvij} = "# times $u=i$ and $v=j$". Therefore, for the case of trees, the prior over all parameters of all possible edges of tree T^k can be described by the set of fictitious marginal counts

$$\{N'^{k}_{uvij},\ u,v\in V,\ i=1,\ldots r_u,\ j=1,\ldots r_v\}$$

Alternatively, one can normalize the counts and express the Dirichlet prior over all trees as a table of fictitious marginal probabilities P'_{uv} for each pair u,v of variables plus an *equivalent sample size* N' that gives the strength of the prior.

However, it is important to note that the fictitious marginal counts or alternatively the pairwise marginals P'_{uv} cannot be set arbitrarily[2]. The values P'_{uv} have to satisfy

$$P'_{uv}(x_u x_v) \;=\; \sum_{x\in\Omega(V):u=x_u,v=x_v} P'_V(x) \;\; \forall u,v,x_u,x_v \tag{4.34}$$

[1] These technical assumptions assumptions amount to the positivity of the joint prior.

[2] Take for example a set of three binary variables a,b,c. Assume that $P_{ab}(01)=P_{ab}(10)=0$ and $P_{bc}(01)=P_{bc}(10)=0$. This implies that $a=b=c$ and thus constrains $P_{ac}(01)=P_{ac}(10)=0$.

for some joint distribution P' over $\Omega(V)$.

The above represent a system of linear equality and inequality constraints. Verifying them directly is theoretically straightforward but computationally intractable on account on the exponential number of columns of the system's matrix (a number of the order of $|\Omega(V)|$). Still one can notice that the *uninformative* prior given by

$$P'_{uv}(x_u x_v) = \frac{1}{r_u r_v} \tag{4.35}$$

is valid since it represents the set of pairwise marginals of the uniform distribution over $\Omega(V)$[3].

If the Dirichlet prior is represented in the natural parameters and the empirical distribution is P, with sample size N, then, from the fact that the Dirichlet prior is a conjugate prior, it follows that finding the MAP tree is equivalent to finding the ML tree for

$$\tilde{P} = \frac{1}{N+N'}(N'P' + NP). \tag{4.36}$$

Consequently, the parameters of optimal tree will be

$$T_{u|v} = \frac{\tilde{P}_{uv}}{\tilde{P}_v} \tag{4.37}$$

and, according to the previous section and equation (4.33) they will also represent the mean of the posterior distribution. Moreover, using the parameter independence assumption, we can conclude that the optimal tree distribution itself is the mean of the posterior distribution given the structure.

For mixtures, maximizing the posterior translates into replacing P by P^k and N by Γ_k in equation (4.36) above. Namely, each step of the EM algorithm fits the optimal tree to the distribution

$$\tilde{P}^k = \frac{1}{\Gamma_k + N'}(N'P' + \Gamma_k P^k). \tag{4.38}$$

If we want to distinguish between trees, then different Dirichlet priors can be used for each tree. By the previous arguments, at each M step the resulting trees T^k and λ_k values represent the mean of their respective posterior distribution. However, this does not imply that the mixture $Q = \sum \lambda_k T^k$ is also the mean of its own posterior. The presence of the hidden variable cancels the independence assumption that allows us to conclude this for the single tree.

[3]This prior is called the BDeu prior in [33]. We note in passing that the uninformative prior denoted there as the K2 metric is not a valid prior from the point of view of equations (4.34).