

732A96/TDDE15 Advanced Machine Learning

Gaussian Process Regression and Classification

Jose M. Peña
IDA, Linköping University, Sweden

Lectures 11: Kernels, Hyperparameter Learning and More

Contents

- ▶ Three Common Covariance Functions
- ▶ Learning the Hyperparameters of the Covariance Function
- ▶ More on Covariance Functions
- ▶ Lab: Algorithm 2.1 in Rasmussen and Williams

Literature

- ▶ Main source
 - ▶ Rasmussen, C. E. and Williams, K. I. *Gaussian Processes for Machine Learning*. MIT Press, 2006. Chapters 2.3, 5.1-5.4.1.
- ▶ Additional source
 - ▶ Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006. Chapters 6.4.3-6.4.4.

Three Common Covariance Functions

- ▶ Let $r = \|\mathbf{x} - \mathbf{x}'\|$.
- ▶ Squared exponential (SE):

$$k_{SE}(r) = \sigma_f^2 \exp \left\{ -\frac{r^2}{2\ell^2} \right\}$$

where $\sigma_f^2 > 0, \ell > 0$. Very smooth.

- ▶ Rational quadratic (RQ):

$$k_{RQ}(r) = \sigma_f^2 \left(1 + \frac{r^2}{2\alpha\ell^2} \right)^{-\alpha}$$

$\sigma_f^2 > 0, \ell > 0, \alpha > 0$. k_{RQ} is an infinite sum of k_{SE} with different ℓ . As $\alpha \rightarrow \infty$, $k_{RQ}(r) \rightarrow k_{SE}(r)$.

- ▶ Matérn:

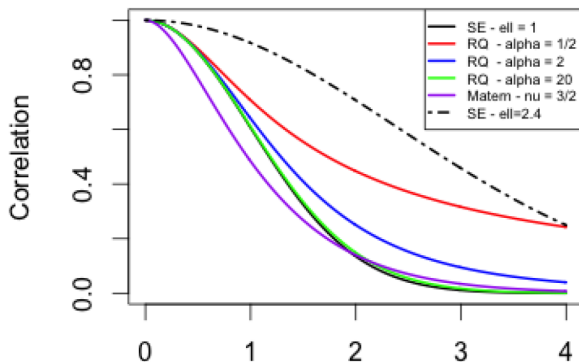
$$k_{Matern} = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}r}{\ell} \right)^\nu K_\nu \left(\frac{\sqrt{2\nu}r}{\ell} \right)$$

where $\sigma_f^2 > 0, \ell > 0, \nu > 0$, and K_ν is the modified Bessel function. As $\nu \rightarrow \infty$, $k_{Matern}(r) \rightarrow k_{SE}(r)$.

- ▶ Demo of `GaussianProcesses.R` and `KernLabDemo.R`.

