Nonparametric regression

Giorgio Gambosi

Fully bayesian regression

We remind that, in fully bayesian regression, no specific model parameters $\mathbf{w}^*$ are identified, to be applied in prediction as

$$y = \mathbf{w}^* \phi(\mathbf{x})$$

Instead the distribution $p(y|\mathbf{x})$ is derived, under the assumption of gaussianity, with

$$p(y|\mathbf{x}, \mathbf{t}, \Phi, \alpha, \beta) = \mathcal{N}(y|m(\mathbf{x}), \sigma^2(\mathbf{x}))$$

and

$$m(\mathbf{x}) = \beta \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \Phi^T \mathbf{t}$$

and variance

$$\sigma^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})$$

Equivalent kernel

- The prediction $y(\mathbf{x})$ can be returned here as the expectation of the predictive distribution

$$y(\mathbf{x}) = \beta \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \Phi^T \mathbf{t} = \sum_{i=1}^{n} \beta \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}_i) t_i$$

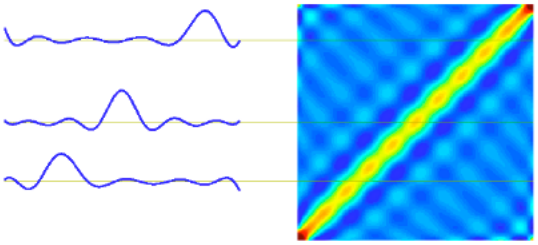
- The prediction is not computed by referring to a set of parameters derived by optimization of a loss function. Instead, it can be seen as a linear combination of the target values t_i of all items in the training set, with weights dependent from the item values $\mathbf{x}_i$ (and from $\mathbf{x}$)

$$y(\mathbf{x}) = \sum_{i=1}^{n} \kappa(\mathbf{x}, \mathbf{x}_i) t_i$$

  The weight function $\kappa(\mathbf{x}, \mathbf{x}') = \beta \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}')$ is said equivalent kernel

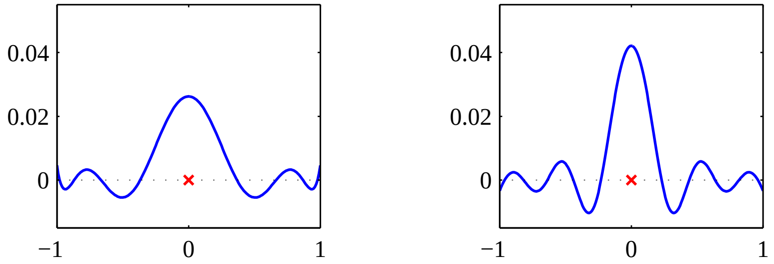
Right: plot on the plane (x, x_i) of a sample equivalent kernel, in the case of gaussian basis functions.

Left: plot as a function of x_i for three different values of x



In deriving y, the equivalent kernel tends to assign greater relevance to the target values t_i corresponding to items x_i near to x.

The same localization property holds also for different base functions.



Left, $\kappa(0, x')$ in the case of polynomial basis functions.

Right, $\kappa(0, x')$ in the case of gaussian basis functions.

- The covariance between $y(\mathbf{x})$ and $y(\mathbf{x}')$ is given by

$$\text{cov}(\mathbf{x}, \mathbf{x}') = \text{cov}(\boldsymbol{\phi}(\mathbf{x})^T\mathbf{w}, \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}')) = \Phi(\mathbf{x})^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}') = \frac{1}{\beta}\kappa(\mathbf{x}, \mathbf{x}')$$

  predicted values are highly correlated at nearby points.

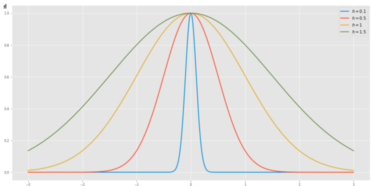
- Instead of introducing base functions which results into a kernel, we may define a localized kernel directly and use it to make predictions

Kernel regression

- In kernel regression methods, the target value corresponding to any item $\mathbf{x}$ is predicted by referring to items in the training set, and in particular to the items which are closer to $\mathbf{x}$.
- This is controlled by referring to a kernel function $\kappa_h(\mathbf{x})$, which is non zero only in an interval around 0
- h is the bandwidth of the kernel, which controls the width of $\kappa_h(\mathbf{x})$

A possible, common kernel, is the gaussian (or RBF) kernel

$$g(\mathbf{x}) = e^{-\frac{\|\mathbf{x}\|^2}{2h^2}}$$



In regression, we are interested in estimating the conditional expectation

$$f(\mathbf{x}) = E[t|\mathbf{x}] = \int p(t|\mathbf{x})tdt = \int \frac{p(\mathbf{x}, t)}{p(\mathbf{x})}tdt = \frac{\displaystyle\int p(\mathbf{x}, t)tdt}{p(\mathbf{x})} = \frac{\displaystyle\int p(\mathbf{x}, t)tdt}{\displaystyle\int p(\mathbf{x}, t)dt}$$

The joint distribution $p(\mathbf{x}, t)$ is approximated by means of a kernel function as

$$p(\mathbf{x}, t) \approx \frac{1}{n}\sum_{i=1}^{n} \kappa_h(\mathbf{x} - \mathbf{x}_i)\kappa_h(t - t_i)$$

