Gaussian Process Classification

Just as we could use the kernel trick to extend Bayesian linear regression to Gaussian processes for general-purpose nonlinear regression, we may also extend Bayesian linear classification in the same way. In Gaussian process classification, we assume there is a latent function $f\colon \mathcal{X} \to \mathbb{R}$ that is commensurate with the probability of a positive observation; higher latent function values correspond to higher probabilities of positive observations. In Bayesian linear classification, we assumed a parametric (linear) form for this latent function:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}.$$

In Gaussian process classification, rather than choosing a parametric form for f, we instead place a Gaussian process prior on f:

$$p(f) = \mathcal{GP}(f; \mu, K).$$

Note that a Gaussian prior on the weight vector $\mathbf{w}$ above induces a Gaussian process prior on f with mean function

$$\mu(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\mu}$$

and covariance function

$$K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \boldsymbol{\Sigma} \mathbf{x},$$

where $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The Gaussian process formalism allows us to model arbitrary nonlinear classification boundaries by using any desired mean and covariance function for f.

Likelihood

Suppose we have made binary observations at a set of values $\mathbf{X}$, and define $\mathbf{f} = f(\mathbf{X})$ to be the associated set of latent function values. As in Bayesian linear classification, we assume the following likelihood for a given binary observation y_i associated with $\mathbf{x}_i$:

$$p(y_i = 1 \mid f_i) = \sigma(f_i),$$

where $\sigma\colon \mathbb{R} \to (0, 1)$ is a monotonically increasing sigmoid function such as the logisitic function or the standard normal CDF. We again assume the observations are conditionally independent given the latent function values:

$$p(\mathbf{y} \mid \mathbf{f}) = \prod p(y_i \mid f_i).$$

Inference

Given our prior $p(f) = \mathcal{GP}(f; \mu, K)$ and a set of observations $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, we wish to find the posterior distribution of the latent function values $\mathbf{f} = f(\mathbf{X})$. Note that if we had a Gaussian posterior $\mathbf{f}$, this would induce a Gaussian process posterior for the function f given $\mathcal{D}$. The posterior is

$$p(\mathbf{f} \mid \mathcal{D}) = \frac{1}{Z} p(\mathbf{f} \mid \mathbf{X}) p(\mathbf{y} \mid \mathbf{f}) = \mathcal{N}\big(\mathbf{f}; \mu(\mathbf{X}), K(\mathbf{X}, \mathbf{X})\big) \prod_i p(y_i \mid f_i).$$

Unfortunately, the sigmoid likelihood coupled with the Gaussian prior do not couple to form a tractable posterior. Instead, we must approximate this posterior in some way. Previously we described the Laplace approximation, which approximates the unnormalized log posterior with a second-order Taylor expansion, resulting in a Gaussian approximate posterior centered at the posterior mode

$$\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} p(\mathbf{f} \mid \mathcal{D}).$$

Here we will consider two more general-purpose approximation techniques for approximating intractable posterior distributions. These techniques are useful when the likelihood factorizes into one-dimensional terms, as in GP classification.

Assumed Density Filtering

Consider a posterior distribution of the form

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{1}{Z} p_0(\boldsymbol{\theta}) \prod_{i=1}^{N} t_i(\boldsymbol{\theta}),$$

where the t_i are typically likelihood terms, for example our $p(y_i \mid f_i)$ above. We assume the prior $p_0(\boldsymbol{\theta})$ has been chosen to be some nice form, for example a Gaussian, and we will use the Gaussian case to illustrate the idea below.

In *assumed density filtering* (ADF), we assume that the posterior has the same form as the prior p_0, and we seek an approximating distribution

$$q(\boldsymbol{\theta}) \approx p(\boldsymbol{\theta} \mid \mathcal{D})$$

from the same family as p_0 that approximates the true posterior "as well as possible."

Mechanically, ensuring that our approximate distribution q remains in the same family as p_0 is achieved by selecting an (unnormalized) member $\tilde{t}_i$ from the likelihood conjugate to the prior for each of the likelihood terms t_i. The result is

$$q(\boldsymbol{\theta}) = p_0(\boldsymbol{\theta}) \prod_{i=1}^{N} \tilde{Z}_i \tilde{t}_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_i),$$

and this product will belong to the desired family by conjugacy. Here the constants $\tilde{Z}_i$ and the local parameter vectors $\{\tilde{\boldsymbol{\theta}}_i\}$ are free parameters, called *site parameters*, we may choose for each of the approximating distributions $\tilde{t}_i$ to try to improve the fit of the approximating distribution.

