Gaussian processes

Kernel methods

Francesco Corona

Outline

Gaussian processes
- Linear regression revisited
- Gaussian processes for regression
- Learning the hyper-parameters

Gaussian processes
Kernel methods

Gaussian processes

We applied the concept of duality to a non-probabilistic model for regression

- We extend the role of kernels to probabilistic discriminative models

We considered linear regression models of the form $y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \phi(\mathbf{x})$

- $\mathbf{w}$ is a vector of parameters and $\phi(\mathbf{x})$ is a vector of fixed nonlinear basis functions that depend on the input vector $\mathbf{x}$
- We showed that a prior distribution over $\mathbf{w}$ induced a corresponding prior distribution over functions $y(\mathbf{x}, \mathbf{w})$

Given a training data set, we evaluated the posterior distribution over $\mathbf{w}$

- To obtain a corresponding posterior distribution over regression functions

With noise, it implies a predictive distribution $p(t|\mathbf{x})$ for new inputs $\mathbf{x}$

Gaussian processes (cont.)

In the Gaussian process viewpoint, we dispense with the parametric model and instead define a prior probability distribution over functions directly

It is difficult to work with a distribution over the infinite space of functions

- For a finite training set we only need to consider the values of the function at the discrete set of input values $\mathbf{x}_n$ corresponding to the training set and test set data points, and so in practice we can work in a finite space

Linear regresson revisited
Gaussian processes

Linear regression revisited

To illustrate the Gaussian process viewpoint, we consider linear regression

- We re-derive the predictive distribution
- In terms of distributions over functions $y(\mathbf{x}, \mathbf{w})$

Consider a model defined in terms of a linear combination of M fixed basis functions given by the elements of the vector $\boldsymbol{\phi}(\mathbf{x}) = (\phi_0(\mathbf{x}), \ldots, \phi_{M-1}(\mathbf{x}))^T$

$$y(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) \tag{1}$$

where $\mathbf{x}$ is the input vector and $\mathbf{w}$ is the M-dimensional weight vector

Linear regression revisited (cont.)

Consider a prior distribution over $\mathbf{w}$ given by an isotropic Gaussian of the form

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \tag{2}$$

governed by hyperparameter α, precision (inverse variance) of the distribution

- For any given value of $\mathbf{w}$, $y(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x})$ defines a function of $\mathbf{x}$

The probability distribution over $\mathbf{w}$ induces
a probability distribution over functions $y(\mathbf{x})$

We wish to evaluate this function at specific values of $\mathbf{x}$, say the training data

$$\mathbf{x}_1, \ldots, \mathbf{x}_N$$

Linear regression revisited (cont.)

We are interested in the joint distribution of function values $y(\mathbf{x}_1), \ldots, y(\mathbf{x}_N)$, which we denote by the vector y with elements $y_n = y(\mathbf{x}_n)$, for $n = 1, \ldots, N$

From $y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$, this vector is given by

$$\mathrm{y} = \mathbf{\Phi w} \tag{3}$$

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix} : \text{ Design matrix } (\Phi_{nk} = \phi_k(\mathbf{x}_n))$$

Linear regression revisited (cont.)

We find the probability distribution of y by seeing that y is a linear combination of Gaussian distributed variables, the elements of $\mathbf{w}$ and thus is itself Gaussian

- We need to find its mean and covariance

$$\mathbb{E}[\mathrm{y}] = \mathbf{\Phi}\mathbb{E}[\mathbf{w}] = \mathbf{0} \tag{4}$$

$$\mathrm{cov}[\mathrm{y}] = \mathbb{E}[\mathrm{y}\mathrm{y}^T] = \mathbf{\Phi}\mathbb{E}[\mathbf{w}\mathbf{w}^T]\mathbf{\Phi}^T = \frac{1}{\alpha}\mathbf{\Phi}\mathbf{\Phi}^T = \mathbf{K} \tag{5}$$

- where $\mathbf{K}$ is the Gram matrix with elements

$$k_{nm} = k(\mathbf{x}_n, \mathbf{x}_m) = \frac{1}{\alpha}\phi(\mathbf{x}_n)^T\phi(\mathbf{x}_m) \tag{6}$$

This model provides us only with a particular example of a Gaussian process

Linear regression revisited (cont.)