This results into

$$f(\mathbf{x}) = \frac{\displaystyle\int \frac{1}{n}\sum_{i=1}^{n} \kappa_t(\mathbf{x} - \mathbf{x}_i)\kappa_h(t - t_i)tdt}{\displaystyle\int \frac{1}{n}\sum_{i=1}^{n} \kappa_h(\mathbf{x} - \mathbf{x}_i)\kappa_h(t - t_i)dt} = \frac{\sum_{i=1}^{n} \kappa_h(\mathbf{x} - \mathbf{x}_i)\displaystyle\int \kappa_h(t - t_i)tdt}{\sum_{i=1}^{n} \kappa_h(\mathbf{x} - \mathbf{x}_i)\displaystyle\int \kappa_h(t - t_i)dt}$$

If we assume that the kernel $\kappa(x)$ is a probability distribution with 0 mean, it results $\displaystyle\int \kappa_h(t - t_i)dt = 1$ and $\displaystyle\int t\kappa_h(t - t_i)dt = t_i$, we get

$$f(\mathbf{x}) = \frac{\sum_{i=1}^{n} \kappa_h(\mathbf{x} - \mathbf{x}_i)t_i}{\sum_{i=1}^{n} \kappa_h(\mathbf{x} - \mathbf{x}_i)}$$

By setting

$$w_i(\mathbf{x}) = \frac{\kappa_h(\mathbf{x} - \mathbf{x}_i)}{\sum_{j=1}^{n} \kappa_h(\mathbf{x} - \mathbf{x}_j)}$$

we can write

$$f(\mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x})t_i$$

that is, the predicted value is computed as a normalized linear combination of all target values, weighted by kernels (Nadaraya-Watson)

Locally weighted regression

In Nadaraya-Watson model, the prediction is performed by means of a normalized weighted combination of constant values (target values in the training set).

Locally weighted regression (LOESS) improves that approach by referring to a weighted version of the sum of squared differences loss function used in regression.

If a value t has to be predicted for an item $\mathbf{x}$, a “local” version of the loss function is considered, with weight $\kappa_i(\mathbf{x})$.

$$L(\mathbf{x}) = \sum_{i=1}^{n} \kappa_i(\mathbf{x})(\mathbf{w}^T\overline{\mathbf{x}}_i - t_i)^2 = \sum_{i=1}^{n} \kappa_h(\mathbf{x} - \mathbf{x}_i)(\mathbf{w}^T\overline{\mathbf{x}}_i - t_i)^2$$

Weights $\kappa_i(\mathbf{x})$ are dependent from the “distance” between $\mathbf{x}$ and $\mathbf{x}_i$, as measured by the kernel function

$$\kappa_i(\mathbf{x}) = \kappa_h(\mathbf{x} - \mathbf{x}_i)$$

The minimization of this loss function

$$\hat{\mathbf{w}}(\mathbf{x}) = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{n} \kappa_i(\mathbf{x})(\mathbf{w}^T\overline{\mathbf{x}}_i - t_i)^2$$

has solution

$$\hat{\mathbf{w}}(\mathbf{x}) = (\overline{\mathbf{X}}^T\Psi(\mathbf{x})\overline{\mathbf{X}})^{-1}\overline{\mathbf{X}}^T\Psi(\mathbf{x})\mathbf{t}$$

where $\Psi(\mathbf{x})$ is a diagonal $n \times n$ matrix with $\Psi(\mathbf{x})_{ii} = \kappa_i(\mathbf{x})$.

The prediction is then performed as usual, as

$$y = \hat{\mathbf{w}}(\mathbf{x})^T\overline{\mathbf{x}}$$

Local logistic regression

The same approach applied in the case of local regression can be applied for classification, by defining a weighted loss function to be minimized, with weights dependent from the item whose target must be predicted.

In this case, a weighted version of the cross entropy function is considered, which has to be maximized

$$L(\mathbf{x}) = \sum_{i=1}^{n}\kappa_h(\mathbf{x}-\mathbf{x}_i)(t_i\log p_i - (1-t_i)\log(1-p_i))$$

with $p_i = \sigma(\mathbf{w}^T\overline{\mathbf{x}}_i)$, as usual.

The loss function minimization can be performed, for example, by applying a suitable modification of the IRLS algorithm for logistic regression

Recap: some properties of Gaussian distribution

In order to introduce Gaussian processes and how they can be exploited for regression, let us first provide a short reminder on some properties of multivariate gaussian distributions.

Let $\mathbf{x} = (x_1, \ldots, x_n)^T$ be a random vector with gaussian distribution $p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and let $\mathbf{x} = (\mathbf{x}_A, \mathbf{x}_B)$ be a partition of the components $\mathbf{x}$ such that:

- $\mathbf{x}_A = (x_1, \ldots, x_r)^T$
- $\mathbf{x}_B = (x_{r+1}, \ldots, x_n)^T$

Then, the marginal densities $p(\mathbf{x}_A)$ and $p(\mathbf{x}_B)$ are both gaussian with means $\boldsymbol{\mu}_A, \boldsymbol{\mu}_B$ and covariance matrices Σ_A, Σ_B which can be derived from $\boldsymbol{\mu}, \Sigma$ by observing that

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_A, \boldsymbol{\mu}_B)^T \qquad\qquad \Sigma = \begin{pmatrix} \Sigma_A & \Sigma_{AB} \\ \Sigma_{AB}^T & \Sigma_B \end{pmatrix}$$

