Gaussian processes

The problem:

we will consider real numbers

N data points $\mathbf{X}_N, \mathbf{t}_N = \{\mathbf{x}_n, t_n\}_{n=1}^{N}$

$\mathbf{x}_n$: vector of vectors; t_n: can be vectors

t = targets

assume that a function $y(\mathbf{x})$ underlies the observed data.

What is t_{N+1} at point $\mathbf{x}_{N+1}$?

Parametric approach

Express $y(\mathbf{x})$ in terms of $y(\mathbf{x}; \mathbf{w})$, model with parameters $\mathbf{w}$

E.g. using a set of basis functions $\{\phi_h(\mathbf{x})\}_{h=1}^{H}$ we have

$$y(\mathbf{x}; \mathbf{w}) = \sum_{h=1}^{H} w_h \phi_h(\mathbf{x})$$

[Note: these can be non-linear functions such as

$$\phi_h(\mathbf{x}) = \exp\left[-\frac{(\mathbf{x}-\mathbf{c}_h)^2}{2r^2}\right]$$

(aka "radial basis functions" in this context ... Gaussians)

Still, since $y(\mathbf{x}; \mathbf{w})$ depends linearly on $\mathbf{w}$, this is a linear model]

We have seen this problem before

$$p(w \mid t_N, X_N) = \frac{p(t_N \mid w, X_N)\, p(w)}{p(t_N \mid X_N)}$$

here expressed slightly differently

t_N = target values given X_N positions

do not include 'I' in these notes.

Having performed inference to obtain the posterior, a prediction can be made

$$p(t_{N+1} \mid t_N, X_{(N+1)}) = \int d^H w\; p(t_{N+1} \mid w, X_{(N+1)})\, p(w \mid t_N, X_N)$$

[Note: e.g. from R samples $w_{(r)}$ from $p(w \mid t_N, X_N)$

$$p(t_{N+1} \mid t_N, X_{N+1}) \simeq \frac{1}{R} \sum_{r=1}^{R} p(t_{N+1} \mid w_{(r)}, X_{(N+1)})$$

[Note that the final result does not make explicit reference to our representation (parametric one) of the unknown function $y(x)$.]

Note on notation:

From now on: X_N = vector of N vectors

t_N = vector of N scalars

$x^{(N+1)}$ = vector # $N+1$

$t^{(N+1)}$ = target # $N+1$

We will see that for models with fixed basis functions and Gaussian prior distributions (w. zero mean) on the parameters, the joint probability of all the observed data given the model, $p(t_N | X_N)$ is a multivariate Gaussian with mean zero and with a covariance function determined by the basis functions; this implies that the conditional distribution $p(t_{(N+1)} | t_N, X_{N+1})$ is also a Gaussian distribution whose mean depends linearly on t_N

Consider a linear model

$$R_{nh} \equiv \phi_h(x^{(n)}) \quad \text{, i.e. } R \text{ is } N \times H$$

(N: number of points $\{x_{(n)}\}_{n=1}^{N}$; H: # of basis functions)

y_N is the vector of values $y(x; w)$

$$y_{(n)} \equiv \sum_h R_{nh} w_h \qquad [\, y_N = R w \,]$$

Assume a Gaussian prior distribution for w

$$p(w) = \mathcal{N}(w; 0, \sigma_w^2 I)$$

since y is a linear function of w, it is also Gaussian distributed with mean zero. Its covariance matrix

$$Q = \langle y y^T \rangle = \langle R w w^T R^T \rangle = \sigma_w^2 R R^T$$

$$\Rightarrow p(y) = \mathcal{N}(y; 0, \sigma_w^2 R R^T)$$

This will be true for any selected points X_N; which is the defining property of a Gaussian process

"The probability distribution of a function $y(x)$ is a Gaussian process if for any selection of points: $x_{(1)}, \ldots x_{(N)}$ the density $p(y(x_{(1)}), \ldots, y(x_{(N)}))$ is a Gaussian"

What about the target values?

If $t_{(n)}$ is assumed to differ by additive Gaussian noise of variance σ_ν^2 from $y(n)$, then t also has a Gaussian prior distribution

$$p(t) = \mathcal{N}(t; 0, \underbrace{Q + \sigma_\nu^2 I}_{\equiv C = \sigma_w^2 RR^T + \sigma_\nu^2 I})$$

What does Q and C look like?

$$Q_{nn'} = \sigma_{w^2} \sum_h \phi_h(x^{(n)}) \phi_h(x^{(n')})$$

and

$$C_{nn'} = Q_{nn'} + \delta_{nn'} \sigma_\nu^2$$

i.e. the correlation between target values $t^{(n)}$ and $t^{(n')}$ is determined by the points $x^{(n)}$, $x^{(n')}$ and the behaviour of ϕ.

In fact, we don't need the basis functions and parameters any more.

Instead of a prior distribution on functions $p(y(x))$ in terms of basis functions and priors on parameters, we can summarize the prior simply by a covariance function

$$C(x, x')$$

generating a covariance matrix

$$Q_{nn'} = C(x^{(n)}, x^{(n')})$$

for any set of points $\{x_{(n)}\}_{n=1}^{N}$

Q must be non-negative-definite

$\Rightarrow$ constraint on valid cov. functions.

The cov. matrix for the target values

$$C_{nn'} = C(x^{(n)}, x^{(n')}) + \sigma_\nu^2 \delta_{nn'}$$

$$\Rightarrow p(t) = \mathcal{N}(t; 0, C) = \frac{1}{Z} e^{-\frac{1}{2} t^T C^{-1} t}$$

Covariance functions

Given the constraint on the previous page we can construct a wide range of covariance functions. We can introduce hyperparameters θ so that

$$C_{mn} = C(x^{(m)}, x^{(n)}; \theta) + \delta_{mn} N(x^{(n)}; \theta)$$

noise model

Stationary cov. functions are translationally invariant

$$C(x, x'; \theta) = C(x - x'; \theta)$$

Also known as a kernel

A popular choice (known as radial basis functions; RBF, in this context — but really a Gaussian)

$$C(x, x'; \theta) = \theta_1 \exp\left[-\frac{1}{2}\sum_{i=1}^{I}\frac{(x_i - x_i')^2}{r_i^2}\right] + \theta_2$$

with hyperparameters $\theta = (\theta_1, \theta_2, \{r_i\})$ or just one "r"

Determine θ by optimization (common) or marginalize over them.