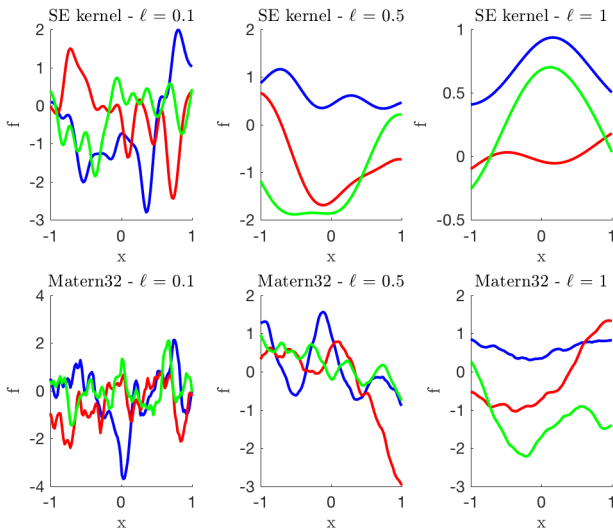
Three Common Covariance Functions

Correlation functions



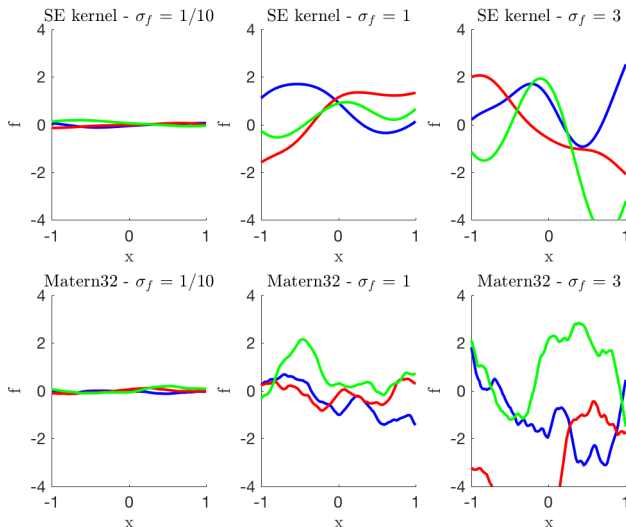
Three Common Covariance Functions

- The length scale ℓ determines the smoothness.



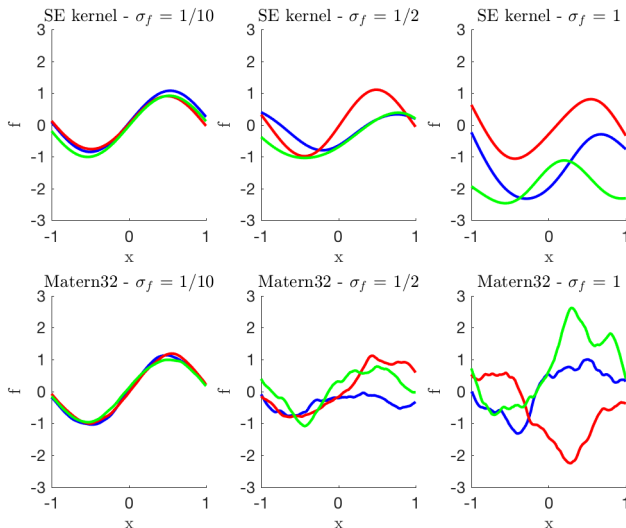
Three Common Covariance Functions

- The scale factor σ_f determines the variance.



Three Common Covariance Functions

- The mean can be arbitrary, e.g. $\sin(3x)$.



Learning the Hyperparameters of the Covariance Function

- ▶ Let θ denote the hyperparameters of the covariance function, i.e. $\theta = (\sigma_f, \ell)$ for k_{SE} , $\theta = (\sigma_f, \ell, \alpha)$ for k_{RQ} , and $\theta = (\sigma_f, \ell, \nu)$ for k_{Matern} .
- ▶ Choose the hyperparameters that maximize the marginal likelihood:

$$p(\mathbf{y}|\mathbf{X}, \theta) = \int p(\mathbf{y}|\mathbf{f}, \mathbf{X}, \theta) p(\mathbf{f}|\mathbf{X}, \theta) d\mathbf{f}$$

where $\mathbf{f}|\mathbf{X}, \theta \sim \mathcal{N}(0, K(\mathbf{X}, \mathbf{X}))$ and $\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma_n^2 I)$, which implies

$$\log p(\mathbf{y}|\mathbf{X}, \theta) = -\frac{1}{2} \mathbf{y}^T (K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I)^{-1} \mathbf{y} - \frac{1}{2} \log |K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I| - \frac{n}{2} \log 2\pi$$

which alternatively can be obtained directly from

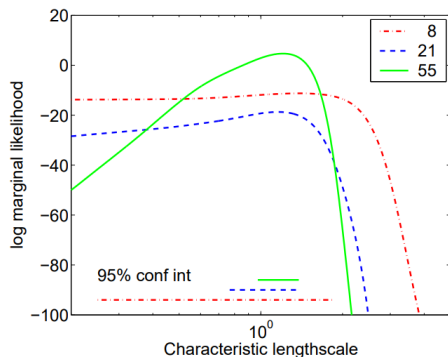
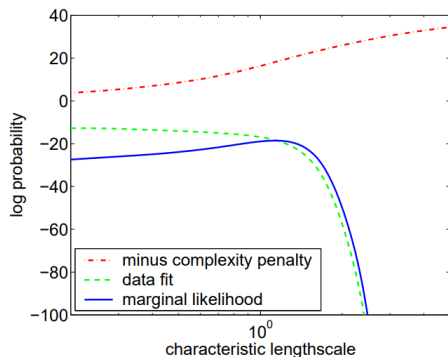
$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix}\right).$$

- ▶ In general, this is a non-convex optimization problem, and gradient methods are typically used. For most common covariance functions, the derivative of $K(\mathbf{X}, \mathbf{X})$ wrt θ is easy to compute.
- ▶ For a Bayesian approach, choose the hyperparameters that maximize the posterior distribution $p(\theta|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{X}, \theta)p(\theta)$. It typically requires MCMC sampling or Laplace approximation.
- ▶ The methods above can also be used to select among covariance functions, i.e. simply include them as hyperparameters. Cross-validation is also an option.