For example, the Gaussian distribution is self-conjugate, so the ADF approximation will in this case take the form

$$q(\boldsymbol{\theta}) = p_0(\boldsymbol{\theta}) \prod_{i=1}^{N} t_i(\boldsymbol{\theta}) \approx \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{i=1}^{N} \tilde{Z}_i \mathcal{N}(\boldsymbol{\theta}; \tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Sigma}}_i),$$

where the site parameters are the constants $\{\tilde{Z}_i\}$ (chosen so that the approximation normalizes) as well as the local mean vectors $\{\tilde{\boldsymbol{\mu}}_i\}$ and covariance matrices $\tilde{\boldsymbol{\Sigma}}_i$.

Note that in the GP classification case, the likelihood terms in the product are in fact one dimensional, because the likelihood for y_i only depends on f_i:

$$q(\mathbf{f}) = \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{i=1}^{N} \tilde{Z}_i \mathcal{N}(f_i; \tilde{\mu}_i, \tilde{\sigma}_i^2).$$

Consider just the first two terms in this product:

$$p(\boldsymbol{\theta} \mid \mathcal{D}_1) \propto p_0(\boldsymbol{\theta}) t_1(\boldsymbol{\theta}).$$

This product will not have the same nice form of the prior, but has only been warped "slightly" away from the prior via a single likelihood term. In many cases, this distribution will be at least partially manageable. Perhaps there will not be a nice closed expression, but we might still be able to compute the normalizing constant or the moments of the posterior.

In assumed density filtering, we will approximate this product with a member of the desired family:

$$q_1(\boldsymbol{\theta}) = \tilde{Z}_1 p_0(\boldsymbol{\theta}) \tilde{t}_1(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_1) \approx p(\boldsymbol{\theta} \mid \mathcal{D}_1) \propto p_0(\boldsymbol{\theta}) t_1(\boldsymbol{\theta}).$$

This approximation is done by matching the moments between the approximation $q_1(\boldsymbol{\theta})$ and the true posterior $p(\boldsymbol{\theta} \mid \mathcal{D}_1)$.

For example, consider $p_0(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. We select the site parameters

$$\tilde{Z}_1^{-1} \qquad \tilde{\boldsymbol{\mu}}_1 \qquad \tilde{\boldsymbol{\Sigma}}_1$$

such that

$$\begin{aligned} \int q_1(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} &= 1 \\ \mathbb{E}\big[q_1(\boldsymbol{\theta})\big] &= \mathbb{E}\big[p(\boldsymbol{\theta} \mid \mathcal{D}_1)\big] \\ \operatorname{cov}\big[q_1(\boldsymbol{\theta})\big] &= \operatorname{cov}\big[p(\boldsymbol{\theta} \mid \mathcal{D}_1)\big]. \end{aligned}$$

Now the approximate distribution is in the desired density family (Gaussians) and matches the true posterior up to the second moment.

Now consider the first three terms of the product:

$$p(\boldsymbol{\theta} \mid \mathcal{D}_1, \mathcal{D}_2) \propto p_0(\boldsymbol{\theta}) t_1(\boldsymbol{\theta}) t_2(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta} \mid \mathcal{D}_1) t_2(\boldsymbol{\theta}).$$

The idea in assumed density filtering is to substitute in our approximation $q_1(\boldsymbol{\theta})$, giving:

$$p(\boldsymbol{\theta} \mid \mathcal{D}_1, \mathcal{D}_2) \overset{\propto}{\approx} q_1(\boldsymbol{\theta}) t_2(\boldsymbol{\theta}).$$

