Gaussian process classification: a message-passing viewpoint

Filipe Rodrigues
fmpr@dei.uc.pt

November 2014

Abstract

The goal of this short paper is to provide a message-passing viewpoint of the Expectation Propagation (EP) algorithm commonly used for Gaussian process (GP) classification with probit likelihood. By presenting this EP algorithm as message-passing in the factor graph gives the reader a different (more unified) perspective on what the algorithm is doing and facilitates the use of GP classification (and regression) as a building block for larger factor graphs.

1 Problem setup and factor graph

Figure 1 shows the factor graph for Gaussian process classification with probit likelihood.

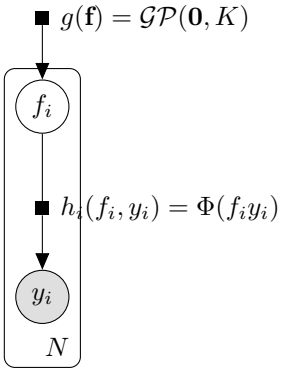

Figure 1: Factor graph for GP classification.

Given a dataset of N observations $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$, where $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ and $\mathbf{y} = \{y_i\}_{i=1}^N$, our goal is to estimate the posterior on $\mathbf{f}$

$$p(\mathbf{f}|\mathbf{X}, \mathbf{y}) = \frac{1}{Z} p(\mathbf{f}|\mathbf{X}) \prod_{i=1}^{N} p(y_i|f_i). \tag{1}$$

We will approximate the posterior on $\mathbf{f}$ by making use of the Expectation Propagation algorithm [3], which approximates each likelihood term $p(y_i|f_i)$ in turn with an unnormalised Gaussian on f_i

$$p(y_i|f_i) \approx \tilde{Z}_i \mathcal{N}(f_i|\tilde{\mu}_i, \tilde{\sigma}_i^2). \tag{2}$$

The approximate posterior on $\mathbf{f}$ is then given by

$$q(\mathbf{f}|\mathbf{X}, \mathbf{y}) = \frac{1}{Z_{EP}} p(\mathbf{f}|\mathbf{X}) \prod_{i=1}^{N} \mathcal{N}(f_i|\tilde{\mu}_i, \tilde{\sigma}_i^2) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \Sigma) \tag{3}$$

with

$$\boldsymbol{\mu} = \Sigma \tilde{\Sigma}^{-1} \tilde{\boldsymbol{\mu}} \tag{4}$$

$$\Sigma = (K^{-1} + \tilde{\Sigma}^{-1})^{-1} \tag{5}$$

where $\tilde{\boldsymbol{\mu}}$ is the vector of $\tilde{\mu}_i$ and $\tilde{\Sigma}$ is diagonal with $\tilde{\Sigma}_{ii} = \tilde{\sigma}_i^2$.

2 Message-passing for classification

We will now present a message-passing viewpoint of the typical EP algorithm presented for example in [1] and [2]. The main advantage of this is that is allows us to easily generalize GP classification (and regression) to be part of larger factor graphs.

The EP algorithm, as a message-passing algorithm in the factor graph of Figure 1, comprises the following steps:

Step 1: Compute message from the factor $g(\mathbf{f})$ to the f_i variables

$$m_{g \to f_i}(f_i) = \int g(\mathbf{f}) \prod_{j \neq i} m_{f_j \to g}(f_j) df_j \tag{6}$$

$$= \int p(\mathbf{f}|\mathbf{X}) \prod_{j \neq i} \mathcal{N}(f_j|\tilde{\mu}_j, \tilde{\sigma}_j^2) df_j \tag{7}$$

Conceptually, one can think of the combination of prior $g(\mathbf{f})$ and the $n-1$ approximate likelihoods in eq. 7 in two ways, either by explicitly multiplying out the terms, or (equivalently) by removing approximate likelihood i from the approximate posterior in eq. 3. Here we will follow the latter approach. The marginal for f_i from $q(\mathbf{f}|\mathbf{X}, \mathbf{y})$ is given by

$$q(f_i|\mathbf{X}, \mathbf{y}) = \mathcal{N}(f_i|\mu_i, \sigma_i^2) \tag{8}$$

The message from the factor $g(\mathbf{f})$ to the f_i variables is then given by

