

Regularization in Gaussian Processes

Risi Kondor

August 10, 2003

Let \mathcal{X} be a countable or uncountable set and let \mathcal{G} be a corresponding set of real valued random variables $\{Y_x\}$ indexed by $x \in \mathcal{X}$. \mathcal{G} is said to form a **Gaussian Process** if for any finite sample $x_1, x_2, \dots, x_m \in \mathcal{X}$, the random variables $Y_{x_1}, Y_{x_2}, \dots, Y_{x_m}$ are jointly Gaussian distributed. Similarly to finite collections of jointly Gaussian distributed variables, \mathcal{G} is completely specified by its mean $\mu(x) = E[Y_x]$ and covariance function $K(x, x') = \text{Cov}[Y_x, Y_{x'}]$. Conversely, any $\mu : \mathcal{X} \mapsto \mathbb{R}$ and any symmetric and positive definite $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ gives rise to a corresponding Gaussian Process over \mathcal{X} . Recall that K is said to be positive definite, if for any $m \in \mathbb{N}$, any x_1, x_2, \dots, x_m and any coefficients c_1, c_2, \dots, c_m not all zero, Definition

$$\sum_{i=1}^m \sum_{j=1}^m c_i c_j K(x_i, x_j) > 0.$$

In the following K will be called the **kernel**, and whenever we refer to a kernel, we shall imply symmetry and positive definiteness. An important example of a kernel is the **Gaussian kernel** on $\mathcal{X} = \mathbb{R}^N$ (not to be confused with the “Gaussian” in “Gaussian Processes”)

$$K(x, x') = e^{-\|x-x'\|^2/(2\sigma_G^2)} \tag{1}$$

with length scale parameter σ .

Another way to regard \mathcal{G} is as a distribution over functions $f : \mathcal{X} \mapsto \mathbb{R}$. RKHS view
To specify $p(f)$ we need to construct an appropriate space of functions from which to choose f and endow it with a measure. To this end, we consider the linear space of functions $k_x = K(x, \cdot)$ and define an inner product by $\langle k_x, k_{x'} \rangle = K(x, x')$. After adjoining the limits of Cauchy sequences with respect to the corresponding norm $\|f\| = \langle f, f \rangle$, the resulting space becomes a Hilbert space of functions, denoted \mathcal{H} . For any linear combination $f = \sum c_i k_{x_i}$,

$$\langle f, k_x \rangle = \left\langle \sum c_i k_{x_i}, k_x \right\rangle = \sum c_i K(x_i, x) = f(x)$$

giving rise to the surprising-looking but important property $\langle f, k_x \rangle = f(x)$ for any $f \in \mathcal{H}$. \mathcal{H} is referred to as the **Reproducing Kernel Hilbert Space** (RKHS) associated with K . Also note that by appropriate choice of K , \mathcal{H} can be made an almost arbitrarily rich space of functions. In particular, for the Gaussian kernel, \mathcal{H} can be shown to be a dense subset of $L_2(\mathcal{X})$.

Without loss of generality assuming zero mean, we now assert

$$p(f) = e^{-\langle f, f \rangle / 2}. \quad (2)$$

Equivalence of two views

To show that this is equivalent to the \mathcal{G} we had before, we need to show that it reproduces the marginals, i.e. that for any x_1, x_2, \dots, x_m and $t = (t_1, t_2, \dots, t_m)^\top$

$$p(f(x_1) = t_1, \dots, f(x_m) = t_m) = \frac{1}{(2\pi)^{m/2} |K|^{1/2}} \exp\left(-\frac{1}{2} t^\top K^{-1} t\right) \quad (3)$$

where K is the so-called **Gram matrix** with elements $K_{ij} = K(x_i, x_j)$. To this end we decompose \mathcal{H} into $V = \text{span}\{k_{x_1}, k_{x_2}, \dots, k_{x_m}\}$ and its orthogonal complement V^\perp and note that any f obeying $f(x_1) = t_1, \dots, f(x_m) = t_m$ may be written as

