Notes on GVI in FS for Image Data

1. Gaussian Processes for Classification

Notes taken from chapter 4 from Matthews (2017).

For Gaussian process regression (GPR), a class of models is defined:

$$f \sim \mathcal{GP}(0, K(\theta))$$

where $f : X \to \mathbb{R}$, mapping to the set of real numbers $\mathbb{R}$ and K is the covariance function $K : X \times X \to \mathbb{R}$ parameterised by θ.

For binary Gaussian process classification (GPC), a mapping is defined:

$$g : \mathbb{R} \to [0, 1]$$

transforming a value on the real line to the unit interval to represent a probability. A bernoulli random variable $\mathcal{B}$ can be defined such that:

$$f_c \sim \mathcal{B}(g(f))$$

where $f_c : X \to \{0, 1\}$, the desired binary classifier.

For multiclass classification of J different classes, models are defined:

$$f^{(j)} \sim \mathcal{GP}(0, K(\theta^{(j)}))$$

where $j = 1, \ldots, J$, defining J i.i.d. Gaussian processes. Concatenating $\mathbf{f} = [f_1 \cdots f_J]^T$, the classification operation can be defined:

$$\mathbf{f}_c \sim Cat(\mathcal{S}(\mathbf{f}))$$

where $\mathbf{f}_c : X \to \{0, \ldots, J\}$, the desired multiclass classifier. $\mathcal{S} : \mathbb{R}^J \to \Delta(J)$, a mapping from a J dimensional real vector to a J dimensional probability simplex. Cat is the categorical distribution (generalisation of Bernoulli distribution for Categorical Data).

There are different possible choices for $\mathcal{S}$. The multiclass generalisation of the logit likelihood:

$$\mathcal{S}_{softmax}(\mathbf{f})_i = \frac{\exp(f^{(i)})}{\sum_{j=1}^{J} \exp(f^{(j)})}$$

The robust max function:

$$\mathcal{S}^{(\epsilon)}_{robust}(\mathbf{f})_i = \begin{cases} 1 - \epsilon, & \text{if } i = \arg\max(\mathbf{f}) \\ \epsilon, & \text{otherwise} \end{cases}$$

taking class label of the maximum value with probability of $1-\epsilon$ and probability ϵ of picking one of the other classes uniformly at random, where ϵ is chosen. This formulation provides robustness to outliers, as it only considers the ranking of the GPR models for each class.

A benefit of the robust max function is that the variational expectation is analytically tractable with respect to the normal CDF ($q(\mathbf{f}) = \mathcal{N}(\mu, C), \mathbf{f} \in \mathbb{R}^J$) and one dimensional quadrature ($\mathcal{S}^{(\epsilon)}_{robust}(\mathbf{f})_i \in \mathbb{R}$):

$$\int_{\mathbb{R}^J} q(\mathbf{f}) \log(\mathcal{S}^{(\epsilon)}_{robust}(\mathbf{f})_y) d\mathbf{f} = \log(1-\epsilon) S + \log\left(\frac{\epsilon}{J-1}\right)(1-S)$$

where S is the probability that the function value corresponding to observed class y is larger than the other function values at that point:

$$S = \mathbb{E}_{\mathbf{f}^{(y)} \sim \mathcal{N}(\mathbf{f}^{(y)}|\mu^{(y)}, C^{(y)})}\left[\prod_{i \neq y} \phi\left(\frac{\mathbf{f}^{(y)} - \mu^{(i)}}{\sqrt{C^{(i)}}}\right)\right]$$

where ϕ is the standard normal CDF. This one dimensional integral can be evaluated using Gauss-Hermite quadrature.

2. GWI for Multiclass Classification

Notes taken from A.6 from Wild et al. (2022).

2.1 Objective Function

The likelihood:

$$p(y|f_1, \dots, f_J) = \prod_{n=1}^{N} p(y_n|f_1, \dots, f_J)$$

where $p(y_n|f_1, \dots, f_J) \coloneqq h^{\epsilon}_{y_n}(f_1(x_n), \dots, f_J x_N)$ and $y_n \in \{1, \dots, J\}$. $h^{\epsilon}_{y_n}$ is the robust max function $\mathcal{S}^{(\epsilon)}_{robust}$ as described in Matthews (2017). Wild et al. (2022) used $\epsilon = 1\%$.

