Short Note: Gaussian Processes

François Fleuret

October 16, 2009

We have a bunch of training points

$$(x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}, \quad n = 1 \dots N,$$

and we want to predict the value y associated to another x, or ideally we would like to have a posterior distribution that would tell us both the most likely value associated to x, but also the dispersion around it, etc.

Consider the following model: we define a functional basis

$$f_k : \mathbb{R}^d \to \mathbb{R}, \quad k = 1, \dots, K,$$

and an unknown vector of coefficients following a centered normal Gaussian distribution (i.e. expectation 0 and covariance matrix identity)

$$(A_1, \dots, A_K) \sim \mathcal{N}(\mathbf{0}, \mathbf{1}).$$

We postulate that the training points are of the form

$$\forall n, \ y_n = F(x_n),$$

where F is a random functional (hence a "process" in the probabilistic terminology) defined by

$$F = \sum_k A_k f_k.$$

Since the vector

$$(F(x_1), \dots, F(x_N), F(x)) \tag{1}$$

is equal to

$$\begin{array}{rcccccc} & A_1 & . & (f_1(x_1), & \dots & f_1(x_N), f_1(x)) \\ + & A_2 & . & (f_2(x_1), & \dots & f_2(x_N), f_2(x)) \\ & & & & \dots & & \\ + & A_K & . & (f_K(x_1), & \dots & f_K(x_N), f_K(x)), \end{array}$$

it is a Gaussian vector of dimension K multiplied by a matrix, which results obviously in a vector following a Gaussian density for which we know the covariance matrix.

From that we can compute explicitly

$$P(F(x) = y \,|\, F(x_1) = y_1, \ldots, F(x_N) = y_N) \\ = \frac{P(F(x) = y,\, F(x_1) = y_1, \ldots, F(x_N) = y_N)}{P(F(x_1) = y_1, \ldots, F(x_N) = y_N)}.$$

Since the denominator is a normalization quantity which does not depend on y, and the numerator is obviously Gaussian in y, we have the desired (Gaussian) posterior on the unknown value y associated to x.

Two remarks:

1. If you were looking for the values associated to several test points $x'_1 \ldots x'_L$, you would do exactly the same and would get a (Gaussian) joint posterior for them.

2. What matters at the end is not the functional basis $f_1, \ldots, f_K$, but a mean to compute the covariance matrix of (1). And you can do so just by knowing for every pair (n, m) the quantity $\kappa(x_n, x_m) = \sum_k f_k(x_n) f_k(x_m)$.

   Hence the possibility to "kernelize" all this: forget the f_k, that you actually never need, and just chose a

   $$\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}.$$

   Note that the $f_1, \ldots, f_K$ may not exist for an arbitrary κ. So for this description of things to be consistent, κ has to be a "Mercer kernel", which means that it is semi-definite positive. Also, since $\kappa(x, x')$ accounts for how much $F(x)$ and $F(x')$ "fluctuate together", it makes sense to see it as a similarity measure.

So, despite the fact that it sounds impressive to have a "functional basis", a "kernel" and a "Gaussian prior", at the end it boils down to a Gaussian vector of dimension $N + L$, for which we know N components and want to estimate the L others.