Computational Statistics
Gaussian Processes

Luca Bortolussi

Department of Mathematics and Geosciences
University of Trieste

Office 238, third floor, H2bis
luca@dmi.units.it

Trieste, Winter Semester 2015/2016

Outline

1. Random functions and Bayesian regression
2. Gaussian Processes
3. Kernel functions
4. Hyperparameters
5. GP classification

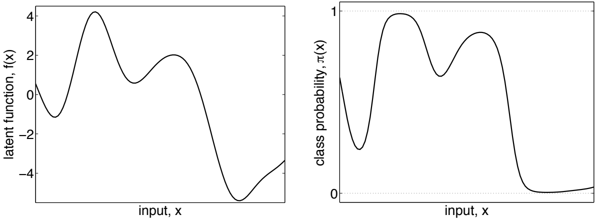
From logistic regression to GP classification

- The idea behind GP classification is to extend logistic (or probit) regression, by assuming the following model for the class conditionals:

LOGIT / PROBIT

$$\pi(\mathbf{x}) = \boxed{p(C_1|\mathbf{x})} = \sigma(f(\mathbf{x})) \text{ where } f \sim GP(\mu, k)$$

- f is often call latent function. Note that π is a random function, as f is.



(a) (b)

Figure 3.2: Panel (a) shows a sample latent function $f(x)$ drawn from a Gaussian process as a function of x. Panel (b) shows the result of squashing this sample function through the logistic logit function, $\lambda(z) = (1 + \exp(-z))^{-1}$ to obtain the class probability $\pi(x) = \lambda(f(x))$.

GP CLASSIFICATION

- f is often call latent or nuisance function. It is not observed directly. We only observe at a point $\mathbf{x}$ the realisation of a Bernoulli random variable with probability $\pi(\mathbf{x})$.
- Inference at a test point $\mathbf{x}^*$ is done, as usual in a Bayesian setting, in two steps:
  1. Compute the posterior f^* of f at the prediction point $\mathbf{x}^*$.

APPROXIMATE WITH A GAUSSIAN

$$p(f_*|X,\mathbf{y},\mathbf{x}_*) = \int p(f_*|X,\mathbf{x}_*,\mathbf{f})p(\mathbf{f}|X,\mathbf{y})\,d\mathbf{f}, \tag{3.9}$$

  with $p(\mathbf{f}|X,\mathbf{y}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X)/p(\mathbf{y}/X)$ by Bayes theorem.

  2. Compute the predictive distribution at $\mathbf{x}^*$

$$\bar{\pi}_* \triangleq p(y_*=+1|X,\mathbf{y},\mathbf{x}_*) = \int \sigma(f_*)p(f_*|X,\mathbf{y},\mathbf{x}_*)\,df_*. \tag{3.10}$$

Laplace Approximation

- As in Bayesian logistic regression, the computation of the posterior $p(\mathbf{f}|X, \mathbf{y})$ cannot be carried out analytically.
- However, we can do a Laplace approximation of the posterior around the MAP $\hat{f}$. The unnormalised log posterior is:

$$\begin{aligned}\Psi(\mathbf{f}) &\triangleq \log p(\mathbf{y}|\mathbf{f}) + \log p(\mathbf{f}|X) \\ &= \log p(\mathbf{y}|\mathbf{f}) - \frac{1}{2}\mathbf{f}^\top K^{-1}\mathbf{f} - \frac{1}{2}\log|K| - \frac{n}{2}\log 2\pi.\end{aligned} \tag{3.12}$$

Differentiating eq. (3.12) w.r.t. $\mathbf{f}$ we obtain

$$\nabla\Psi(\mathbf{f}) = \nabla \log p(\mathbf{y}|\mathbf{f}) - K^{-1}\mathbf{f}, \tag{3.13}$$

$$\nabla\nabla\Psi(\mathbf{f}) = \nabla\nabla \log p(\mathbf{y}|\mathbf{f}) - K^{-1} = -W - K^{-1}, \tag{3.14}$$

where W is diagonal, as observations are i.i.d.