In the same situation, the conditional densities $p(\mathbf{x}_A|\mathbf{x}_B)$ and $p(\mathbf{x}_B|\mathbf{x}_A)$ are also gaussian with means

$$\boldsymbol{\mu}_{A|B} = \boldsymbol{\mu}_A + \Sigma_{AB}\Sigma_B^{-1}(\mathbf{x}_B - \boldsymbol{\mu}_B)$$
$$\boldsymbol{\mu}_{B|A} = \boldsymbol{\mu}_B + \Sigma_{BA}\Sigma_A^{-1}(\mathbf{x}_A - \boldsymbol{\mu}_A)$$

and covariance matrices

$$\Sigma_{A|B} = \Sigma_A - \Sigma_{AB}\Sigma_B^{-1}\Sigma_{BA}$$
$$\Sigma_{B|A} = \Sigma_B - \Sigma_{BA}\Sigma_A^{-1}\Sigma_{AB}$$

Gaussian processes

- Multivariate gaussians on random vectors are useful for modeling finite collections of real-valued variables. They have nice analytical properties (see previous slides).
- Gaussian processes: extension of multivariate gaussians to infinite-sized collections of real-valued variables.
- We may think of gaussian processes as distributions not just over random vectors but over random real functions.

Probability distributions over functions with finite domains

Let us first consider the case of functions defined over finite vectors.

- Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ be a vector of m points in $\mathbb{R}^d$, and let $\mathcal{H}$ be the set of functions $f : \mathbb{R}^d \mapsto \mathbb{R}$
  - any such functions assigns a value $f(\mathbf{x}_i)$ to each $\mathbf{x}_i \in \mathbf{X}$ and can be described by the vector $(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_m))$
  - at the same time, any vector $\mathbf{y} = (y_1, \ldots, y_m)$ can be seen as the description of a function $f \in \mathcal{H}$, the one with $f(\mathbf{x}_i) = y_i$
  - hence, the set $\mathcal{H}$ is in 1-to-1 correspondence with the set of vectors in $\mathbb{R}^m$
- A probability distribution $p(\mathbf{y}), \mathbf{y} \in \mathbb{R}^m$ is also a distribution $p(f)$ of functions in $\mathcal{H}$

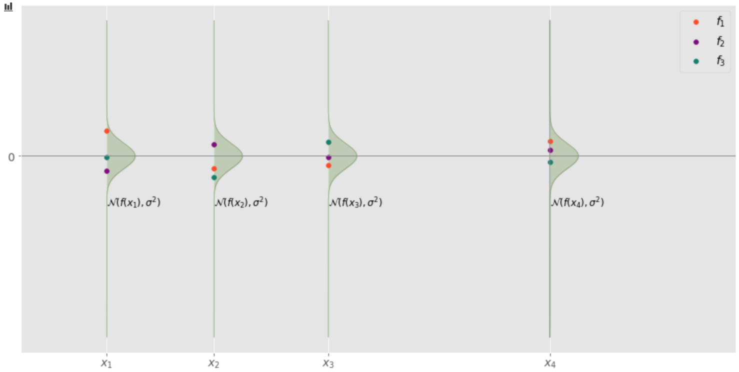
Gaussian distributions over functions with finite domains

We assume that $p(\mathbf{y})$ (or, equivalently, $p(f)$) is a (multivariate, m-dimensional) Gaussian distribution with mean $\mathbf{0}$ and diagonal covariance matrix $\sigma^2\mathbf{I}$, that is

$$p(\mathbf{y}|\mathbf{X}; \sigma^2) = \mathcal{N}(\mathbf{y}|\mathbf{X}; \mathbf{0}, \sigma^2\mathbf{I}) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y_i^2}{2\sigma^2}}$$

- This is equivalent to assuming that each function value $y_i = f(\mathbf{x}_i)$ has normal distribution with mean 0 and variance σ^2, and that values are independent
- A dependence between function values at different points could be modeled through a non-diagonal covariance matrix

In the figure below, a possible situation is given with $d = 1$, $m = 4$: three functions in $\mathcal{H}$ are reported.



- Assume now that the targets $\mathbf{t} = (t_1, \ldots, t_m)$ corresponding to points in $\mathbf{X}$ are available.
- Observe that $p(\mathbf{y}|\sigma^2)$ is only dependent on $\mathbf{y}$ and $\mathbf{X}$, and does not take into account the targets $\mathbf{t}$. We may than consider it as a prior distribution of functions, with respect to the observation of the targets $\mathbf{t}$ associated to $\mathbf{X}$

- By applying Bayes rule, we may derive the posterior (with respect to $\mathbf{t}$) distribution $p(\mathbf{y}|\mathbf{X},\mathbf{t})$ of functions. To this aim, a likelihood model has to be defined

$$p(\mathbf{X},\mathbf{t}|\mathbf{y}) = \prod_{i=1}^{m} p(\mathbf{x}_i, t_i|y_i) = \prod_{i=1}^{m} p(t_i|\mathbf{x}_i, y_i)p(\mathbf{x}_i|y_i) \propto \prod_{i=1}^{m} p(t_i|\mathbf{x}_i, y_i)$$

