AutoML: Gaussian Processes

Mean Functions for Gaussian Processes

Bernd Bischl Frank Hutter Lars Kotthoff
Marius Lindauer Joaquin Vanschoren

The Role of Mean Functions I

- It is common but by no means necessary to consider GPs with zero-mean functions:

$$m(\mathbf{x}) \equiv 0$$

- This is not necessarily a drastic limitation, since the mean of the posterior process is not confined to be zero:

$$\boldsymbol{f}_*|\mathbf{X}_*, \mathbf{X}, \boldsymbol{f} \sim \mathcal{N}(\mathbf{K}_*^\top \mathbf{K}^{-1} \boldsymbol{f}, \mathbf{K}_{**} - \mathbf{K}_*^\top \mathbf{K}^{-1} \mathbf{K}_*).$$

- There are several reasons why one might wish to explicitly model a mean function, including interpretability, convenience of expressing prior informations, etc.

The Role of Mean Functions II

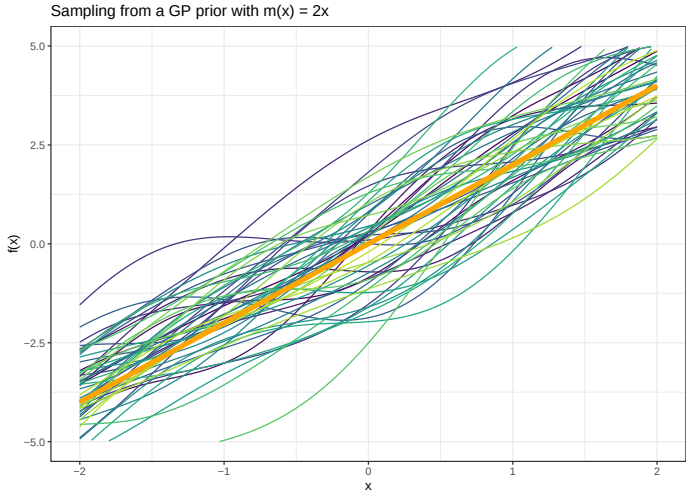
- When assuming a non-zero mean GP prior $\mathcal{G}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')$ with mean $m(\mathbf{x})$, the predictive mean becomes:

$$m(\mathbf{X}_*) + \mathbf{K}_* \mathbf{K}_y^{-1} \left(\boldsymbol{y} - m(\mathbf{X}) \right)$$

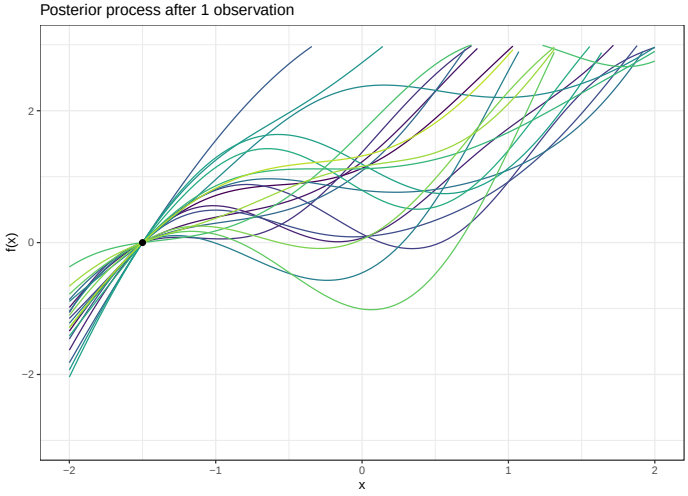
  but the predictive variance remains unchanged.

- Gaussian processes with non-zero mean Gaussian process priors are also called **Gaussian processes with trend**.

