Abstract

We present an extension to the Mixture of Experts (ME) model, where the individual experts are Gaussian Process (GP) regression models. Using an input-dependent adaptation of the Dirichlet Process, we implement a gating network for an infinite number of Experts. Inference in this model may be done efficiently using a Markov Chain relying on Gibbs sampling. The model allows the effective covariance function to vary with the inputs, and may handle large datasets – thus potentially overcoming two of the biggest hurdles with GP models. Simulations show the viability of this approach.

1 Introduction

Gaussian Processes [Williams & Rasmussen, 1996] have proven to be a powerful tool for regression. They combine the flexibility of being able to model arbitrary smooth functions if given enough data, with the simplicity of a Bayesian specification that only requires inference over a small number of readily interpretable hyperparameters – such as the length scales by which the function varies along different dimensions, the contributions of signal and noise to the variance in the data, etc. However, GPs suffer from two important limitations. First, because inference requires inversion of an $n \times n$ covariance matrix where $n$ is the number of training data points, they are computationally impractical for large datasets. Second, the covariance function is commonly assumed to be stationary, limiting the modeling flexibility. For example, if the noise variance is different in different parts of the input space, or if the function has a discontinuity, a stationary covariance function will not be adequate. Goldberg et al [1998] discussed the case of input dependent noise variance.

Several recent attempts have been aimed at approximate inference in GP models [Williams & Seeger 2001, Smola & Bartlett 2001]. These methods are based on selecting a projection of the covariance matrix onto a smaller subspace (e.g. a subset of the data points) reducing the overall computational complexity. There have also been attempts at deriving more complex covariance functions [Gibbs 1997] although it can be difficult to decide a priori on a covariance function of sufficient complexity which guarantees positive definiteness.

In this paper we will simultaneously address both the problem of computational complexity and the deficiencies in covariance functions using a divide and conquer strategy inspired by the Mixture of Experts (ME) architecture [Jacobs et al, 1991]. In this model the input
space is (probabilistically) divided by a gating network into regions within which specific separate experts make predictions. Using GP models as experts we gain the double advantage that computation for each expert is cubic only in the number of data point in its region, rather than in the entire dataset, and that each GP-expert may learn different characteristics of the function (such as lengths scales, noise variances, etc). Of course, as in the ME, the learning of the experts and the gating network are intimately coupled.

Unfortunately, it may be (practically and statistically) difficult to infer the appropriate number of experts for a particular dataset. In the current paper we sidestep this difficult problem by using an infinite number of experts and employing a gating network related to the Dirichlet Process, to specify a spatially varying Dirichlet Process. An infinite number of experts may also in many cases be more faithful to our prior expectations about complex real-world datasets. Integrating over the posterior distribution for the parameters is carried out using a Markov Chain Monte Carlo approach.

Tresp [2001] presented an alternative approach to mixtures of GPs. In his approach both the $M$ experts and the gating network were implemented with GPs; the gating network being a softmax of $M$ GPs. Our new model avoids several limitations of the previous approach, which are covered in depth in the discussion.

## 2 Infinite GP mixtures

The traditional ME likelihood does not apply when the experts are non-parametric. This is because in a normal ME model the data is assumed to be iid given the model parameters:

$$p(y|x, \theta) = \prod_i \sum_j p(y_i = j | x_i, \theta_j) p(c_i = j | x_i, \phi),$$

where $x$ and $y$ are inputs and outputs (boldface denotes vectors), $\theta_j$ are the parameters of expert $j$, $\phi$ are the parameters of the gating network and $c_i$ are the discrete indicator variables assigning data points to experts.

This iid assumption is contrary to GP models which solely model the dependencies in the joint distribution (given the hyperparameters). There is a joint distribution corresponding to every possible assignment of data points to experts; therefore the likelihood is a sum over (exponentially many) assignments:

$$p(y|x, \theta) = \sum_c p(y|c, x, \theta) p(c|x, \phi)$$

$$= \sum_c \left[ \prod_j p(y_i = j | \{x_i : c_i = j\}, \theta_j) \right] p(c|x, \phi).$$

(1)

Given the configuration $c = (c_1, \ldots, c_n)$, the distribution factors into the product, over experts, of the joint Gaussian distribution of all data points assigned to each expert. Whereas the original ME formulation used expectations of assignment variables called responsibilities, this is inadequate for inference in the mixture of GP experts. Consequently, we directly represent the indicators, $c_i$, and Gibbs sample for them to capture their dependencies.