Learning the Hyperparameters of the Covariance Function

$$\log p(\mathbf{y}|\mathbf{X}, \theta) = -\frac{1}{2}\mathbf{y}^T (K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y} - \frac{1}{2} \log |K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}| - \frac{n}{2} \log 2\pi$$

= data fit - model complexity - normalization constant.



More on Covariance Functions

- ▶ Anisotropic version of isotropic covariance function (i.e., it depends only on r) by setting $r^2 = (\mathbf{x} - \mathbf{x}')^T \mathbf{M} (\mathbf{x} - \mathbf{x}')$ where \mathbf{M} is positive definite.
- ▶ Automatic Relevance Determination: $\mathbf{M} = \text{diag}(\ell_1^{-2}, \dots, \ell_D^{-2})$, i.e. different length scales for different dimensions. In other words, ARM performs variable selection since a large ℓ_j means that the j -th dimension is essentially irrelevant for $f(\mathbf{x})$.
- ▶ Periodic kernel with period d : $k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp \left\{ - \frac{2 \sin^2(\pi |\mathbf{x} - \mathbf{x}'|/d)}{\ell^2} \right\}$.
- ▶ The sum and product of two kernels is a kernel. For instance:
 - ▶ $k_{ARD}(\mathbf{x}, \mathbf{x}') = \prod_{d=1}^D k_{SE, \ell_d}(x_d, x'_d)$.
 - ▶ $k_{Periodic}(\mathbf{x}, \mathbf{x}') \cdot k_{SE}(\mathbf{x}, \mathbf{x}')$: Close peaks more dependent than distant ones.

Lab: Algorithm 2.1 in Rasmussen and Williams

input: X (inputs), \mathbf{y} (targets), k (covariance function), σ_n^2 (noise level), \mathbf{x}_* (test input)

2: $L := \text{cholesky}(K + \sigma_n^2 I)$

$\alpha := L^\top \backslash (L \backslash \mathbf{y})$ } predictive mean eq. (2.25)

4: $\bar{f}_* := \mathbf{k}_*^\top \alpha$

$\mathbf{v} := L \backslash \mathbf{k}_*$ } predictive variance eq. (2.26)

6: $\mathbb{V}[f_*] := k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{v}^\top \mathbf{v}$

$\log p(\mathbf{y}|X) := -\frac{1}{2} \mathbf{y}^\top \alpha - \sum_i \log L_{ii} - \frac{n}{2} \log 2\pi$ eq. (2.30)

8: **return:** \bar{f}_* (mean), $\mathbb{V}[f_*]$ (variance), $\log p(\mathbf{y}|X)$ (log marginal likelihood)

- ▶ The algorithm uses Cholesky decomposition instead of matrix inversion because it is faster and numerically more stable.
- ▶ It returns the predictive distribution for noise-free test data, i.e. \mathbf{f}_* . Add σ_n^2 to the predictive variances to obtain the distribution for noisy test data, i.e. \mathbf{y}_*
- ▶ It is presented for a single test case but it also works for several test cases.
- ▶ $K = K(X, X)$.
- ▶ $K_* = K(X, X_*)$.
- ▶ $\mathbf{k}_* = k(\mathbf{x}_*) = K(X, \mathbf{x}_*)$.
- ▶ $L = \text{cholesky}(A) \Rightarrow A = LL^\top \Rightarrow A^{-1} = (L^\top)^{-1} L^{-1} = (L^{-1})^\top L^{-1}$ and $|A| = \det(A) = \det(L) \det(L^\top) = 2 \prod_i L_{ii}$.
- ▶ $L \mathbf{y} = \text{solve}(L, \mathbf{y}) = L^{-1} \mathbf{y}$.

Contents

- ▶ Three Common Covariance Functions
- ▶ Learning the Hyperparameters of the Covariance Function
- ▶ More on Covariance Functions
- ▶ Lab: Algorithm 2.1 in Rasmussen and Williams

Thank you