$$f = f_V + f_\perp = \left(\sum_{i=1}^m \alpha_i k_{x_i}\right) + f_\perp$$

with $f_\perp \in V^\perp$. The vector of coefficients $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_m)^\top$ can be found from

$$f(x_j) = \sum_{i=1}^m \alpha_i k_{x_i}(x_j) = \sum_{i=1}^m \alpha_i K(x_i, x_j) = t_j$$

leading to the matrix equation $K\alpha = t$. We can marginalize p to V just as in the finite dimensional case by

$$p_V(f_V) = p(f(x_1) = t_1, \dots) = \int_{V^\perp} p(f_V + f_\perp) df_\perp = e^{-\langle f_V, f_V \rangle / 2} \quad (4)$$

and expand

$$\langle f_V, f_V \rangle = \sum_{i=1}^m \sum_{j=1}^m [K^{-1}t]_i [K^{-1}t]_j \langle k_{x_i}, k_{x_j} \rangle = t^\top K^{-1} K K^{-1} t = t^\top K^{-1} t$$

proving (3).

In the following sections we will make extensive use of Dirac's bra-ket notation. Functions $f \in \mathcal{H}$ will be denoted by "kets" $|f\rangle$ and members of the dual-space by the "bras" $\langle f|$. Hence the inner product becomes $\langle f, g \rangle = \langle f|g\rangle$. In contrast, linear combinations of the form $\sum_{ij} |f_i\rangle F_{i,j} \langle f_j|$ for some matrix F , or more generally, $\int \int dx dx' |f_x\rangle F(x, x') \langle f_{x'}|$ where F is now a function are bilinear operators on \mathcal{H} .

Dirac Notation

Inference with Gaussian Processes

For inference Gaussian Processes are used as a Bayesian tool for estimating functions $f: \mathcal{X} \mapsto \mathbb{R}$ from observations $D = ((x_1, y_1), (x_2, y_2), \dots, (x_m, y_m))$ with $x_i \in \mathcal{X}$ and $y_i \in \mathbb{R}$. For notational simplicity assuming that the mean is zero, we assert a prior over functions

$$p(f) = e^{-\langle f|f\rangle / 2}$$

by specifying the kernel (covariance function) corresponding to the inner product. The noise model for the data is typically also Gaussian:

$$p(y|x, f) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-f(x))^2/(2\sigma^2)}.$$

The posterior can then be computed via Bayes rule

Computing
the posterior

$$p(f|D) = \frac{p(D|f)p(f)}{\int p(D|f)p(f)df} \quad (5)$$

with $p(D|f) = \prod_{i=1}^m p(y_i|x_i, f)$. Since all distributions here are Gaussian, we need not worry about normalizing factors and we can write the log posterior straight away as

$$\log p(f|D) \propto \frac{1}{2} \langle f|f \rangle + \frac{1}{2\sigma^2} \sum_{i=1}^m (f(x_i) - y_i)^2.$$

At this point the reproducing property of \mathcal{H} becomes crucial, allowing us to write

$$\log p(f|D) \propto \frac{1}{2} \langle f|f \rangle + \frac{1}{2\sigma_0^2} \sum (\langle f|k_{x_i} \rangle - y_i)^2 \quad (6)$$

and to complete the square

$$2 \log p(f|D) \propto \left[\langle f| - \langle u|\hat{S} \right] \hat{S}^{-1} \left[|f\rangle - \hat{S}|u\rangle \right]$$

with

$$u = \frac{1}{\sigma^2} \sum y_i k_{x_i} \quad \text{and} \quad \hat{S}^{-1} = I + \frac{1}{\sigma^2} \sum |k_{x_i}\rangle \langle k_{x_i}|.$$

To express \hat{S} , we extend k_{x_1}, k_{x_2} to a countable basis of \mathcal{H} and define the matrix S^{-1} with elements

$$S_{i,j}^{-1} = \langle k_{x_i}|\hat{S}^{-1}|k_{x_j}\rangle = K(I + \frac{1}{\sigma^2}K).$$

This matrix can readily be inverted with ordinary matrix algebra to give another matrix $S = K^{-1}(I + \frac{1}{\sigma^2}K)^{-1}$. Writing

$$\sum_i \sum_j \langle k_{x_a}|k_{x_i}\rangle S_{i,j} \langle k_{x_j}|\hat{S}^{-1}|k_{x_b}\rangle = \sum_i \langle k_{x_a}|k_{x_i}\rangle [SS^{-1}]_{i,b} = \langle k_{x_a}|k_{x_b}\rangle \quad (7)$$

then shows that

$$\hat{S} = \sum_i \sum_j |k_{x_i}\rangle S_{i,j} \langle k_{x_j}|$$

is the correct operator inverse of \hat{S}^{-1} .

