ϵ-Gaussian Process Regression

Arnold Salas

Dept. of Engineering Science,
University of Oxford

arnold.salas@eng.ox.ac.uk

October 1, 2015

Support Vector Regression (SVR)

- $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N} \subset \mathbb{R}^D \times \mathbb{R}$
- SVR finds a model $\mathbf{w}$ such that

$$\min_{\mathbf{w}} \quad C \sum_{n=1}^{N} \left| y_n - \mathbf{w}^\top \phi\left(\mathbf{x}_n\right) \right|_\epsilon + \frac{1}{2}\mathbf{w}^\top \mathbf{w} \tag{1}$$

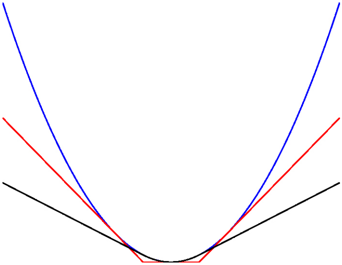
- $\phi(\cdot)$ is a feature map, $C > 0$ is the regularisation parameter and

$$|\eta|_\epsilon \equiv \max\left(|\eta| - \epsilon,\, 0\right) \qquad (\epsilon \geq 0) \tag{2}$$

is the ϵ-insensitive loss function (ILF)

Main features

- Sparsity
- Outlier robustness



MAP interpretation

Solving the SVR objective is equivalent to finding the MAP estimate of **w**, whose posterior distribution is

$$p\left(\mathbf{w}|\mathbf{y}, C, \epsilon\right) \propto \prod_{n=1}^{N} p\left(y_n|\mathbf{w}, C, \epsilon\right) p\left(\mathbf{w}\right), \tag{3}$$

with $\mathbf{y} \equiv (y_1, \ldots, y_N)^\top$ and

$$p\left(y_n|\mathbf{w}, C, \epsilon\right) \equiv \mathcal{Z} \exp\left[-C \left|y_n - \mathbf{w}^\top \phi\left(\mathbf{x}_n\right)\right|_\epsilon\right] \tag{4}$$

$$p\left(\mathbf{w}\right) \propto \exp\left(-\frac{1}{2}\mathbf{w}^\top \mathbf{w}\right), \tag{5}$$

where $\mathcal{Z} \equiv C/[2(1 + C\epsilon)]$ is a normalising factor

Non-parametric extension: ϵ-GPR

Instead of $\mathbf{w}^\top \phi(\mathbf{x})$, we consider a non-parametric prediction function with a GP prior

$$f(\mathbf{x})\,|\boldsymbol{\theta} \sim \mathcal{GP}\left(m(\mathbf{x}),\ k\left(\mathbf{x}, \mathbf{x}'|\boldsymbol{\theta}\right)\right) \tag{6}$$

BUT with a *non-Gaussian* observation model:

$$p\left(y_n|f_n, C, \epsilon\right) = \mathcal{Z} \exp\left(-C\left|y_n - f_n\right|_\epsilon\right), \tag{7}$$

where $f_n \equiv f(\mathbf{x}_n)$

Inference is challenging!

The conditional posterior

$$p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta}, \psi) = \frac{\prod_{n=1}^{N} p(y_n|f_n, C, \epsilon)\, p(\mathbf{f}|\boldsymbol{\theta})}{p(\mathbf{y}|\boldsymbol{\theta}, \psi)}, \tag{8}$$

where $\psi \equiv (C, \epsilon)^\top$, cannot be evaluated in closed form, since the marginal likelihood

$$p(\mathbf{y}|\boldsymbol{\theta}, \psi) = \int \prod_{n=1}^{N} p(y_n|f_n, C, \epsilon)\, p(\mathbf{f}|\boldsymbol{\theta})\, \mathrm{d}\mathbf{f} \tag{9}$$

is analytically intractable

Analogy with Student-t observation model

- same intractability issue
- common approach: use Gaussian scale-mixture representation of Student-t distribution $\Rightarrow$ enables Gibbs MCMC and mean-field VB
- similar approach may be taken for the ϵ-GPR observation model b/c it can be expressed as a continuous mixture of Gaussians (cMoG)

Mixture representation

Theorem

The ϵ-GPR observation model may be represented as a cMoG. Specifically:

$$p\left(y_n|f_n, C, \epsilon\right) = \int\int \mathcal{N}\left(y_n|f_n + \mu_n,\ \sigma_n^2\right) p(\mu_n|\epsilon) p(\sigma_n^2|\mu_n, C, \epsilon) \mathrm{d}\mu_n \mathrm{d}\sigma_n^2, \tag{10}$$

with

$$p\left(\mu_n|\epsilon\right) \equiv \mathcal{U}\left(\mu_n|-\epsilon, \epsilon\right), \quad p\left(\sigma_n^2|\mu_n, C, \epsilon\right) \equiv \mathcal{GIG}\left(\sigma_n^2|3/2,\ \chi_n,\ \omega_n\right), \tag{11}$$

where $\chi_n \equiv \epsilon^2(1 - \mu_n^2/\epsilon^2)$, $\omega_n \equiv (C\epsilon)^2/\chi_n$ *and*

$$\mathcal{GIG}\left(x|\nu, \chi, \omega\right) = \frac{(\omega/\chi)^{\nu/2}}{2K_\nu\left(\sqrt{\chi\omega}\right)} x^{\nu-1} \exp\left(-\frac{\chi x^{-1} + \omega x}{2}\right) \tag{12}$$

is the density of the Generalised Inverse Gaussian distribution.

Proof (sketch)

The proof relies on the following identities:

1. $\max(|u| - \epsilon, 0) = \max(u - \epsilon, 0) + \max(-u - \epsilon, 0)$
2. $2\max(u, 0) = |u| + u$
3. $\exp(-|u| - u) = \int \mathcal{N}(u| - \lambda, \lambda)\mathrm{d}\lambda$


THANK YOU
FOR
YOUR
ATTENTION!
ANY QUESTIONS?