A Gaussian process is a probability distribution over functions $y(\mathbf{x})$ such that the set of values of $y(\mathbf{x})$ evaluated at an arbitrary set of points $\{\mathbf{x}_n\}$ jointly have a Gaussian distribution

A key point about Gaussian processes is that the joint distribution over the N variables $y_1, \ldots, y_N$ is specified completely by second-order statistics

- **Mean**: In most applications, we will not have any prior knowledge about the mean of $y(\mathbf{x})$ and so by symmetry we take it to be zero

This is equivalent to choosing the mean of the prior over weight values $p(\mathbf{w}|\alpha)$ to be zero in the basis function viewpoint

- **Covariance**: The specification of the Gaussian process is completed by giving the covariance of $y(\mathbf{x})$ evaluated at any two values of $\mathbf{x}$

This is given by the kernel function $\mathbb{E}[y(\mathbf{x}_n)y(\mathbf{x})_m] = k(\mathbf{x}_n, \mathbf{x}_m)$

Linear regression revisited (cont.)

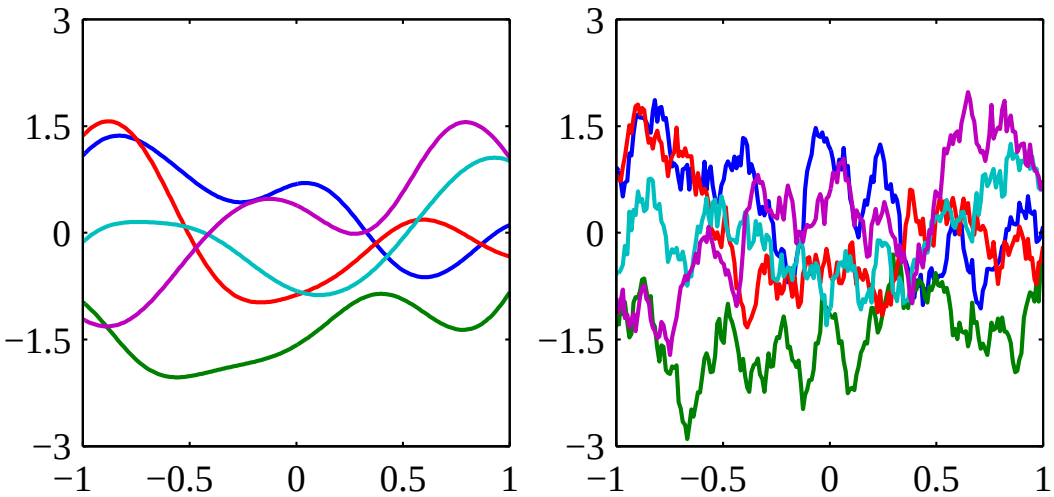
For the specific case of a Gaussian process defined by the linear regression model $y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$ with a weight prior $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$,

- the kernel function is given by $k_{nm} = k(\mathbf{x}_n, \mathbf{x}_m) = \frac{1}{\alpha}\phi(\mathbf{x}_n)^T \phi(\mathbf{x}_n)$

Linear regression revisited (cont.)

We can also define the kernel function directly, rather than indirectly

- By-pass the choice of basis functions
- Draw samples of functions from the GP



Gaussian processes for regression

Gaussian processes

Gaussian processes for regression

In order to apply Gaussian process models to the problem of regression, we need to take account of the noise on the observed target values

$$t_n = y_n + \varepsilon_n \tag{7}$$

where $y_n = y(\mathbf{x}_n)$ and ε_n is a random noise variable whose value is chosen independently for each observation n

We consider noise processes that have a Gaussian distribution, so that

$$p(t_n|y_n) = \mathcal{N}(t_n|y_n, \beta^{-1}) \tag{8}$$

where β is a hyperparameter representing for the precision of the noise

Gaussian processes for regression (cont.)

Because the noise is independent for each point, the joint distribution of the target values $\mathrm{t} = (t_1, \ldots, t_N)^T$ conditioned on the values of $y = (y_1, \ldots, y_N)^T$ is given by an isotropic Gaussian of the form

$$p(\mathrm{t}|\mathrm{y}) = \mathcal{N}(\mathrm{t}|\mathrm{y}, \beta^{-1}\mathbf{I}_N) \tag{9}$$

From the definition of Gaussian process, the marginal distribution $p(\mathrm{y})$ is given by a Gaussian whose mean is zero and whose covariance is a Gram matrix $\mathbf{K}$

$$p(\mathrm{y}) = \mathcal{N}(\mathrm{y}|\mathbf{0}, \mathbf{K}) \tag{10}$$

The kernel function that determines $\mathbf{K}$ can be chosen to express the property that, for points $\mathbf{x}_n$ and $\mathbf{x}_m$ that are similar, corresponding values $y(\mathbf{x}_n)$ and $y(\mathbf{x}_m)$ will be more strongly correlated than for dissimilar points

Gaussian processes for regression (cont.)

