Sparse GP Regression

Suppose we have training data $\{\mathbf{x}_i, y_i\}_{i=1}^n$ where the $\mathbf{x}_i \in \mathbb{R}^d$ are drawn from $p(\mathbf{x})$ and each y_i is a noisy observation of a latent function $f : \mathbb{R}^d \to \mathbb{R}$ applied to $\mathbf{x}_i$. In other words, $y_i = f(\mathbf{x}_i) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

We wish to model these data with a Sparse Gaussian Process with N inducing inputs $\{\mathbf{z}_i\}_{i=1}^N$, where each $\mathbf{z}_i \in \mathbb{R}^d$. We suppose that appropriate hyperparameters θ (including σ^2) are known/fixed and therefore omit them where possible in subsequent notation.

Letting X, $\mathbf{y}$, and Z represent the collections of data and inducing variables, respectively, the posterior $p(Z|\mathbf{y}, X)$ may be expressed as follows:

$$p(Z|\mathbf{y}, X) = \frac{p(\mathbf{y}|Z, X)p(Z)}{p(\mathbf{y}|X)} \tag{1}$$

Sparse GP Regression: Notes

We notice that:

- $p(\mathbf{y}|X)$ is intractable
- $p(\mathbf{y}|Z, X)$ can be approximated by several methods or bounded from below using a variational approach (see next two slides).
- Thus $p(Z|\mathbf{y}, X)$ is an unknown, unnormalized distribution.

We'd like to approximate $p(Z|\mathbf{y}, X)$ with $q_\psi(Z)$. But really this should be for a single pseudo-input, not the collection. We parameterize $q(z)$ with ψ. Also remember it's unnormalized even before we apply the prior.

Approximate SGP Likelihoods: PP and FITC

In the projected process (PP) and fully independent training conditional (FITC) formulations, $p(\mathbf{y}|Z, X)$ is an approximation to the full GP likelihood $p(\mathbf{y}|X)$. The PP approximation takes the following form:

$$F_{PP} = \mathcal{N}(\mathbf{y} \mid 0, \sigma^2 I + K_{XZ} K_{ZZ}^{-1} K_{ZX}) \tag{2}$$

And the FITC approximation corrects the PP approximation to match the full GP covariance on the diagonal:

$$F_{FITC} = \mathcal{N}(y \mid 0, \sigma^2 I + \mathrm{diag}[K_{XX} - K_{XZ} K_{ZZ}^{-1} K_{ZX}] + K_{XZ} K_{ZZ}^{-1} K_{ZX}) \tag{3}$$

Variational Lower Bound on the Full GP Likelihood

In the variational formulation due to Titsias (2009), inducing inputs are selected to maximize the following lower bound on the full GP log-likelihood:

$$\log F_V = \log[\mathcal{N}(\mathbf{y}|0, \sigma^2 I + K_{XZ} K_{ZZ}^{-1} K_{ZX})] + \frac{1}{2\sigma^2} Tr[K_{XX} - K_{XZ} K_{ZZ}^{-1} K_{ZX}] \tag{4}$$

Selecting inducing inputs

- Typically a greedy selection procedure is used to sequentially select inducing inputs that provide the greatest increase in (2), (3), or (4).
- Alternatively, a set of N inducing inputs might be initialized to approximate $p(X)$, then jointly tuned to maximize (2), (3), or (4).
- However, both approaches focus on the likelihood rather than the posterior $p(Z|\mathbf{y}, X)$.
- The key idea: using this posterior (via adversarial learning) will:
    1. Ensure that inducing inputs are selected consistent with both the data X and the likelihood, leading to better predictive performance.
    2. Allow us specify an alternative prior over Z that is more appropriate for a specific, anticipated prediction task.

Adversarial Approach

Learning from unnormalized distribution: look to Chunyuan paper or AVB.

Neural Network Parameterization

Mixture model parameterized by NN.

Baseline Models

We are comparing selection of inducing variables by:

1. Greedy selection
2. Models