

Discussion of the paper

Bayesian Treed Generalized Linear Models

by H. A. Chipman, E. I. George and R. E. McCulloch.

Christopher M. Bishop

Microsoft Research

7 J J Thomson Avenue,

Cambridge, CB3 0FB, U.K.

cmbishop@microsoft.com

<http://research.microsoft.com/~cmbishop>

In *Bayesian Statistics 7*, J. M. Bernardo *et al.* (editors),
Oxford University Press (2003) 98–101.

In this stimulating paper, the authors have successfully exploited Markov chain Monte Carlo methods to explore the space of graphs for CART-like trees in which the terminal nodes represent generalized linear models (GLMs). Integration over the parameters of the terminal GLMs, in order to compute the marginal likelihood (probability of data given the model) for the MCMC search, is accomplished using the Laplace approximation. Hyper-parameters (such as those governing the GLM parameters) are either set by hand or fixed after a brief empirical search.

The underlying CART model on which this approach is based, however, suffers from some significant limitations, namely (i) the splits are axis-aligned, i.e. dependent on only one of the input variables at a time (this limitation is removed in some other variants of CART), (ii) the splits are binary, (iii) the splits are hard, so that each region of input space is associated with one, and only one, leaf node.

An alternative tree-based model, which avoids these limitations, is the *Hierarchical Mixtures of Experts* (HME) proposed by Jordan and Jacobs (1994). Each non-terminal node in the tree, called a *gating* node, corresponds to a multi-way indicator variable $Z = \{Z_1, \dots, Z_M\}$, where $Z_i \in \{0, 1\}$ and $\sum_i Z_i = 1$. The conditional distribution of Z is given by a normalized exponential, or *softmax*, function

$$P(Z_i = 1|V, X) = \frac{\exp(V_i^T X)}{\sum_{j=1}^M \exp(V_j^T X)}$$

where X is the vector of explanatory, or input, variables, and $\{V_i\}$ is a set of parameter vectors governing the orientation and steepness of the gating function.

In the case of binary splits this is equivalent to a single binary indicator variable Z with $P(Z = 1|V, X) = \sigma(V^T X)$, where $\sigma(X) \equiv 1/(1 + \exp(-X))$ is the logistic sigmoid function. For a given X , the probability of selecting a particular terminal node is obtained by multiplying together all of the conditional probabilities along the unique path from the root node to the terminal node. Thus each point of the input space is assigned probabilistically to each of the terminal nodes through a partition of unity.

The terminal nodes for the HME model follow those of Chipman, George, and McCulloch (2002), namely softmax for multi-way classification, Gaussian for regression and so on. Jordan and Jacobs (1994) proposed an efficient EM algorithm for setting the model parameters of the HME, once the architecture of the tree has been prescribed.

It would be straightforward to use the MCMC approach of Chipman, George, and McCulloch (2002) to explore the tree structure of the Hierarchical Mixture of Experts model. While this could again be accomplished using a Laplace approximation to determine the likelihood score for each graph, a more appealing approach is to use variational methods (Jordan, Ghahramani, Jaakkola, and Saul 1998) to marginalize over the model parameters.

Variational methods optimize an analytical approximation to the posterior distribution by maximization of a lower bound on the log marginal likelihood (in contrast to the Laplace approximation which simply fits the second order moments at a mode of the distribution). The variational posterior distribution is chosen to have some factorization property with respect to the hidden variables, but is otherwise unconstrained.

The application of variational methods to the HME model is complicated by the fact that the softmax gating function does not lie within the conjugate exponential family. Previous attempts to apply variational methods to the HME have either resorted to mode fitting to circumvent this problem (Waterhouse, MacKay, and Robinson 1996) thereby losing the appealing property of the lower bound, or else have modelled the joint distribution of input and output variables (Ueda and Ghahramani 2002) which may be wasteful of resources and data particularly if the input space has high dimensionality.

Recently Jaakkola and Jordan (2000) have developed a variational bound for the logistic sigmoid function which takes the form

$$\sigma(x) \geq F(x, \xi) \equiv \sigma(\xi) \exp \left\{ (x - \xi)/2 - \lambda(\xi)(x^2 - \xi^2) \right\}$$

where $\lambda(\xi) = \tanh(\xi/2)/(4\xi)$, and ξ is a variational parameter. For any given value of x we can make this bound exact by an appropriate choice of the variational parameter ξ , namely $\xi = x$. (In fact the bound is exact at both $x = \xi$ and $x = -\xi$.) The bound is illustrated in Figure 1(a) for the case of $\xi = 2$, where the solid curve shows the logistic sigmoid function $\sigma(x)$, and the dashed curve shows the lower bound $F(x, \xi)$.

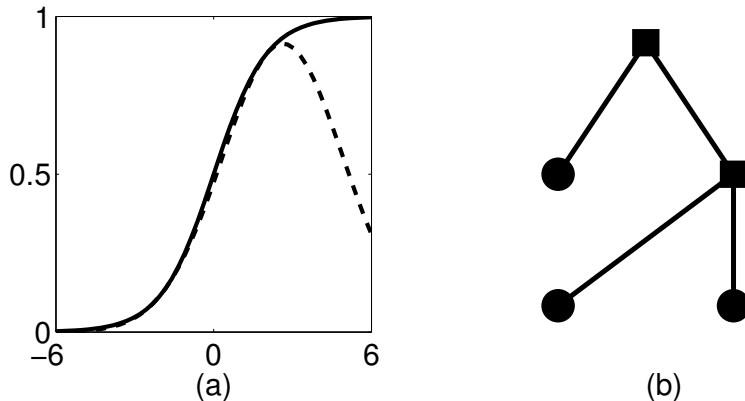


Figure 1: (a) Logistic sigmoid function and variational bound. (b) Example HME model.