In order to find the marginal distribution $p(\mathrm{t})$, conditioned on the input values $\mathbf{x}_1, \ldots, \mathbf{x}_N$, we need to integrate $p(\mathrm{t}|\mathrm{y})$ over y

$$p(\mathrm{t}) = \int p(\mathrm{t}|\mathrm{y})p(\mathrm{y})d\mathrm{y} = \mathcal{N}(\mathrm{t}|\mathbf{0}, \mathbf{C}) \tag{11}$$

where the covariance matrix $\mathbf{C}$ has elements

$$C(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m) + \beta^{-1}\delta_{nm} \tag{12}$$

- δ_{nm} is a Kronecker delta (1 iff $n = m$, 0 otherwise)

The covariance matrix $\mathbf{C}$ reflects the fact that the two Gaussian sources of randomness (one associated with $y(\mathbf{x})$ and one to ε) are independent

- their covariances ($\mathbf{K}$ and $\beta^{-1}\mathbf{I}$) simply add

Gaussian processes for regression (cont.)

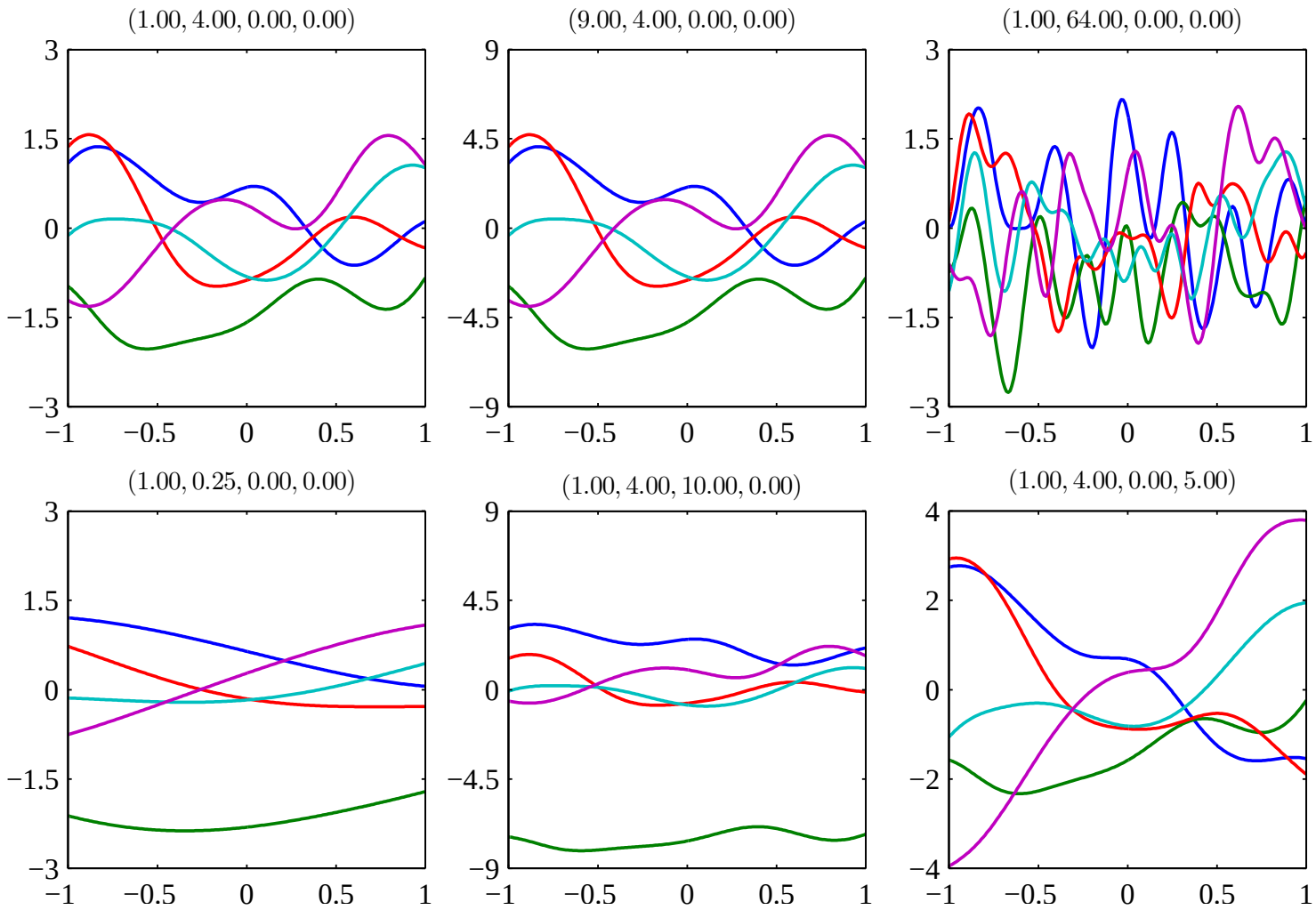
One widely used kernel function for Gaussian processes is the exponential of a quadratic form with the addition of constant and linear terms to give