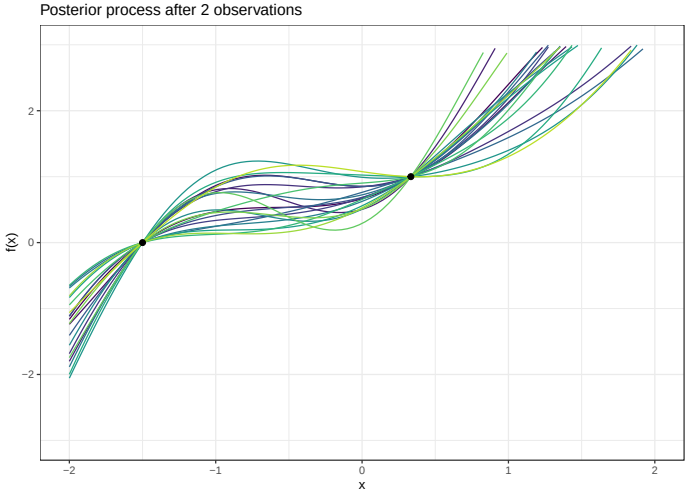
The Role of Mean Functions III



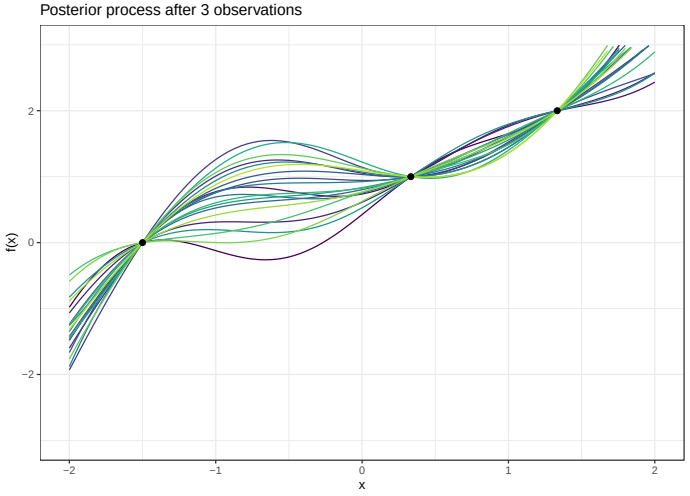
The Role of Mean Functions IV



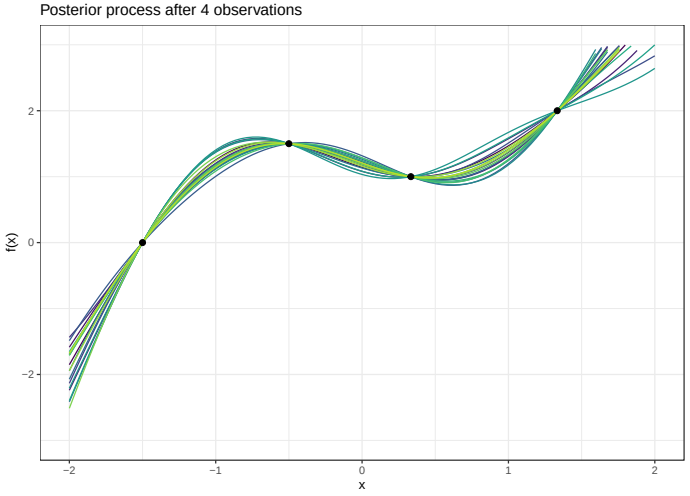
The Role of Mean Functions V



The Role of Mean Functions VI



The Role of Mean Functions VII



The Role of Mean Functions VIII

- In practice it is often difficult to specify a fixed mean function. Instead, it may be more convenient to specify a few fixed basis functions, whose coefficients, $\boldsymbol{\beta}$, are to be inferred from the data.
- Consider
$$g(\mathbf{x}) = f(\mathbf{x}) + \boldsymbol{h}(\mathbf{x})^\top \boldsymbol{\beta}, \text{ with } f(\mathbf{x}) \sim \mathcal{G}\left(0, k(\mathbf{x}, \mathbf{x}')\right)$$
where $f(\mathbf{x})$ is a zero mean GP, $\boldsymbol{h}(\mathbf{x})$ are a set of fixed basis functions, and $\boldsymbol{\beta}$ are additional parameters.
- This formulation expresses that the data is close to a global linear model with the residuals being modelled by a GP.
- By taking the prior on $\boldsymbol{\beta}$ to be Gaussian, $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{b}, B)$, another GP with an added contribution in the covariance function can be obtained:
$$g(\mathbf{x}) \sim \mathcal{G}\left(\mathbf{h}(\mathbf{x}^\top \mathbf{b}, k\left(\mathbf{x}, \mathbf{x}'\right) + \mathbf{h}(\mathbf{x})^\top B\ \mathbf{h}\left(\mathbf{x}'\right)\right)$$

The Role of Mean Functions IX

- The mean and covariance functions of $g(\mathbf{x})$ will be then:

$$\overline{\mathbf{g}}\left(X_*\right) = \overline{\mathbf{f}}\left(X_*\right) + R^\top \overline{\boldsymbol{\beta}},$$
$$\operatorname{cov}\left(\mathbf{g}_*\right) = \operatorname{cov}\left(\mathbf{f}_*\right) + R^\top (B^{-1} + HK_y^{-1}H^\top)^{-1} R,$$

  where $\overline{\boldsymbol{\beta}} = (B^{-1} + HK_y^{-1}H^\top)^{-1}(HK_y^{-1}\mathbf{y} + B^{-1}\mathbf{b})$ and $R = H_* - HK_y^{-1}K_*$.