- we refer to the usual gaussian likelihood introduced for probabilistic modeling linear regression $p(t|\mathbf{x}, y, \beta) = \mathcal{N}(t|f(\mathbf{x}), \beta)$, which results into

$$p(\mathbf{X},\mathbf{t}|\mathbf{y}, \beta) \propto \prod_{i=1}^{m} \mathcal{N}(t_i|f(\mathbf{x}_i), \beta)$$

- the posterior distribution is then

$$p(f|\mathbf{X},\mathbf{t}, \beta, \sigma^2) \propto \prod_{i=1}^{m} \mathcal{N}(t_i|f(\mathbf{x}_i), \beta)p(f|\sigma^2)$$

Both the prior and the posterior distributions of f are gaussian: this implies that the predictive distribution

$$p(t|\mathbf{x},\mathbf{X},\mathbf{t}, \beta, \sigma^2) = \int p(t|\mathbf{x}, f, \beta)p(f|\mathbf{X},\mathbf{t}, \beta, \sigma^2)df$$

is itself a gaussian.

That would the case also in the more general case when some dependency between function points is assumed. In this case, a general covariance matrix Σ is defined for the prior distribution

$$p(\mathbf{y}|\Sigma) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \Sigma)$$

- In the case of an infinite domain χ, we have to deal with an infinite collection of random variables.
- In this case, the role of multidimensional distributions is covered by stochastic processes.
  - A *stochastic process* is a collection of random variables, $\{f(\mathbf{x}) : \mathbf{x} \in \chi\}$, indexed by elements from some set $\mathbf{X}$, known as the index set.
- A Gaussian process is a stochastic process such that for any finite subset $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of χ, the function values $f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)$ have joint multivariate Gaussian distribution

In order to specify the gaussian process, we must introduce two rules which, for any set of points $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, define the distribution $p(\mathbf{y})$ of the corresponding values $y_1, \ldots, y_n = f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)$.

- We already know that, by assumption, $p(\mathbf{y})$ is a multivariate normal distribution, hence characterized by a mean vector $\boldsymbol{\mu}(\mathbf{X})$ and covariance matrix $\Sigma(\mathbf{X})$
- We assume that $\boldsymbol{\mu}(\mathbf{X})$ is indeed a constant independent from $\mathbf{X}$. In particular, $\boldsymbol{\mu}(\mathbf{X}) = \mathbf{0}$
- The covariance matrix derives from the application of a predefined covariance function $\kappa : \chi \times \chi \mapsto \mathbb{R}$ which associates a real value to any pair of points in χ and, in particular, to any pair in $\mathbf{X}$, hence to all elements of Σ

Kernels in gaussian processes

The covariance function κ is assumed to be a positive definite kernel.

- This means that for any set of distinct points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ it must be

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \kappa(\mathbf{x}_i, \mathbf{x}_j) > 0$$

  for any choice of the constants $c_1, \ldots, c_n$ such that not all c_i are equal to 0.

- Equivalently, the square Gram matrix G defined as

$$G = \begin{pmatrix} \kappa(\mathbf{x}_1, \mathbf{x}_1) & \kappa(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \kappa(\mathbf{x}_1, \mathbf{x}_n) \\ \kappa(\mathbf{x}_2, \mathbf{x}_1) & \kappa(\mathbf{x}_2, \mathbf{x}_2) & \cdots & \kappa(\mathbf{x}_2, \mathbf{x}_n) \\ \cdots & \cdots & \cdots & \cdots \\ \kappa(\mathbf{x}_n, \mathbf{x}_1) & \kappa(\mathbf{x}_n, \mathbf{x}_2) & \cdots & \kappa(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix}$$

  must have positive eigenvalues.

- A collection of positive definite kernels is known in the literature and can be constructed by applying suitable rules.

Gaussian processes

Given a gaussian process $p(f) = \mathcal{GP}(m, \kappa)$, then for any set of items $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, the distribution of $f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)$ is a gaussian

$$(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)) \sim \mathcal{N}(\boldsymbol{\mu}(\mathbf{X})|\Sigma(\mathbf{X}))$$

where

- $\boldsymbol{\mu}(\mathbf{X}) = (m(\mathbf{x}_1), \ldots, m(\mathbf{x}_n))^T$
- $\Sigma(\mathbf{X})$ is the Gram matrix wrt $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of a kernel function $\kappa(\mathbf{x}, \mathbf{x}')$

As stated before, it is usually assumed that the mean vector is $\mathbf{0}$: different processes are then characterized only by their covariance kernel κ.

Sampling functions from gaussian processes

Given $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, a probability distribution on $f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)$ is then defined, as

$$p(f|\mathbf{X}) =$$

- For any finite subset $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of χ it is possible to sample from $p(f)$ the values of $f(\mathbf{x}_1), \ldots, f(\mathbf{x}_m)$ by gaussian sampling from $\mathcal{N}(f|\mathbf{0}, \Sigma(\mathbf{X}))$, where, as stated before

$$\Sigma(\mathbf{X})_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$$

RBF kernel

Clearly, different kernels provide different processes.