$$m_{g \to f_i}(f_i) = \frac{q(f_i|\mathbf{X}, \mathbf{y})}{m_{f_i \to g}(f_i)} = \frac{\mathcal{N}(f_i|\mu_i, \sigma_i^2)}{\mathcal{N}(f_i|\tilde{\mu}_i, \tilde{\sigma}_i^2)} = \mathcal{N}(f_i|\mu_{-i}, \sigma_{-i}^2) \tag{9}$$

where

$$\mu_{-i} = \sigma_{-i}^2 (\sigma_i^{-2} \mu_i - \tilde{\sigma}_i^{-2} \tilde{\mu}_i) \tag{10}$$

$$\sigma_{-i}^2 = (\sigma_i^{-2} - \tilde{\sigma}_i^{-2})^{-1} \tag{11}$$

where we made use of the property in A.1.

This step is therefore equivalent to the computation of the cavity distribution as presented in [1] and [2]. However, note that we need to make use of equations 4 and 5 in order to compute $\boldsymbol{\mu}$ and Σ, from which we get the values of μ_i and σ_i^2.

Step 2: Compute approximate posterior on f_i

$$q(f_i) = m_{g \to f_i}(f_i) m_{h_i \to f_i}(f_i) \tag{12}$$
$$= \mathcal{N}(f_i | \mu_{-i}, \sigma_{-i}^2) \Phi(f_i y_i) \tag{13}$$
$$\approx \hat{Z}_i \mathcal{N}(f_i | \hat{\mu}_i, \hat{\sigma}_i^2) \tag{14}$$

where the approximation is done by matching the moments of the two distributions [3]. The moments are then given by [1, 4]:

$$\hat{Z}_i = \Phi(z_i) \tag{15}$$
$$\hat{\mu}_i = \mu_{-i} + \frac{y_i \sigma_{-i}^2 \mathcal{N}(z_i)}{\Phi(z_i)\sqrt{1 + \sigma_{-i}^2}} \tag{16}$$
$$\hat{\sigma}_i^2 = \sigma_{-i}^2 - \sigma_{-i}^4 \frac{\mathcal{N}(z_i)}{(1 + \sigma_{-i}^2)\Phi(z_i)} \left(z_i + \frac{\mathcal{N}(z_i)}{\Phi(z_i)} \right) \tag{17}$$

where we defined

$$z_i = \frac{y_i \mu_{-i}}{\sqrt{1 + \sigma_{-i}^2}} \tag{18}$$

See [1] for the derivation of this moments.

Notice that we are combining the cavity distribution with the exact likelihood $p(y_i|f_i)$ to get the desired (non-Gaussian) marginal, which we then approximate with a Gaussian with moment matching. Hence, this step is equivalent to the computation of the site parameters as presented in [1] and [2].

Step 3: Compute message from f_i to the factor $g(\mathbf{f})$

$$m_{f_j \to g}(f_j) = \frac{q(f_i)}{m_{g \to f_i}(f_i)} = \frac{\mathcal{N}(f_i | \hat{\mu}_i, \hat{\sigma}_i^2)}{\mathcal{N}(f_i | \mu_{-i}, \sigma_{-i}^2)} = \tilde{Z}_i \mathcal{N}(f_i | \tilde{\mu}_i, \tilde{\sigma}_i^2) \tag{19}$$

with

$$\tilde{\mu}_i = \tilde{\sigma}_i^2 (\hat{\sigma}_i^{-2} \hat{\mu}_i - \sigma_{-i}^{-2} \mu_{-i}) \tag{20}$$
$$\tilde{\sigma}_i^2 = (\hat{\sigma}_i^{-2} - \sigma_{-i}^{-2})^{-1} \tag{21}$$
$$\tilde{Z}_i = \hat{Z}_i \sqrt{2\pi} \sqrt{\sigma_{-i}^2 + \tilde{\sigma}_i^2} \exp\left(\frac{1}{2} \frac{(\mu_{-i} - \tilde{\mu}_i)^2}{\sigma_{-i}^2 + \tilde{\sigma}_i^2} \right) \tag{22}$$

where we made use of the property in A.1.

Notice that this step corresponds to the final of EP as presented in [1] and [2], where we compute the parameters of the approximation $\mathcal{N}(f_i|\tilde{\mu}_i, \tilde{\sigma}_i^2)$ which achieves a match with the desired moments.

Figure 2 provides an overview of the message-passing algorithm on the factor graph.