The model consists of J independent Gaussian Random Elements such that:

$$f_j \sim P_j = \mathcal{N}(m_{\mathbb{P},j}, C_{\mathbb{P},j})$$

with the corresponding variational measures:

$$Q_j = \mathcal{N}(m_{\mathbb{Q},j}, C_{\mathbb{Q},j})$$

The objective to minimise:

$$\mathcal{L} = -\mathbb{E}_{\mathbb{Q}}\left[\log p(y_n|F_1, \dots, F_J)\right] + \sum_{j=1}^{J} W_2^2(P_j, Q_j)$$

The variational (posterior) approximation of the probability of $\{(F_1(x), \dots, F_J(x)) \in A\}$ will be denoted:

$$\mathbb{Q}\left((F_1(x), \dots, F_J(x)) \in A\right)$$

where $A \subset \mathbb{R}^J$. We get the expected log-likelihood:

$$\mathbb{E}_{\mathbb{Q}}\left[\log p(y_n|F_1,\ldots,F_J)\right] \approx \sum_{n=1}^{N} \log(1-\epsilon)S(x_n,y_n) + \log\left(\frac{\epsilon}{J-1}\right)(1-S(x_n,y_n))$$

where:

$$S(x,j) \coloneqq \frac{1}{\sqrt{\pi}}\sum_{i=1}^{I} w_i \prod_{l\neq j} \phi\left(\frac{\sqrt{2r_j(x,x)}\xi_i + m_{Q,j}(x) - m_{Q,l}(x)}{\sqrt{r_l(x,x)}}\right)$$

for any $x \in \mathcal{X}$, $j = 1,\ldots,J$ where $(w_i,\xi_i)_{i=1}^{I}$ are the weights and roots of the Hermite polynomial of order $I \in \mathbb{N}$., calculated with `scipy.special.roots_hermite`. ϕ is the standard normal cumulative distribution function.

The Wasserstein distance $W_2^2(P_j,Q_j)$ can be estimated in the same way as for regression:

$$\begin{aligned}\hat{W}^2 \coloneqq &\frac{1}{N}\sum_{n=1}^{N}(m_{\mathbb{P}}(x_n) - m_{\mathbb{Q}}(x_n))^2 + \frac{1}{N}\sum_{n=1}^{N} k(x_n,x_n)\\ &+ \frac{1}{N}\sum_{n=1}^{N} r(x_n,x_n) - \frac{2}{\sqrt{NN_S}}\sum_{s=1}^{N_S}\sqrt{\lambda_s(r(X_S,X)k(X,X_S))}\end{aligned}$$

where:

- $X_S \coloneqq (x_{S,1},\ldots,x_{S,N_S})$ with $x_{S,1},\ldots,x_{S,N_S} \in \mathbb{R}^D$, a set of N_S points sub-sampled from the input data X
- $r(X_S,X) \coloneqq (r(x_{S_s,x_n}))_{s,n} \in \mathbb{R}^{N_s\times N}$
- $k(X,X_S) \coloneqq (k(x_{x_n,S_s}))_{n,s} \in \mathbb{R}^{N\times N_s}$
- $\lambda_s(\cdot)$ calculates the s-th eigenvalue

and $n = 1,\ldots,N$, $s = 1,\ldots,N_S$, k is the kernel for $\mathbb{P}$, r is the kernel for $\mathbb{Q}$

2.2 Prediction

For an unseen point $x^* \in \mathcal{X}$, the probability that it belongs to class $j \in \{1,\ldots,J\}$:

$$\mathbb{Q}(Y^* = j) = (1-\epsilon)S(x^*,j) + \frac{\epsilon}{J-1}(1-S(x^*,j))$$

where the predicted label class is the maximiser of this probability:

$$Cat(\mathbb{Q}(Y^*)) = \underset{j\in\{1,\ldots J\}}{\arg\max}\, \mathbb{Q}(Y^* = j)$$

References

Alexander Graeme de Garis Matthews. *Scalable Gaussian process inference using variational methods.* PhD thesis, University of Cambridge, 2017.

Veit D Wild, Robert Hu, and Dino Sejdinovic. Generalized variational inference in function spaces: Gaussian measures meet bayesian deep learning. *arXiv preprint arXiv:2205.06342*, 2022.