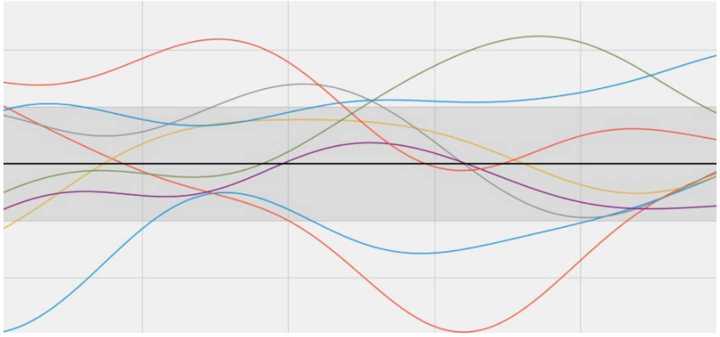
- One of the most applied kernel is the RBF kernel

$$\kappa(\mathbf{x}_1, \mathbf{x}_2) = \sigma^2 e^{-\frac{||\mathbf{x}_1 - \mathbf{x}_2||^2}{2\tau^2}}$$

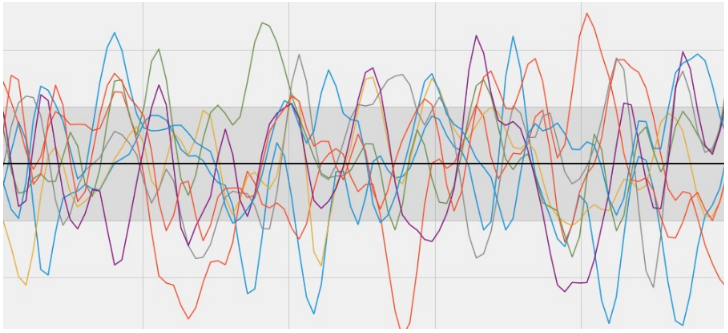
  which tends to assign higher covariance between $f(\mathbf{x}_1)$ and $f(\mathbf{x}_2)$ if $\mathbf{x}_1$ and $\mathbf{x}_2$ are nearby points.

- Functions drawn from a Gaussian process with RBF kernel tend to be smooth, since values computed for nearby points tend to be similar. Smoothing is larger for larger τ.

Samples of functions from $p(f)$. RBF kernel, larger τ and smoothing

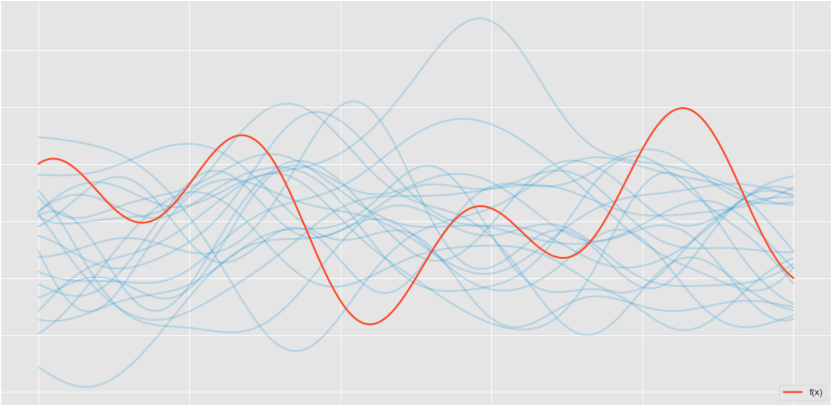


Samples of functions from $p(f)$. RBF kernel, smaller τ and smoothing



Gaussian process regression: no noise

- By the gaussian process definition, f is distributed as a multivariate gaussian such that the mean of any value $f(\mathbf{x})$ is $m(\mathbf{x})$ and the covariance of any pair $f(\mathbf{x}), f(\mathbf{x}')$ is $\kappa(\mathbf{x}, \mathbf{x}')$
- as a consequence, for any finite set of points $\mathbf{X}$, we have that $f(\mathbf{X})$ is distributed as a multivariate gaussian with mean $\boldsymbol{\mu}(\mathbf{X})$ defined as $\boldsymbol{\mu}(\mathbf{X})_i = m(\mathbf{x}_i)$ and covariance matrix $\Sigma(\mathbf{X})$, defined as $\Sigma(\mathbf{X})_{i,j} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$



- Let us now assume that for a set of points $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ the corresponding values $\mathbf{t} = (t_1, \ldots, t_n)^T$ are known
- that is, we assume that a training set $\mathbf{X}, \mathbf{t}$ is available, and we assume that the target values in the training set correspond exactly to the function value $t_i = f(\mathbf{x}_i)$, that is, there is no noise in the observations
- Note that in the probabilistic model of regression this is not true, since a (gaussian) error is assumed

By the model assumptions, if we consider an additional set of points $\overline{\mathbf{X}} = (\overline{\mathbf{x}}_1, \dots, \overline{\mathbf{x}}_m)^T$, the joint distribution of $f(\mathbf{X})$ and $f(\overline{\mathbf{X}})$ is a multivariate gaussian distribution with a certain mean $\boldsymbol{\mu}(\mathbf{X}, \overline{\mathbf{X}})$ and covariance $\Sigma(\mathbf{X}, \overline{\mathbf{X}})$ that, by the properties of gaussian distributions are

$$\boldsymbol{\mu}(\mathbf{X}, \overline{\mathbf{X}}) = (\boldsymbol{\mu}(\mathbf{X}), \boldsymbol{\mu}(\overline{\mathbf{X}}))^T$$

$$\Sigma(\mathbf{X}, \overline{\mathbf{X}}) = \begin{pmatrix} \Sigma(\mathbf{X}) & \Sigma(\overline{\mathbf{X}}, \mathbf{X}) \\ \Sigma(\overline{\mathbf{X}}, \mathbf{X})^T & \Sigma(\overline{\mathbf{X}}) \end{pmatrix}$$