In Gibbs sampling we need the posterior conditional distribution for each indicator given all the remaining indicators and the data:

$$p(c_i = j | c_{-i}, x, y, \theta, \phi) \propto p(y | c_{-i} = j, x, \theta) p(c_i = j | x_{-i}, x, \phi),$$

where $c_{-i}$ denotes all indicators except number $i$. We defer discussion of the second term defining the gating network to the next section. As discussed, the first term being the likelihood given the indicators factors into independent terms for each expert. For Gibbs sampling we therefore need the probability of output $y_i$ under GP number $j$: $p(y_i | \{y_\ell : \ell \neq i, c_\ell = j\}, \{x_\ell : c_\ell = j\}, \theta_j)$. 

For a GP model, this conditional density is the well known Gaussian [Williams & Rasmussen, 1996]:

\[ p(y_i | y_{-i}, x, \theta) \sim N(\mu, \sigma^2), \quad \begin{cases} 
\mu = Q(x_i, x)^T Q^{-1} y_{-i} \\
\sigma^2 = Q(x_i, x_i) - Q(x_i, x)^T Q^{-1} Q(x_i, x) 
\end{cases} \tag{2} \]

where the covariance matrix \( Q \) depends on the parameters \( \theta \). Thus, for the GP expert, we compute the above conditional density by simply evaluating the GP on the data assigned to it. Although this equation looks computationally expensive, we can keep track of the inverse covariance matrices and reuse them for consecutive Gibbs updates by performing rank one updates (since Gibbs sampling changes at most one indicator at a time).

We are free to choose any valid covariance function for the experts. In our simulations we employed the following Gaussian covariance function:

\[ Q(x_i, x_{i'}) = \nu_0 \exp \left( -\frac{1}{2} \sum_d (x_{id} - x_{i'd})^2 / \omega_d^2 \right) + \nu_1 \delta(i, i') \tag{3} \]

with hyperparameters \( \nu_0 \) controlling the signal variance, \( \nu_1 \) controlling the noise variance, and \( \omega_d \) controlling the length scale or (inverse) relevance of the \( d \)-th dimension of \( x \) in relation to predicting \( y \); \( \delta \) is the Kronecker delta function (i.e. \( \delta(i, i') = 1 \) if \( i = i' \), o.w. 0).

### 3 The Gating network

The gating network assigns probability to different experts based entirely on the input. We will derive a gating network based on the Dirichlet Process which can be defined as the limit of a Dirichlet distribution when the number of classes tends to infinity. The standard Dirichlet Process is not input dependent, but we will modify it to serve as a gating network based on the Dirichlet Process which can be defined as the

\[ p(\pi_1, \ldots, \pi_n | \alpha) \sim \text{Dirichlet}(\alpha / k) = \frac{\Gamma(\alpha)}{\Gamma(\alpha / k)^k} \prod_j \pi_j^{\alpha / k - 1}, \]

where \( \alpha \) is the (positive) concentration parameter. It can be shown [Rasmussen, 2000] that the conditional probability of a single indicator when integrating over the \( \pi_j \) variables and letting \( k \) tend to infinity is given by:

\[ p(c_i = j | \pi_{-i}, \alpha) = \frac{n_{-i,j}}{n - 1 + \alpha} \]

\[ p(c_i \neq c_i \text{ for all } i' \neq i | \pi_{-i}, \alpha) = \frac{\alpha}{n - 1 + \alpha}, \tag{4} \]

where \( n_{-i,j} (= \sum_{i' \neq i} \delta(c_{i'}, j)) \) is the occupation number of expert \( j \) excluding observation \( i \), and \( n \) is the total number of data points. This shows that the probabilities are proportional to the occupation numbers. To make the gating network input dependent, we will simply employ a local estimate \(^1\) for this occupation number using a kernel classifier:

\[ n_{-i,j} = (n - 1) \frac{\sum_{i' \neq i} K_\phi(x_i, x_{i'}) \delta(c_{i'}, j)}{\sum_{i' \neq i} K_\phi(x_i, x_{i'})}, \tag{5} \]

where the delta function selects data points assigned to class \( j \), and \( K \) is the kernel function parametrized by \( \phi \). As an example we use a Gaussian kernel function:

\[ K_\phi(x_i, x_{i'}) = \exp \left( -\frac{1}{2} \sum_d (x_{id} - x_{i'd})^2 / \phi_d^2 \right), \tag{6} \]