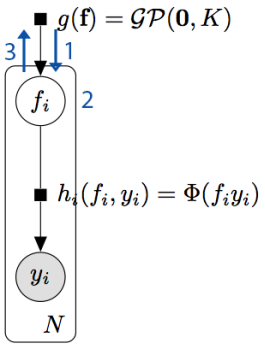

Figure 2: Overview of the message-passing algorithm.

3 Marginal likelihood

The EP approximation to the marginal likelihood Z_{EP} is given by

$$Z_{EP} = \int p(\mathbf{f}|\mathbf{X}) \prod_{i=1}^{N} \tilde{Z}_i \mathcal{N}(f_i|\tilde{\mu}_i, \tilde{\sigma}_i^2) d\mathbf{f} \tag{23}$$

$$= \int \mathcal{N}(\mathbf{f}|\mathbf{0}, K) \mathcal{N}(\mathbf{f}|\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}) d\mathbf{f} \prod_{i=1}^{N} \tilde{Z}_i \tag{24}$$

Making use of the results for the product of two Gaussians in A.1, we get

$$Z_{EP} = (2\pi)^{-D/2} |K + \tilde{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2} \tilde{\boldsymbol{\mu}}^T (K + \tilde{\Sigma})^{-1} \tilde{\boldsymbol{\mu}} \right) \prod_{i=1}^{N} \tilde{Z}_i \tag{25}$$

$$= (2\pi)^{-D/2} |K + \tilde{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2} \tilde{\boldsymbol{\mu}}^T (K + \tilde{\Sigma})^{-1} \tilde{\boldsymbol{\mu}} \right)$$
$$\times \prod_{i=1}^{N} \Phi\left(\frac{y_i \mu_{-i}}{\sqrt{1 + \sigma_{-i}^2}} \right) \sqrt{2\pi} \sqrt{\sigma_{-i}^2 + \tilde{\sigma}_i^2} \exp\left(\frac{1}{2} \frac{(\mu_{-i} - \tilde{\mu}_i)^2}{\sigma_{-i}^2 + \tilde{\sigma}_i^2} \right) \tag{26}$$

Taking the logarithm gives

$$\log Z_{EP} = -\frac{D}{2} \log(2\pi) - \frac{1}{2} \log |K + \tilde{\Sigma}| - \frac{1}{2} \tilde{\boldsymbol{\mu}}^T (K + \tilde{\Sigma})^{-1} \tilde{\boldsymbol{\mu}}$$
$$+ \sum_{i=1}^{N} \log \Phi\left(\frac{y_i \mu_{-i}}{\sqrt{1 + \sigma_{-i}^2}} \right) + \frac{N}{2} \log(2\pi) + \frac{1}{2} \sum_{i=1}^{N} \log(\sigma_{-i}^2 + \tilde{\sigma}_i^2) + \sum_{i=1}^{N} \frac{(\mu_{-i} - \tilde{\mu}_i)^2}{2(\sigma_{-i}^2 + \tilde{\sigma}_i^2)} \tag{27}$$

$$= -\frac{1}{2} \log |K + \tilde{\Sigma}| - \frac{1}{2} \tilde{\boldsymbol{\mu}}^T (K + \tilde{\Sigma})^{-1} \tilde{\boldsymbol{\mu}} + \sum_{i=1}^{N} \log \Phi\left(\frac{y_i \mu_{-i}}{\sqrt{1 + \sigma_{-i}^2}} \right)$$
$$+ \frac{1}{2} \sum_{i=1}^{N} \log(\sigma_{-i}^2 + \tilde{\sigma}_i^2) + \sum_{i=1}^{N} \frac{(\mu_{-i} - \tilde{\mu}_i)^2}{2(\sigma_{-i}^2 + \tilde{\sigma}_i^2)} + const. \tag{28}$$