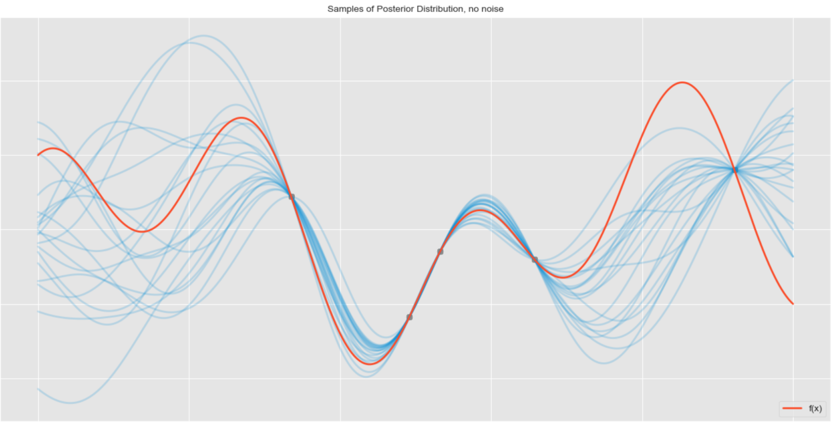
where

$$\Sigma(\overline{\mathbf{X}}, \mathbf{X}) = \begin{pmatrix} \kappa(\overline{\mathbf{x}}_1, \mathbf{x}_1) & \kappa(\overline{\mathbf{x}}_1, \mathbf{x}_2) & \cdots & \kappa(\overline{\mathbf{x}}_1, \mathbf{x}_n) \\ \kappa(\overline{\mathbf{x}}_2, \mathbf{x}_1) & \kappa(\overline{\mathbf{x}}_2, \mathbf{x}_2) & \cdots & \kappa(\overline{\mathbf{x}}_2, \mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa(\overline{\mathbf{x}}_m, \mathbf{x}_1) & \kappa(\overline{\mathbf{x}}_m, \mathbf{x}_2) & \cdots & \kappa(\overline{\mathbf{x}}_m, \mathbf{x}_n) \end{pmatrix}$$

The posterior distribution of $\mathbf{y} = f(\overline{\mathbf{X}})$, given $\mathbf{X}, \mathbf{t}$ can be derived by the gaussian distribution properties recalled above, and turns out to be a m-dimensional gaussian distribution itself with mean and covariance defined as

- $\overline{\boldsymbol{\mu}}_p = \mu(\mathbf{y}|\mathbf{X}, \mathbf{t}) = \boldsymbol{\mu}(\overline{\mathbf{X}}) + \Sigma(\mathbf{x}, \mathbf{X})\Sigma(\mathbf{X})^{-1}(\mathbf{t} - \boldsymbol{\mu}(\mathbf{X}))$
- $\overline{\Sigma}_p = \Sigma(\overline{\mathbf{X}}) - \Sigma(\mathbf{x}, \mathbf{X})\Sigma(\mathbf{X})^{-1}\Sigma(\mathbf{x}, \mathbf{X})^T$

Sample of functions from the posterior distribution



In particular, for the prediction of a single test point $\mathbf{x}$, the joint distribution of $(\mathbf{t}, f(\mathbf{x}))$ is a multivariate gaussian distribution with mean $\boldsymbol{\mu}(\mathbf{X}, \mathbf{x})$ and covariance $\Sigma(\mathbf{X}, \mathbf{x})$

$$\boldsymbol{\mu}(\mathbf{X}, \mathbf{x}) = (\boldsymbol{\mu}(\mathbf{X}), \mu(\mathbf{x}))^T$$

$$\Sigma(\mathbf{X}, \mathbf{x}) = \begin{pmatrix} \Sigma(\mathbf{X}) & \Sigma(\mathbf{x}, \mathbf{X}) \\ \Sigma(\mathbf{x}, \mathbf{X})^T & \Sigma(\mathbf{x}, \mathbf{x}) \end{pmatrix}$$

where

$$\Sigma(\mathbf{x}, \mathbf{X}) = (\kappa(\mathbf{x}, \mathbf{x}_1), \kappa(\mathbf{x}, \mathbf{x}_2), \dots, \kappa(\mathbf{x}_n, \mathbf{x}_n))^T$$

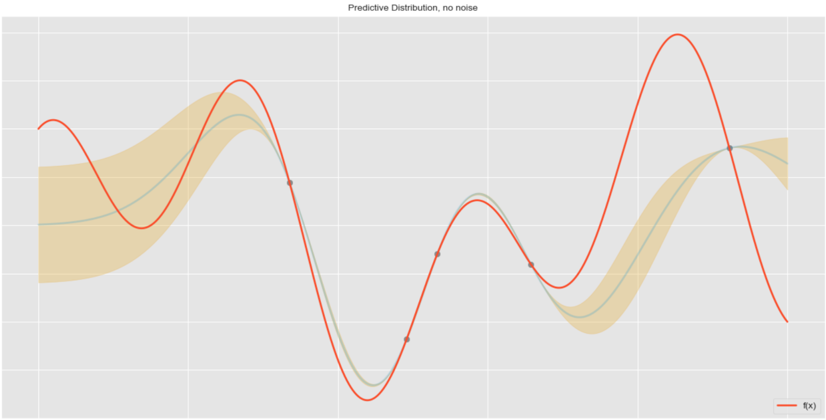
and

$$\Sigma(\mathbf{x}, \mathbf{x}) = \kappa(\mathbf{x}, \mathbf{x})$$

As a consequence, the predictive distribution of $y = f(\mathbf{x})$ is

$$m_p(y|\mathbf{X}, f) = m(\mathbf{x}) + \Sigma(\mathbf{x}, \mathbf{X})\Sigma(\mathbf{X})^{-1}(\mathbf{t} - \boldsymbol{\mu}(\mathbf{X}))$$