$$k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 \exp\left(-\frac{\theta_1}{2}||\mathbf{x}_n - \mathbf{x}_m||^2 \right) + \theta_2 + \theta_3 \mathbf{x}_n^T \mathbf{x}_m \tag{13}$$

The term involving θ_3 corresponds to a parametric model that is a linear function of the input variables.

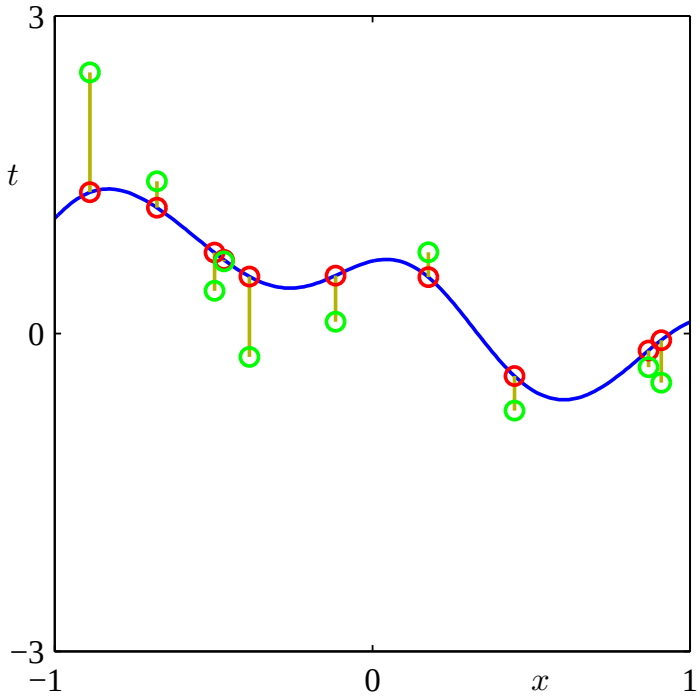
Gaussian processes for regression (cont.)

From a GP prior with covariance function $k(\mathbf{x}_n, \mathbf{x}_m)$, we can sample functions



Gaussian processes for regression (cont.)

A sampled function (blue line) drawn from the Gaussian process prior over functions is evaluated at a set of points $\{x_n\}$ to give points $\{y_n\}$ (red dots)



The corresponding values of $\{t_n\}$ (green dots) are obtained by adding independent Gaussian noise to each of the points in $\{y_n\}$

Gaussian processes for regression (cont.)

Our goal in regression is to make predictions of target variables for new inputs

- given a set of training data

Let us suppose that $\mathbf{t}_N = (t_1, \ldots, t_N)^T$ for input values $\mathbf{x}_1, \ldots, \mathbf{x}_N$, comprise the observed training data set and our goal is to predict the target variable t_{N+1}