\(^1\) this local estimate won’t generally be an integer, but this doesn’t have any adverse consequences
parameterized by length scales $\phi_d$ for each dimension. These length scales allow dimensions of $x$ space to be more or less relevant to the gating network classification.

We Gibbs sample from the indicator variables by multiplying the input-dependent Dirichlet process prior eq. (4) and (5) with the GP conditional density eq. (2). Gibbs sampling in an infinite model requires that the indicator variables can take on values that no other indicator variable has already taken, thereby creating new experts. We use the auxiliary variable approach of Neal [1998] (algorithm 8 in that paper). In this approach hyperparameters for new experts are sampled from their prior and the likelihood is evaluated based on these. This requires finding the likelihood of a Gaussian process with no data. Fortunately, for the covariance function eq. (3) this likelihood is Gaussian with zero mean and variance $v_0 + v_1$.

If all $n$ data points are assigned to a single GP, the likelihood calculation will still be cubic in the number of data points (per Gibbs sweep over all indicators). We can reduce the computational complexity by introducing the constraint that no GP expert can have more than $n_{\text{max}}$ data points assigned to it. This is easily implemented\(^2\) by modifying the conditionals in the Gibbs sampler.

The hyperparameter $\alpha$ controls the prior probability of assigning a data point to a new expert, and therefore influences the total number of experts used to model the data. As in Rasmussen [2000], we give a vague inverse gamma prior to $\alpha$, and sample from its posterior using Adaptive Rejection Sampling (ARS) [Gilks & Wild, 1992]. Allowing $\alpha$ to vary gives the model more freedom to infer the number of GPs to use for a particular dataset.

Finally we need to do inference for the parameters of the gating function. Given a set of indicator variables one could use standard methods from kernel classification to optimize the kernel widths in different directions. These methods typically optimize the leave-one-out pseudo-likelihood (ie the product of the conditionals), since computing the likelihood in a model defined purely from conditional distributions as in eq. (4), (5) & (6) is generally difficult (and as pointed out in the discussion section there may not even be a single likelihood). In our model we multiply the pseudo-likelihood by a (vague) prior and sample from the resulting pseudo-posterior.

\section{The Algorithm}

The individual GP experts are given a stationary Gaussian covariance function, with a single length scale per dimension, a signal variance and a noise variance, i.e. $D + 2$ (where $D$ is the dimension of the input) hyperparameters per expert, eq. (3). The signal and noise variances are given inverse gamma priors with hyper-hypers $a$ and $b$ (separately for the two variances). This serves to couple the hyperparameters between experts, and allows the priors on $v_0$ and $v_1$ (which are used when evaluating auxiliary classes) to adapt. Finally we give vague independent log normal priors to the length scale parameters $w$ and $\phi$.

The algorithm for learning an infinite mixture of GP experts consists of the following steps:

\begin{enumerate}
\item Initialize indicator variables $c_i$ to a single value (or a few values if individual GPs are to be kept small for computational reasons).
\item Do a Gibbs sampling sweep over all indicators.
\item Do Hybrid Monte Carlo (HMC) [Duane et al, 1987] for hyperparameters of the GP covariance function, $v_0, v_1, w_d$, for each expert in turn. We used 10 leapfrog iterations with a stepsize small enough that rejections were rare.
\item Optimize the hyper-hypers, $a$ & $b$, for each of the variance parameters.
\item Sample the Dirichlet process concentration parameter, $\alpha$ using ARS.
\end{enumerate}

\(^2\)We simply set the conditional probability of joining a class which has been deemed full to zero.