$$\sigma^2 = \Sigma_p(\mathbf{x}, \mathbf{x}) = \kappa(\mathbf{x}, \mathbf{x}) - \Sigma(\mathbf{x}, \mathbf{X})\Sigma(\mathbf{X})^{-1}\Sigma(\mathbf{x}, \mathbf{X})^T$$



In this case, an interpolation of the given values has been performed: $f(\mathbf{x}_i) = t_i$ for all possible functions, sampled from $f(\mathbf{x}|\mathbf{X}, f)$.

It results, in fact, for all $\mathbf{x}_i \in \mathbf{X}$,

$$m(\mathbf{x}_i|\mathbf{X}, f) = t_i$$
$$\sigma^2 = 0$$

Gaussian process regression: gaussian noise

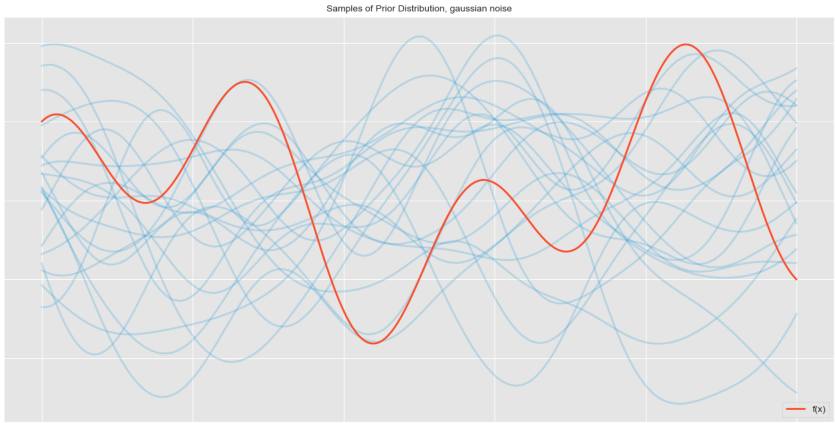
Let us now assume, as usual, that $p(t_i|f, \mathbf{x}_i) = \mathcal{N}(f(\mathbf{x}_i), \sigma_f^2)$

That is, the value t_i observed for variable $\mathbf{x}_i$ differs from the one obtained as $f(\mathbf{x}_i)$ by a gaussian and independent noise

$$t_i = f(\mathbf{x}_i) + \varepsilon \qquad\qquad p(\varepsilon) = \mathcal{N}(\varepsilon|0, \sigma_f^2)$$

that is, $p(\mathbf{t}|f) = \mathcal{N}(\mathbf{t}|f, \sigma^2\mathbf{I})$

- f is now distributed as a multivariate gaussian with known mean $\boldsymbol{\mu}(\mathbf{X}) = (m(\mathbf{x}_i), \ldots, m(\mathbf{x}_n))^T$ and covariance matrix $\hat{\Sigma}(\mathbf{X}) = \Sigma(\mathbf{X}) + \sigma_f^2\mathbf{I}$, defined by κ and σ_f^2



- Let us now assume that a training set $\mathbf{X}, \mathbf{t}$ is available such that the target values in the training set correspond approximately to the function value $t_i = f(\mathbf{x}_i) + \varepsilon$.

- In this case, for any new set of points $\overline{\mathbf{X}}$, the joint distribution of $(\mathbf{t}, f(\overline{\mathbf{X}}))$ is a multivariate gaussian distribution with mean $\boldsymbol{\mu}(\mathbf{X}, \overline{\mathbf{X}})$ and covariance $\Sigma(\mathbf{X}, \overline{\mathbf{X}})$

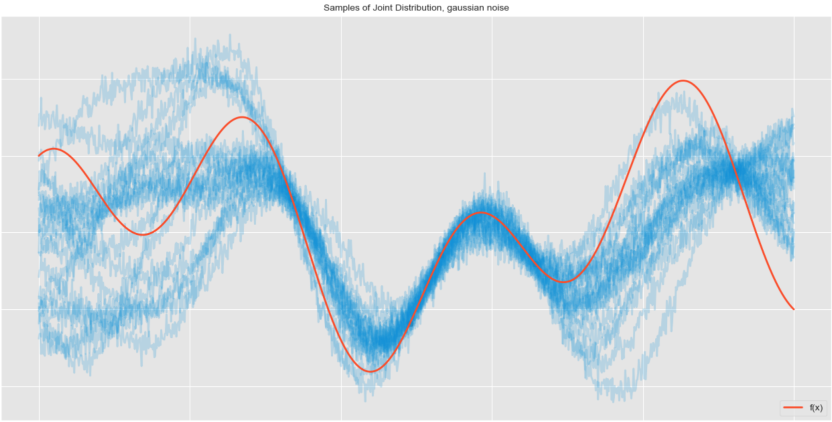
$$\boldsymbol{\mu}(\mathbf{X}, \overline{\mathbf{X}}) = (\boldsymbol{\mu}(\mathbf{X}), \boldsymbol{\mu}(\overline{\mathbf{X}}))^T$$