We can now read off the posterior mean

$$\mathbb{E}[|f\rangle] = \hat{S}|u\rangle = \frac{1}{\sigma^2} \sum_{i=1}^m \sum_{j=1}^m \sum_{l=1}^m |k_{x_i}\rangle S_{i,j} \langle k_{x_j}|k_{x_l}\rangle y_l = \sum_{i=1}^m |k_{x_i}\rangle [(\sigma^2 I + K)^{-1}y]_i$$

and the posterior variance

$$\text{Var}[[f]] = \hat{S} = \sum_i \sum_j |k_{x_i}\rangle S_{i,j} \langle k_{x_j}|.$$

It is sometimes helpful to describe the posterior in terms of the mean and variance of $f(x)$ for fixed x it induces. Note that this is not the whole story, though: the posterior will also have a new covariance structure (effectively a new K) which this does not shed light on. Computing the mean is easy:

Pointwise
mean and
variance

$$\begin{aligned} \text{E}[f(x)] &= \text{E}[\langle k_x | f \rangle] = \langle k_x | \text{E}[[f]] \rangle = \\ &= \sum_{i=1}^m \langle k_x | k_{x_i} \rangle [(\sigma^2 I + K)^{-1} y]_i = k^\top (\sigma^2 I + K)^{-1} y \end{aligned} \quad (8)$$

with $k = (K(x, x_1), K(x, x_2), \dots, K(x, x_m))^\top$. Computing the variance requires somewhat more work, since now that the inner product and covariance do not match any more, marginalization is not quite as simple as applying (4) to a one dimensional subspace. Instead, we opt to take advantage of the property that any finite subset of variables is jointly Gaussian distributed. In particular, the distribution over $y_x^* = (y_1, y_2, \dots, y_m, f(x))$ is

$$p(y^*) \propto e^{-y^{*\top} \overline{K}^{-1} y^*} \quad (9)$$

with covariance matrix

$$\overline{K}^* = \left[\begin{array}{c|c} K + \sigma^2 I & k \\ \hline k^\top & \kappa \end{array} \right]$$

where $\kappa = K(x, x)$. Given y_1, y_2, \dots, y_m , the variance of $f(x)$ can be read off from (9) as $(K_{m+1, m+1}^{*-1})^{-1}$, which, using the technique of partitioned inverses can be written as

$$\text{Var}[f(x)] = (K_{m+1, m+1}^{*-1})^{-1} = \kappa - k^\top K^{-1} k. \quad (10)$$

From a practitioner's point of view it is significant that (8) and (10) only depend on x through k and κ . For given data $((x_1, y_1), (x_2, y_2), \dots, (x_m, y_m))$ only a single $m \times m$ matrix has to be inverted. Afterwards, to probe the posterior at various points is only a matter of matrix/vector multiplications.

To sample the posterior (i.e. approximate an entire function distributed according to $p(f|D)$) choose an arbitrarily fine mesh z_1, z_2, \dots, z_R on \mathcal{X} and apply

Sampling the
Posterior

$$\begin{aligned} p(f(z_1) = t_1, \dots, f(z_{r+1}) = t_{r+1}, D) &= \\ p(f(z_{r+1}) = t_{r+1} | f(z_1) = t_1, \dots, f(z_r) = t_r, D) &\cdot \\ p(f(z_1) = t_1, \dots, f(z_r) = t_r, D) & \end{aligned}$$

recursively, sampling similarly to (8) and (10) at each iteration, but inverting ever larger covariance matrices.