ADVANCED MACHINE LEARNING

Non-linear regression techniques
Part – II

Gaussian Process Regression

Probabilistic Regression (PR)

Creating the model

PR is a statistical approach to classical linear regression that estimates the relationship between zero-mean variables y and x by building a linear model:

$$y = f(x, w) = w^T x, \quad w, x \in \mathbb{R}^N$$

Probabilistic Regression (PR)

Creating the model

If one assumes that the observed values of *y* differ from f(x) by an additive noise ε that follows a zero-mean Gaussian distribution (such an assumption consists of putting a *prior distribution* over the noise), then:

$$y = w^T x + \varepsilon, \quad \text{with } \varepsilon = N\left(0, \sigma_\varepsilon^{\,2}\right)$$

The addition of a Normal random variable to a constant variable leads *y* to become a Normal random variable

Where have we seen this before?

Answer: RVM / RVR

Probabilistic Linear Regression

Maximum Likelihood Estimation

Training set of M pairs of data points $\{X, \mathbf{y}\} = \{x^i, y^i\}_{i=1}^{M}$.

Likelihood of the regressive model

$$\mathbf{y} = w^T X + N\left(0, \sigma_\varepsilon^{\,2}\right)$$

$$\Rightarrow \mathbf{y} \sim p\left(\mathbf{y} \mid X, w, \sigma_\varepsilon\right)$$

Parameters of the model.

Assume that the data points are independently and identically distributed (i.i.d), the likelihood is:

$$p\left(\mathbf{y} \mid X, w, \sigma_\varepsilon\right) \sim \prod_{i=1}^{M} p\left(y^i \mid x^i, w, \sigma_\varepsilon\right)$$

$$= \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} \exp\left(-\frac{\left(y^i - w^T x^i\right)^2}{2\sigma_\varepsilon^{\,2}}\right)$$

Probabilistic Linear Regression

Maximum Likelihood Estimation

Training set of M pairs of data points $\{X, \mathbf{y}\} = \{x^i, y^i\}_{i=1}^{M}$.

Likelihood of the regressive model

$$\mathbf{y} = w^T X + N\left(0, \sigma_\varepsilon{}^2\right)$$

$$\Rightarrow \mathbf{y} \sim p\left(\mathbf{y} \mid X, w, \sigma_\varepsilon\right)$$

Fix σ_ε

$$\mathbf{w}_{\text{MLE}} = \arg\max_{\mathbf{w}} p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \boldsymbol{\sigma}_\epsilon^2 \mathbf{I}_M)$$

Closed-form solution:

$$\mathbf{w}_{\text{MLE}} = \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{X}\mathbf{y} \rightarrow \text{ Same with the OLS estimator}$$

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{N}\left(\mathbf{w}^T_{\text{MLE}}\mathbf{X} - \mathbf{y}\right)^T\left(\mathbf{w}^T_{\text{MLE}}\mathbf{X} - \mathbf{y}\right) \rightarrow \text{ Mean squared prediction error}$$

Probabilistic Linear Regression

Maximum a Posteriori (MAP) Estimation

Training set of M pairs of data points $\{X, \mathbf{y}\} = \{x^i, y^i\}_{i=1}^{M}$.

Likelihood of the regressive model

$$\mathbf{y} = w^T X + N(0, \sigma_\varepsilon^2)$$

$$\Rightarrow \mathbf{y} \sim p(\mathbf{y} \mid X, w, \sigma_\varepsilon)$$

Hyperparameter set by the user.

w is treated as a random variable.
Set a prior on distribution of w:

$$p(w) = N(0, \Sigma_w) \propto \exp\left(-\frac{1}{2} w^T \Sigma_w^{-1} w\right)$$

Probabilistic Linear Regression

Maximum a Posteriori (MAP) Estimation

Prior distribution over the weights: $\mathbf{w} \sim \mathcal{N}\left(\mu_{\mathbf{w}}, \Sigma_{\mathbf{w}}\right)$

Model Likelihood: $p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \boldsymbol{\sigma}_{\epsilon}^{2}\mathbf{I}_{M})$