- for a new input vector $\mathbf{x}_{N+1}$

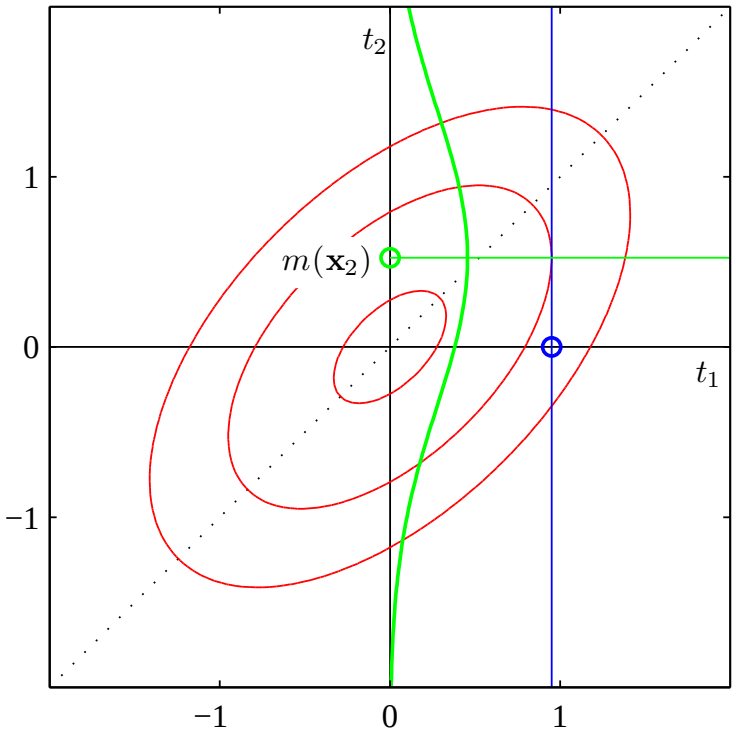
As always, this requires that we evaluate the predictive distribution $p(t_{N+1}|\mathbf{t}_N)$

- This distribution is conditioned also on the variables $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and $\mathbf{x}_{N+1}$

Gaussian processes for regression (cont.)

To find the conditional distribution $p(t_{N+1}|\mathbf{t}_N)$, we begin by writing down the joint distribution $p(\mathbf{t}_{N+1})$, where $\mathbf{t}_{N+1}$ is the vector $\mathbf{t}_{N+1} = (t_1, \ldots, t_N, t_{N+1})^T$

We then apply results from the Gaussian distribution to obtain the conditional



For one training point t_1 and one test point t_2

- Contours of the joint distribution $p(t_1, t_2)$

We condition on (fix) the value of t_1 (blue line)

- We obtain $p(t_2|t_1)$

We know that the conditional distribution $p(t_{N+1}|\mathbf{t})$ will also be a Gaussian distribution

Gaussian processes for regression (cont.)

From $p(\mathrm{t}) = \int p(\mathrm{t}|\mathrm{y})p(\mathrm{y})d\mathrm{y}$, the joint distribution over $t_1, \ldots, t_{N+1}$ is given by

$$p(\mathrm{t}_{N+1}) = \mathcal{N}(\mathrm{t}_{N+1}|\mathbf{0}, \mathbf{C}_{N+1}) \tag{14}$$

where $\mathbf{C}_{N+1}$ is a $(N+1) \times (N+1)$ covariance matrix with elements given by

$$C(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m) + \beta^{-1}\delta_{nm}$$

Because this joint distribution is Gaussian, we can apply the results from the Gaussian distribution to characterise this conditional Gaussian distribution

Gaussian processes for regression (cont.)

Remembering that when two sets of variables are jointly Gaussian also the conditional distribution of one set conditioned is Gaussian

For an arbitrary vector $\mathbf{x}$ with Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- We first partitioned $\mathbf{x}$ into two disjoint subsets $\mathbf{x}_a$ and $\mathbf{x}_b$

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}$$

- We partitioned mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

From these, we obtained the expressions for the mean and covariance of the conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ in terms of the partitioned covariance matrix

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \end{aligned}$$

Gaussian processes for regression

We first partition the covariance matrix of the joint distribution

$$\mathbf{C}_{N+1} = \begin{pmatrix} \mathbf{C}_N & \mathbf{k} \\ \mathbf{k}^T & c \end{pmatrix} \tag{15}$$

- $\mathbf{C}_N$ is a $N \times N$ covariance matrix with elements given by

$$C(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m) + \beta^{-1}\delta_{nm}, \quad \text{with } n, m = 1, \ldots, N$$

- Vector $\mathbf{k}$ has elements $k(\mathbf{x}_n, \mathbf{x}_{N+1})$ for $n = 1, \ldots, N$
- Scalar $c = k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}) + \beta^{-1}$

Gaussian processes for regression (cont.)