- Notice that $\overline{\boldsymbol{\beta}}$ is the mean of the global linear model parameters and the H matrix collects the $\mathbf{h}(\mathbf{x})$ vectors for all training and test cases.

- Hence, the predictive mean is the sum of the mean linear output and what the GP model predicts from the residuals. The covariance is the usual covariance term plus a non negative contribution.

The Role of Mean Functions X

- In the limiting case where the prior on the $\boldsymbol{\beta}$ becomes vague ($B^{-1} \to O$), the predictive distribution becomes independent of $\boldsymbol{b}$:

$$\overline{\mathbf{g}}(X_*) = \overline{\mathbf{f}}(X_*) + R^\top \overline{\boldsymbol{\beta}},$$
$$\operatorname{cov}(\mathbf{g}_*) = \operatorname{cov}(\mathbf{f}_*) + R^\top (H K_y^{-1} H^\top)^{-1} R,$$

  where the limiting $\overline{\boldsymbol{\beta}} = (H K_y^{-1} H^\top)^{-1} H K_y^{-1} \mathbf{y}$.

- To make predictions under the limit $B^{-1} \to O$, it is important to use the above equation. Otherwise, by naïvely plugging the modified covariance function into the standard prediction equations, the entries of the covariance function tend to infinity making it unsuitable for numerical implementation.

The Role of Mean Functions XI

- The marginal likelihood for the model with a Gaussian prior $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{b}, B)$ can be expressed as what follows (the explicit mean has been included).

$$\begin{aligned}\log p(\mathbf{y}|X, \mathbf{b}, B) = &-\frac{1}{2}\left(H^\top \mathbf{b} - \mathbf{y}\right)^\top \left(K_y + H^\top B H\right)^{-1} \left(H^\top \mathbf{b} - \mathbf{y}\right) \\ &-\frac{1}{2}\log\left|K_y + H^\top B H\right| - \frac{n}{2}\log 2\pi.\end{aligned}$$

- We are interested in exploring the limit where $B^{-1} \to O$, i.e. when the prior is vague. In this limit the mean of the prior is irrelevant, so without loss of generality we assume for now that the mean is zero, $\mathbf{b} = \mathbf{0}$.

The Role of Mean Functions XII

- This assumption gives:

$$\log p(\mathbf{y}|X, \mathbf{b} = \mathbf{0}, B) = -\frac{1}{2}\mathbf{y}^\top K_y^{-1}\mathbf{y} + \frac{1}{2}\mathbf{y}^\top C\mathbf{y} - \frac{1}{2}\log|K_y| - \frac{1}{2}\log|B| - \frac{1}{2}\log|A| - \frac{n}{2}\log 2\pi,$$

  where $A = B^{-1} + HK_y^{-1}H^\top$ and $C = K_y^{-1}H^\top A^{-1}HK_y^{-1}$.

- The behaviour of the log marginal likelihood in the limiting case can be explored now. The variances of the Gaussian in the directions spanned by columns of $H^\top$ will become infinite, and it will require special treatment.

- The log marginal likelihood consists of three terms: a quadratic form in $\mathbf{y}$, a log determinant term, and a term involving $\log 2\pi$.

The Role of Mean Functions XIII

- Performing an eigendecomposition of the covariance matrix, the contributions to quadratic form term from the infinite-variance directions will be zero. However, the log determinant term will tend to minus infinity.

- The standard solution in this case is to project $\mathbf{y}$ onto the directions orthogonal to the span of $H^\top$ and compute the marginal likelihood in this subspace.

- By taking m to denote the rank of $H^\top$, we can discard the terms $-\frac{1}{2}\log|B| - \frac{m}{2}\log 2\pi$ from the previous equation to give:

$$\log p(\mathbf{y}|X) = -\frac{1}{2}\mathbf{y}^\top K_y^{-1}\mathbf{y} + \frac{1}{2}\mathbf{y}^\top C\mathbf{y} - \frac{1}{2}\log|K_y| - \frac{1}{2}\log|A| - \frac{n-m}{2}\log 2\pi,$$

where $A = HK_y^{-1}H^\top$ and $C = K_y^{-1}H^\top A^{-1}HK_y^{-1}$.