Bayes Rule: $\text{posterior} = \dfrac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$

$$p\left(\mathbf{w} \mid \mathbf{X}, \mathbf{y}\right) = \frac{p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \boldsymbol{\sigma}_{\epsilon}^{2})\, p\left(\mathbf{w}\right)}{p\left(\mathbf{y} \mid \mathbf{X}\right)}$$

Prior and likelihood are conjugate distributions. The posterior has a closed form solution:

$$p\left(\mathbf{w} \mid \mathbf{X}, \mathbf{y}\right) \sim \mathcal{N}\left(\mu_{\mathbf{w}|\mathbf{X},\mathbf{y}}, \mathbf{A}^{-1}\right)$$

Probabilistic Linear Regression

Maximum a Posteriori (MAP) Estimation

MAP estimates derive from the expected value of the posterior distribution:

$$\mathbf{w}_{\text{MAP}} = \mathbb{E}\left\{ p\left(\mathbf{w} \mid \mathbf{x}, \mathbf{Y}\right)\right\}$$

$$\Sigma_{\text{MAP}}^{-1} = \mathbf{A} = \frac{1}{\sigma_\epsilon^2}\mathbf{X}\mathbf{X}^T + \Sigma_{\mathbf{w}}^{-1}$$

$$\mathbf{w}_{\text{MAP}} = \mu_{\mathbf{w}|\mathbf{X},\mathbf{y}} = \mathbf{A}^{-1}\Sigma_{\mathbf{w}}^{-1}\mu_{\mathbf{w}} + \frac{1}{\sigma_\epsilon^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{y}$$

In the special case when the distribution is zero mean and $\Sigma_w = \tau I$, probabilistic regression reduces to ridge regression with $\lambda = \frac{\sigma_\epsilon^2}{\tau}$.

Ridge optimal regressor: $w^* = \left(XX^T + \lambda I\right)^{-1} X\mathbf{y}$

Probabilistic Linear Regression

Posterior Predictive Distribution

The predictive distribution is a distribution over output y.

$$p(y \mid x, X, \mathbf{y}) = \int \underbrace{p(y \mid x, w)}_{\text{Likelihood}} \underbrace{p(w \mid X, \mathbf{y})}_{\textit{Posterior}} dw$$

The integral has a closed form solution in the case of Normal/Gauss distributions

$$p(y \mid x, X, \mathbf{y}) \sim N\left(\mu_y(x), \sigma_y^2(x)\right),$$

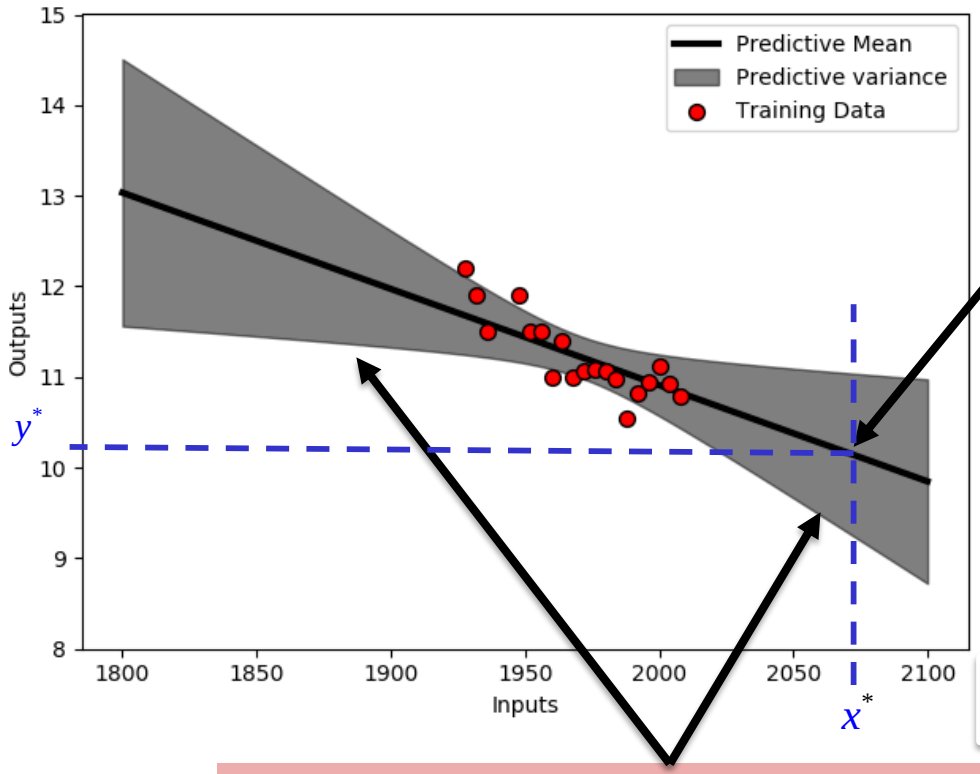
$$\mu_y(x) = x^T \mathbf{w}_{\text{MAP}} = \frac{1}{\sigma^2} x^T A^{-1} X\mathbf{y},$$