Using expressions from the conditional Gaussian distribution on $p(t_n|\mathbf{t})$ yields

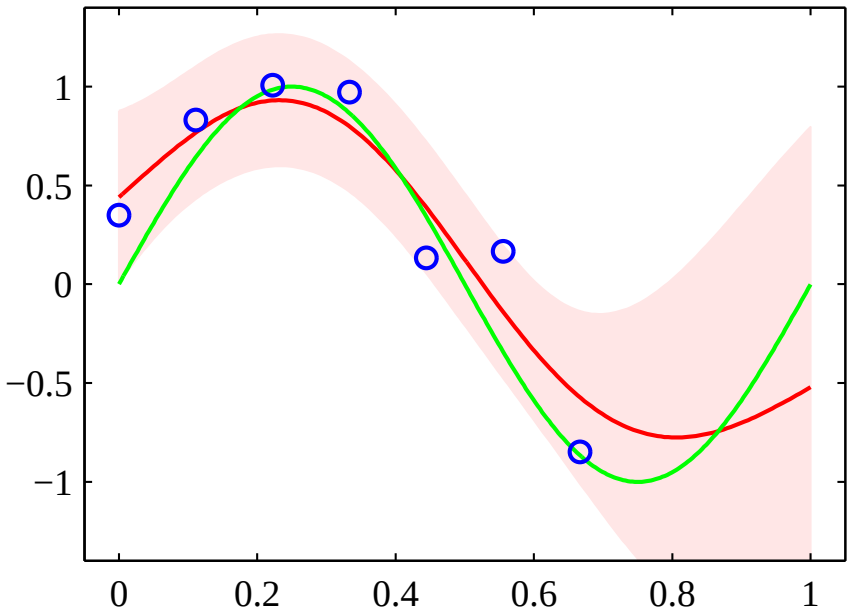
$$m(\mathbf{x}_{N+1}) = \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{t} \quad (16)$$

$$\sigma^2(\mathbf{x}_{N+1}) = c - \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{k} \quad (17)$$

Because vector $\mathbf{k}$ is a function of test point input value $\mathbf{x}_{N+1}$, the predictive distribution is a Gaussian whose mean and variance both depend on $\mathbf{x}_{N+1}$

Gaussian processes for regression (cont.)

Illustration of Gaussian process regression applied to the sinusoidal data set



Three right-most points were omitted

Green curve: The sine function from which data points (blue) are obtained by sampling and adding Gaussian noise

Red line: The mean of the Gaussian process predictive distribution

- $\pm$ two standard deviations

Note how the uncertainty increases in the region to the right of the data points

Gaussian processes for regression (cont.)

The only restriction on the kernel function is that the covariance matrix $C(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m) + \beta^{-1}\delta_{nm}$ must be positive definite

- If λ_i is an eigenvalue of $\mathbf{K}$, then the associated eigenvalue of $\mathbf{C}$ will be $\lambda_i + \beta^{-1}$

It is therefore sufficient that the kernel matrix $k(\mathbf{x}_n, \mathbf{x}_m)$ be positive semidefinite for any pair of points $\mathbf{x}_n$ and $\mathbf{x}_m$, so that $\lambda_i \geq 0$

- any eigenvalue λ_i that is zero will still give rise to a positive eigenvalue for $\mathbf{C}$ because $\beta > 0$

This is the same restriction on the kernel function discussed earlier

- We can exploit all of the techniques to construct suitable kernels

Gaussian processes for regression (cont.)

Mean $m(\mathbf{x}_{N+1}) = \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{t}$ of the predictive distribution is a function of $\mathbf{x}_{N+1}$

$$m(\mathbf{x}_{N+1}) = \sum_{n=1}^{N} a_n k(\mathbf{x}_N, \mathbf{x}_{N+1}), \quad a_n \text{ is the } n\text{-th component of } \mathbf{C}_N^{-1}\mathrm{t} \qquad (18)$$

For a kernel function $k(\mathbf{x}_n, \mathbf{x}_m)$ depends only on the distance $||\mathbf{x_n} - \mathbf{x}_m||$, we obtain an expansion in radial basis functions

Gaussian processes for regression (cont.)