- It can be optimised with a Newton-Rapson scheme:

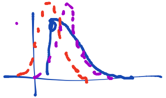
$$\begin{aligned}\mathbf{f}^{\text{new}} = \mathbf{f} - (\nabla\nabla\Psi)^{-1}\nabla\Psi &= \mathbf{f} + (K^{-1} + W)^{-1}(\nabla \log p(\mathbf{y}|\mathbf{f}) - K^{-1}\mathbf{f}) \\ &= (K^{-1} + W)^{-1}\big(W\mathbf{f} + \nabla \log p(\mathbf{y}|\mathbf{f})\big).\end{aligned} \tag{3.18}$$

LAPLACE APPROXIMATION

$p(f^* | X, y, x_*)$ is GAUSSIAN

- The Laplace approximation around the MAP $\hat{f}$ is a Gaussian q with mean

$$\mathbb{E}_q[f_*|X,\mathbf{y},\mathbf{x}_*] = \boxed{\mathbf{k}(\mathbf{x}_*)^\top K^{-1}\hat{\mathbf{f}}} = \mathbf{k}(\mathbf{x}_*)^\top \nabla \log p(\mathbf{y}|\hat{\mathbf{f}}). \qquad (3.21)$$



and variance

$$\begin{aligned}\mathbb{V}_q[f_*|X,\mathbf{y},\mathbf{x}_*] &= k(\mathbf{x}_*,\mathbf{x}_*) - \mathbf{k}_*^\top K^{-1}\mathbf{k}_* + \mathbf{k}_*^\top K^{-1}(K^{-1}+W)^{-1}K^{-1}\mathbf{k}_* \\ &= k(\mathbf{x}_*,\mathbf{x}_*) - \mathbf{k}_*^\top (K + W^{-1})^{-1}\mathbf{k}_*, \qquad (3.24)\end{aligned}$$

- The prediction π^* can be computed by the integral

$$\bar{\pi}_* \simeq \mathbb{E}_q[\pi_*|X,\mathbf{y},\mathbf{x}_*] = \int \sigma(f_*)\overbrace{q(f_*|X,\mathbf{y},\mathbf{x}_*)}^{\mathcal{N}}\, df_*, \qquad (3.25)$$

$\psi(\pi|x_*)$

which can be approximated with the same logit-probit-logit trick used for Bayesian logistic regression.

EXPECTATION PROPAGATION

$0 \leq KL[p,q] = \int p \log \frac{p}{q}$

- A (better) alternative to Laplace approximation is to use a variational method, typically for the probit activation function.
- A first option is to approximate the posterior distribution by a Gaussian q, minimising the (reversed) KL divergence $KL(q(\mathbf{f}|X, \mathbf{y}), p(\mathbf{f}|X, \mathbf{y}))$ (the minimisation of the KL divergence $KL(p(\mathbf{f}|X, \mathbf{y}), q(\mathbf{f}|X, \mathbf{y}))$ is intractable).
- Alternatively, one can use the Expectation Propagation algorithm, which constructs iteratively (over obs i, until convergence) a Gaussian approximation of the posterior by
  1. taking the current Gaussian approximation and factoring out the term for the i-th likelihood $p(y_i|f_i)$, obtaining a distribution for all observations but the i-th one.
  2. multiplying the cavity by the exact likelihood of the i-th observation, and finding a Gaussian approximation by moment matching of such a (non-Gaussian) distribution.
- EP is more accurate than Laplace approximation, and provides also an approximation of the Marginal likelihood.

PITFALLS OF GP PREDICTION

- Addition of a new observation *always* reduces uncertainty at all points $\rightarrow$ vulnerable to outliers
- Optimisation of hyperparameters often tricky: works well if σ^2 is known, otherwise it can be seriously multimodal
- **MAIN PROBLEM: GP prediction relies on a matrix inversion which scales cubically with the number of points!**
- Sparsification methods have been proposed but in high dimension GP regression is likely to be tricky nevertheless