$$\sigma_y^2(x) = x^T A^{-1} x$$

$$A = \frac{1}{\sigma^2} XX^T + \Sigma_w^{-1}$$

Probabilistic Linear Regression

Posterior Predictive Distribution



Estimate y given a test point x^* :

$$y^* = E\left\{ p(y \mid x^*, X, \mathbf{y}\right\} = \frac{1}{\sigma^2} x^{*T} A^{-1} X\mathbf{y}$$

Training datapoints

Testing point

The variance gives a measure of the uncertainty of the prediction:

$$\text{var}\left\{ p(y \mid x\right\} = x^T A^{-1} x$$

Probabilistic Linear Regression

Marginal Likelihood

Bayesian regression depends on hyperparameter for priors on noise and *w*.

The marginal likelihood provides a metric to evaluate how well our model explains the observed data and can be used to select these hyperparameters.

$$p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \boldsymbol{\sigma}_\epsilon^2)\, p(\mathbf{w})}{\underbrace{p(\mathbf{y} \mid \mathbf{X})}_{MarginalLikelihood}}$$

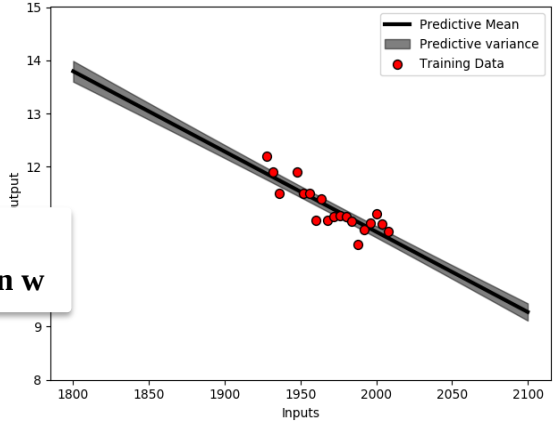
$$p(\mathbf{y} \mid \mathbf{X}) = \int p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \boldsymbol{\sigma}_\epsilon^2)\, p(\mathbf{w})\, d\mathbf{w}$$

With a Gaussian prior and Gaussian likelihood, the marginal is also Gaussian and has a closed form solution.