$$\Sigma(\mathbf{X}, \overline{\mathbf{X}}) = \begin{pmatrix} \hat{\Sigma}(\mathbf{X}) & \Sigma(\overline{\mathbf{X}}, \mathbf{X}) \\ \Sigma(\overline{\mathbf{X}}, \mathbf{X})^T & \Sigma(\overline{\mathbf{X}}) \end{pmatrix}$$

where in particular $\hat{\Sigma}(\mathbf{X}) = \begin{pmatrix} \kappa(\mathbf{x}_1, \mathbf{x}_1) + \sigma_f^2 & \kappa(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \kappa(\mathbf{x}_1, \mathbf{x}_n) \\ \kappa(\mathbf{x}_2, \mathbf{x}_1) & \kappa(\mathbf{x}_2, \mathbf{x}_2) + \sigma_f^2 & \cdots & \kappa(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa(\mathbf{x}_n, \mathbf{x}_1) & \kappa(\mathbf{x}_n, \mathbf{x}_2) & \cdots & \kappa(\mathbf{x}_n, \mathbf{x}_n) + \sigma_f^2 \end{pmatrix}$

The posterior distribution of $\mathbf{y} = f(\overline{\mathbf{X}})$, given $\mathbf{X}, \overline{\mathbf{X}}, \mathbf{t}$ can be again derived by the gaussian distribution properties, and turns out again to be a gaussian distribution with mean and covariance defined as

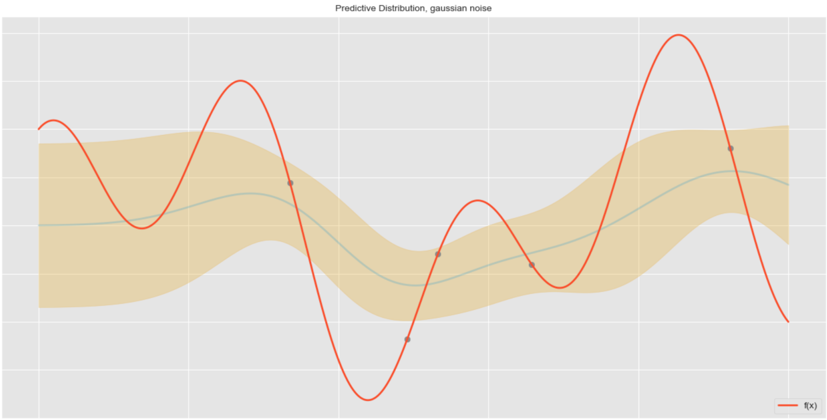
- $\overline{\boldsymbol{\mu}}_p = \boldsymbol{\mu}(\overline{\mathbf{X}}) + \Sigma(\mathbf{x}, \mathbf{X})\hat{\Sigma}(\mathbf{X})^{-1}(\mathbf{t} - \boldsymbol{\mu}(\mathbf{X}))$
- $\overline{\Sigma} = \Sigma(\overline{\mathbf{X}}) - \Sigma(\mathbf{x}, \mathbf{X})\hat{\Sigma}(\mathbf{X})^{-1}\Sigma(\mathbf{x}, \mathbf{X})^T$


Samples of Joint Distribution, gaussian noise

In particular, for a single test point $\mathbf{x}$, we have now that the corresponding predictive distribution is

$$m_p(y|\mathbf{X}, \mathbf{f}) = m(\mathbf{x}) + \Sigma(\mathbf{x}, \mathbf{X})\hat{\Sigma}(\mathbf{X})^{-1}(\mathbf{t} - \boldsymbol{\mu}(\mathbf{X}))$$

$$\sigma^2 = \kappa_p(\mathbf{x}, \mathbf{x}) = \kappa(\mathbf{x}, \mathbf{x}) - \Sigma(\mathbf{x}, \mathbf{X})\hat{\Sigma}(\mathbf{X})^{-1}\Sigma(\mathbf{x}, \mathbf{X})^T$$


Predictive Distribution, gaussian noise

Estimating kernel parameters

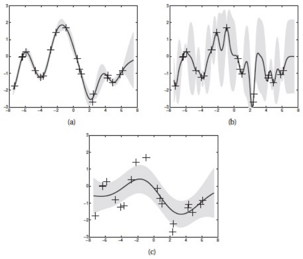
The predictive performance of gaussian processes depends exclusively on the suitability of the chosen kernel.

Let us consider the case of an RBF kernel. Then,

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 e^{-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)} + \sigma_y^2 \delta_{ij}$$

M can be defined in several ways: the simplest one is $\mathbf{M} = l^{-2}\mathbf{I}$.

Even in this simple case, varying the values of σ_f, σ_y, l returns quite different results.



(figure from K.Murphy “Machine learning: a probabilistic perspective” p. 519, with (l, σ_f, σ_y) equal to $(1, 1, 0.1)$, $(0.3, 1.08, 0.00005)$, $(3.0, 1.16, 0.8$