$$\begin{aligned} m(\mathbf{x}_{N+1}) &= \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{t} \\ \sigma^2(\mathbf{x}_{N+1}) &= c - \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{t} \end{aligned}$$

The results above define the predictive distribution for Gaussian process for regression with an arbitrary kernel function $k(\mathbf{x}_n, \mathbf{x}_m)$

In the particular case in which the kernel function $k(\mathbf{x}_n, \mathbf{x}_m)$ is defined in terms of a finite set of basis functions, we can derive the results for linear regression

- from a Gaussian process view point ($\star$)

For such models, we can therefore obtain the predictive distribution either

- by taking a parameter space viewpoint and using linear regression results
- by taking a function space viewpoint and using the Gaussian process result

Gaussian processes for regression (cont.)

The central computational operation in GP is the inversion of a $N \times N$ matrix

- Standard methods require $\mathcal{O}(N^3)$ computations

In the basis function model, we have to invert a matrix $\mathbf{S}_N$ of size $M \times M$

- $\mathcal{O}(M^3)$ computational complexity

For both, the matrix inversion must be performed once for the given training set

For each new test point, both require a vector-matrix multiply, which has cost $\mathcal{O}(N^2)$ for Gaussian process models and $\mathcal{O}(M^2)$ for linear basis function models

If the number M of basis functions is smaller than the number N of points, it is computationally more efficient to work in the basis function framework

Learning the hyper-parameters
Gaussian processes

Learning the hyper-parameters

Predictions of a GP model depends partly on the choice of covariance function

In practice, rather than fixing the covariance function, we may prefer to use a parametric family of functions and then infer the parameter values from data

These parameters govern such things as length scale of correlations and the precision of noise, they are hyper-parameters in a standard parametric model

Learning the hyper-parameters (cont.)

Techniques for learning the hyper-parameters are based on the evaluation of the likelihood function $p(\mathbf{t}|\boldsymbol{\Theta})$ where $\boldsymbol{\Theta}$ denotes the hyper-parameters of the GP

The simplest approach is to make a point estimate of $\boldsymbol{\Theta}$ by maximising the log likelihood function (e.g., efficient gradient-based optimisation algorithms as CG)

Learning the hyper-parameters (cont.)

The log likelihood function for a Gaussian process regression model is evaluated using the standard form for a multivariate Gaussian distribution, giving

$$\ln p(\mathbf{t}|\boldsymbol{\Theta}) = -\frac{1}{2}\mathrm{Tr}|\mathbf{C}_N| - \frac{1}{2}\mathbf{t}^T\mathbf{C}_N^{-1}\mathbf{t} - \frac{N}{2}\ln(2\pi) \tag{19}$$

For nonlinear optimisation, we also need the gradient of the log likelihood function with respect to the parameter vector $\boldsymbol{\Theta}$

$$\frac{\partial \ln p(\mathbf{t}|\boldsymbol{\Theta})}{\partial \theta_i} = -\frac{1}{2}\mathrm{Tr}\left(\mathbf{C}_N^{-1}\frac{\partial \mathbf{C}_N}{\partial \theta_i}\right) + \frac{1}{2}\mathbf{t}^T\mathbf{C}_N^{-1}\frac{\partial \mathbf{C}_N}{\partial \theta_i}\mathbf{C}_N^{-1}\mathbf{t} \tag{20}$$

In general, $\ln p(\mathbf{t}|\boldsymbol{\Theta})$ is a non-convex function, it might have multiple maxima

Learning the hyper-parameters (cont.)

Alternatively, we could introduce a prior over $\boldsymbol{\Theta}$ and maximise the log posterior

- Again, by using gradient-based methods

In a fully Bayesian treatment, we would need to evaluate marginals over $\boldsymbol{\Theta}$ weighted by the product of the prior $p(\boldsymbol{\Theta})$ and the likelihood function $p(\mathbf{t}|\boldsymbol{\Theta})$

- In general, however, exact marginalisation will be intractable
- We must resort to approximations

Learning the hyper-parameters (cont.)

The Gaussian process regression model gives a predictive distribution whose mean and variance are functions of the input vector **x**

However, we have assumed that the contribution to the predictive variance arising from the additive noise, governed by the parameter β, is a constant

For hetero-scedastic problems, the noise variance itself will also depend on **x**

To model this, we can extend the Gaussian process framework by introducing a second Gaussian process to represent the dependence of β on the input **x**