Now we are in the same situation we were in before! We have a nice "prior" distribution q_1 multiplied by a single likelihood term t_2. We proceed as before, replacing the true likelihood term t_2 with an approximate (conjugate) term $\tilde{Z}_2 \tilde{t}_2(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_2)$, where we again choose the site parameters $(\tilde{Z}_2, \tilde{\boldsymbol{\theta}}_2)$ to match the moments between our new approximation

$$q_2(\boldsymbol{\theta}) = \tilde{Z}_2 q_1(\boldsymbol{\theta}) \tilde{t}_2(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_2) = \tilde{Z}_1 \tilde{Z}_2 p_0(\boldsymbol{\theta}) \tilde{t}_1(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_1) \tilde{t}_2(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_2).$$

and the "less-approximate" posterior $q_1(\boldsymbol{\theta}) t_2(\boldsymbol{\theta})$. We proceed in this fashion until we have processed all the local likelihood terms, resulting in the final approximation

$$q(\boldsymbol{\theta}) = p_0(\boldsymbol{\theta}) \prod_{i=1}^{N} \tilde{Z}_i \tilde{t}_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_i).$$

Expectation Propagation

One thing to note about assumed density filtering is that our final approximation is dependent on the sequence in which we process the likelihood terms $\{t_i\}$. Note that we are always matching the moments between

$$q_{i-1}(\boldsymbol{\theta}) t_i(\boldsymbol{\theta}),$$

the approximation using data up to the ith term as well as the true ith likelihood term, and the new approximation

$$q_i(\boldsymbol{\theta}) = \tilde{Z}_i q_{i-1}(\boldsymbol{\theta}) \tilde{t}_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_i).$$

This moment matching at time i therefore never considers data that appears in future terms. For this reason, we might accumulate errors and/or wish to later "revisit" a particular term and update the site parameters $(\tilde{Z}_i, \tilde{\boldsymbol{\theta}}_i)$ in light of future data. This idea leads to *expectation propagation,* a refinement of assumed density filtering that can address some of these issues.

The idea is simple. Once we have processed each of the likelihood terms, resulting in the approximation above

$$q(\boldsymbol{\theta}) = p_0(\boldsymbol{\theta}) \prod_{i=1}^{N} \tilde{Z}_i \tilde{t}_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_i),$$

we repeatedly revisit each term and update its site parameters. The mechanism for doing so is quite simple. First, we select a site $1 \leq i \leq N$ to update, then form the so-called *cavity distribution,* which is our approximation using all but the ith term in the product:

$$q_{-i}(\boldsymbol{\theta}) = p_0(\boldsymbol{\theta}) \prod_{j \neq i} \tilde{Z}_j \tilde{t}_j(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_j),$$

where we have divided by the old site term $\tilde{Z}_i \tilde{t}_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_i)$. Now we replace the removed site term with the true likelihood term t_i, forming the *tilted distribution*

$$q_{-i}(\boldsymbol{\theta}) t_i(\boldsymbol{\theta}).$$

Finally, we select new site parameters to (re)match the moments between the tilted distribution and our new approximation:

$$q_{\text{new}}(\boldsymbol{\theta}) = \tilde{Z}_i q_{-i}(\boldsymbol{\theta}) \tilde{t}_i(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}_i).$$

We proceed continually updating site parameters in this manner until we reach convergence (i.e., none of the site parameters changes very much) or we expend a chosen computational budget.

Theoretical motivation

We conclude with one brief note about the theoretical motivation behind the moment matching used in assumed density filtering and expectation propagation. The Kullback–Leibler (KL) divergence (also called *relative entropy*) is a notion of "distance" between probability distributions, defined by

$$d_{\text{KL}}\big(p(\boldsymbol{\theta}) \parallel q(\boldsymbol{\theta})\big) = \int p(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \, \mathrm{d}\boldsymbol{\theta}.$$

Notice that KL divergence is not a true distance, as it is not symmetric, but it does satisfy $d_{\text{KL}}(p \parallel q) \geq 0$ with $d_{\text{KL}}(p \parallel q) = 0$ if and only if $p(\boldsymbol{\theta}) = q(\boldsymbol{\theta})$ almost everywhere.

A well-known result is that the KL divergence between an arbitrary probability distribution $p(\boldsymbol{\theta})$ and a multivariate Gaussian distribution $q(\boldsymbol{\theta})$ (in the direction $d_{\text{KL}}(p \parallel q)$) is minimized when q is chosen to match the moments of p. Therefore, in the case of Gaussian approximations, these methods can be seen as iteratively building up an approximate posterior in the desired family by minimizing the KL divergence at every step.

Minimizing KL divergence in "the other direction," $d_{\text{KL}}(q \parallel p)$, gives rise to another family of approximation techniques known as *variational Bayesian inference.*