We can use this bound to develop a fully Bayesian variational treatment of the HME model, for the case of binary gating nodes, which optimizes a rigorous lower bound on the marginal likelihood. Since the bound transforms the logistic sigmoid into the exponential of a quadratic form over V , we can obtain a conjugate model by using Gaussian priors $V \sim \mathcal{N}(\mu, \Sigma)$, with conjugate hyper-priors for μ and Σ . Note that, for each gating node, there is a separate variational parameter ξ_n for each observation n .

Optimization of the ξ parameters is achieved by maximizing the lower bound on the marginal likelihood, leading to the re-estimation equations

$$\xi_n^2 = X_n^T \langle VV^T \rangle X_n$$

where $\langle \cdot \rangle$ denotes an average with respect to the variational posterior distribution. Re-estimation of ξ_n is interleaved with re-estimation of the factors in the variational posterior.

Unfortunately, the variational bound given by $F(x, \xi)$ does not extend to multi-way gating nodes governed by softmax functions. However, a complex, multi-way division of the input space can always be represented using binary splits provided the tree structure is sufficiently rich.

We illustrate this approach using the simple HME model shown in Figure 1(b). Here the two square nodes denote logistic sigmoid gating functions, and the three circular terminal nodes correspond to Gaussian conditional distributions over the output variable. In Figure 2(a) we show a simple data set with one input and one output variable, together with the means of the variational posterior distributions over the terminal node variables V . The corresponding outputs of the gating nodes are shown by the two curves in Figure 2(b).

Finally in Figure 2(c) we show the conditional distribution $P(Y|X)$ of the output variable given the input variable, as a function of the input variable, i.e. each vertical slice through this plot represents $P(Y|X)$ for the given value of X .

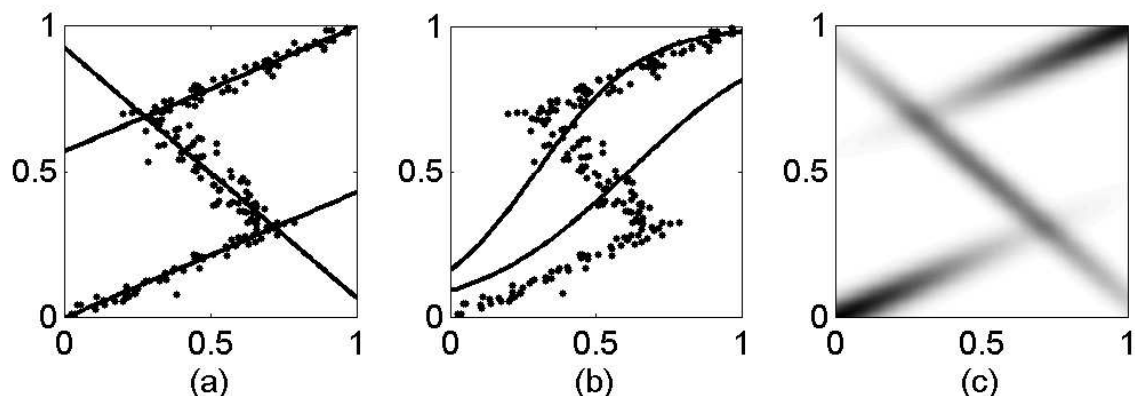


Figure 2: (a) Means of terminal nodes, (b) means of gating nodes, (c) conditional distribution.

A key feature to note is that this conditional distribution is multi-modal. This is possible because the gating node outputs are smooth functions of the input variable. Such multi-modality could *not* be captured in a CART model, since it assigns each point of the input space to one, and only one, of the terminal nodes ('hard' splits).

We have illustrated the variational HME using a regression example with Gaussian terminal nodes. It is straightforward to apply this approach to two-way classification problems for a model with logistic sigmoid terminal nodes simply by making further use of the variational bound $F(x, \xi)$.

It is also straightforward to evaluate the lower bound on the log marginal likelihood under the variational posterior distribution. This could readily be used in the Markov chain Monte Carlo scheme of Chipman, George, and McCulloch (2002) to sample from the posterior distribution over graphs, using operators to add and remove nodes as before.

The use of MCMC methods to explore the space of graphs in tree structured classification and regression models is clearly a rich area for research. The paper of Chipman, George, and McCulloch (2002) provides some important tools for tackling such problems, as well as motivating a number of future research directions.

References

- Chipman, H. A., E. I. George, and R. E. McCulloch (2002). Bayesian treed generalized linear models. In J. M. Bernardo (Ed.), *Proceedings Seventh Valencia International Meeting on Bayesian Statistics*. Oxford University Press. To appear.
- Jaakkola, T. and M. I. Jordan (2000). Bayesian parameter estimation through variational methods. *Statistics and Computing* 10, 25–37.
- Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1998). An introduction to variational methods for graphical models. In M. I. Jordan (Ed.), *Learning in Graphical Models*, pp. 105–162. Kluwer.
- Jordan, M. I. and R. A. Jacobs (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* 6(2), 181–214.
- Ueda, N. and Z. Ghahramani (2002). Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks* 15(10), 1223–1241.
- Waterhouse, S., D. MacKay, and T. Robinson (1996). Bayesian methods for mixtures of experts. In M. C. M. D. S. Touretzky and M. E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems*, pp. 351–357. MIT Press.