4 Message-passing for regression

We will now take a quick look at the (simpler) case of GP regression with a Gaussian likelihood. This change corresponds to replacing the factor $h_i(f_i, y_i) = \Phi(f_i y_i)$ with $h_i(f_i, y_i) = \mathcal{N}(y_i|f_i, \sigma^2)$. Hence, we now have to revise step 2 of the EP algorithm presented in the previous section, which computed the (approximate) posterior on f_i. Now, this posterior is exact, and it is given by

$$q(f_i) = m_{g\to f_i}(f_i) m_{h_i\to f_i}(f_i) \tag{29}$$
$$= \mathcal{N}(f_i|\mu_{-i}, \sigma^2_{-i})\mathcal{N}(y_i|f_i, \sigma^2) \tag{30}$$
$$\propto \mathcal{N}(f_i|\hat{\mu}_i, \hat{\sigma}^2_i) \tag{31}$$

with

$$\hat{\mu}_i = \hat{\sigma}^2_i\Big(\sigma^{-2} y_i + \sigma^{-2}_{-i}\mu_{-i}\Big) \tag{32}$$
$$\hat{\sigma}^2_i = \Big(\sigma^{-2} + \sigma^{-2}_{-i}\Big)^{-1} \tag{33}$$

where we made use of the properties in A.2 to get the posterior on f_i.

An alternative way to arrive at these equations in by taking derivatives of the log partition function $\hat{Z}_i$, in order to compute the moments of the distributions by making use of the ADF updates provided by Minka in [4].

The marginal likelihood for the regression case is simply given by

$$Z_{EP} = \int p(\mathbf{f}|\mathbf{X}) \prod_{i=1}^{N} \mathcal{N}(y_i|f_i, \sigma^2) d\mathbf{f} \tag{34}$$
$$= \int \mathcal{N}(\mathbf{f}|\mathbf{0}, K) \prod_{i=1}^{N} \mathcal{N}(y_i|f_i, \sigma^2) d\mathbf{f} \tag{35}$$
$$\tag{36}$$

Notice that, in this case, the likelihood terms are already Gaussian. Hence, we can proceed by making use of the Gaussian identities in A.1, to evaluate the integral, and get

$$Z_{EP} = \mathcal{N}(\mathbf{y}|\mathbf{0}, K + \sigma^2 I) \tag{37}$$
$$\tag{38}$$

Taking the logarithm gives

$$\log Z_{EP} = -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|K + \sigma^2 I| - \frac{1}{2}\mathbf{y}^T(K + \sigma^2 I)^{-1}\mathbf{y} \tag{39}$$

A Operations with Gaussians

A.1 Product and division

Given two (multivariate) Gaussian distributions $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \Sigma_1)$ and $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \Sigma_2)$, the product is given by

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \Sigma_1)\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \Sigma_2) = Z^{-1}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) \tag{40}$$

where

$$\boldsymbol{\mu} = \Sigma(\Sigma_1^{-1}\boldsymbol{\mu}_1 + \Sigma_2^{-1}\boldsymbol{\mu}_2) \tag{41}$$
$$\Sigma = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}. \tag{42}$$

The normalization constant is given by

$$Z^{-1} = (2\pi)^{-D/2}|\Sigma_1 + \Sigma_2|^{-1/2} \exp\Big(-\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\Sigma_1 + \Sigma_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\Big) \tag{43}$$
$$= \sqrt{\frac{|\Sigma|}{(2\pi)^D|\Sigma_1||\Sigma_2|}} \exp\left(-\frac{1}{2}(\boldsymbol{\mu}_1^T\Sigma_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2^T\Sigma_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu})\right) \tag{44}$$

Similarly, for division we have

$$\frac{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \Sigma_1)}{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \Sigma_2)} = Z^{-1}\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) \tag{45}$$

where

$$\boldsymbol{\mu} = \Sigma(\Sigma_1^{-1}\boldsymbol{\mu}_1 - \Sigma_2^{-1}\boldsymbol{\mu}_2) \tag{46}$$
$$\Sigma = (\Sigma_1^{-1} - \Sigma_2^{-1})^{-1}. \tag{47}$$

The normalization constant is given by

$$Z^{-1} = \sqrt{\frac{|\Sigma||\Sigma_2|}{(2\pi)^D|\Sigma_1|}} \exp\left(-\frac{1}{2}(\boldsymbol{\mu}_1^T\Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T\Sigma_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu})\right) \tag{48}$$

A.2 Bayes rule

Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{49}$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{50}$$

the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T + \mathbf{L}^{-1}) \tag{51}$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\mathbf{S}\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \mathbf{S}) \tag{52}$$

where

$$\mathbf{S} = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}. \tag{53}$$

References

[1] Carl Edward Rasmussen and Christopher K. I. Williams. 2005. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press.

[2] Christopher M. Bishop. 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[3] Thomas Minka. 2001. A Family of Algorithms for Approximate Bayesian Inference. MIT Press.

[4] Thomas Minka. 2008. EP: A quick reference.