Probabilistic Linear Regression

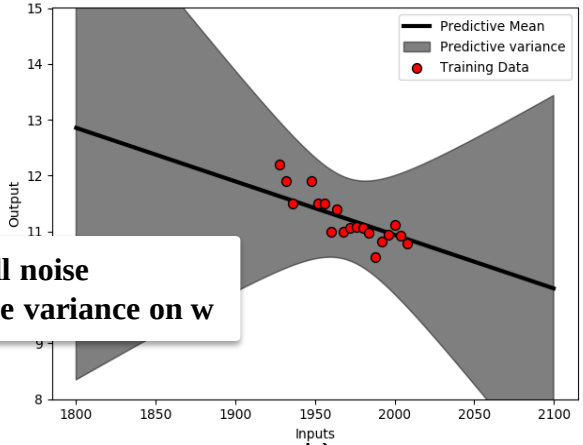
Marginal Likelihood

Small noise
Small variance on w



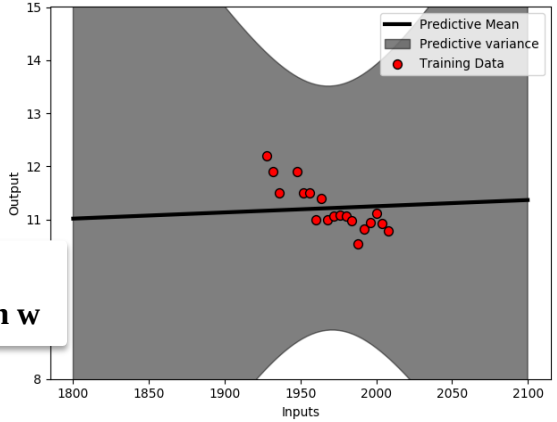
a)
33.38

Small noise
Large variance on w

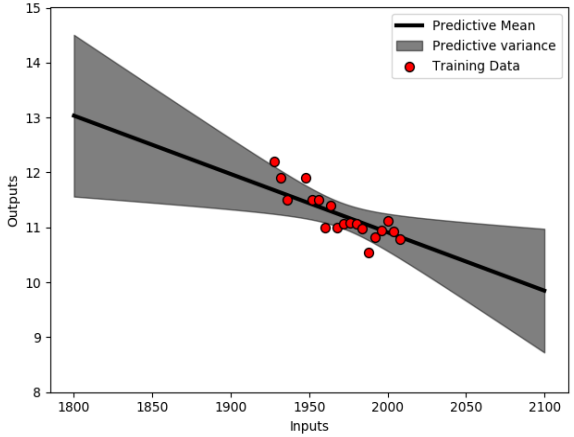


b)
47.19

Large noise
Large variance on w



c)
67.19



c)
27.32

Negative Log-Marginal Likelihood for different models

Gaussian Process Regression (GPR)

From linear to nonlinear probabilistic regression

Linear Probabilistic Regression

$$y = w^T x + N\left(0, \sigma_\varepsilon{}^2\right)$$

$$p(y \mid x, X, \mathbf{y}) = N\left(\frac{1}{\sigma^2} x^T A^{-1} X\mathbf{y}, x^T A^{-1} x\right),$$

$$A = \frac{1}{\sigma_\varepsilon{}^2} XX^T + \Sigma_w^{-1}$$

Linear Probabilistic Regression in feature space

$$y = w^T \phi(x) + \varepsilon, \qquad \varepsilon \sim N\left(0, \sigma_\varepsilon{}^2\right)$$

$$p(y \mid x, X, \mathbf{y}) = N\left(\frac{1}{\sigma_\varepsilon{}^2} \phi(x)^T A^{-1} \Phi(X)\mathbf{y}, \phi(x)^T A^{-1} \phi(x)\right)$$

with $A = \sigma_\varepsilon{}^{-2} \Phi(X)\Phi(X)^T + \Sigma_w^{-1}$

Inner product in feature space

Gaussian Process Regression (GPR)

From linear to nonlinear probabilistic regression

Define the kernel as: $k(x, x') = \phi(x)^T \Sigma_w \phi(x')$

and apply on the mean and the variance

$$y = E\{p(y \mid x, X, \mathbf{y})\} = \sum_{i=1}^{M} \alpha_i k(x, x^i)$$

with $\alpha = \left[K(X, X) + \sigma_\varepsilon{}^2 I\right]^{-1} \mathbf{y}$

$\alpha_i > 0 \quad \forall i$

→ All datapoints are used in the computation!

$$\text{var}(p(y \mid x)) = k(x, x) - K(x, X)\left[K(X, X) + \sigma_\varepsilon{}^2 I\right]^{-1} K(X, x)$$

Gaussian Process Regression (GPR)

Hyperparameters

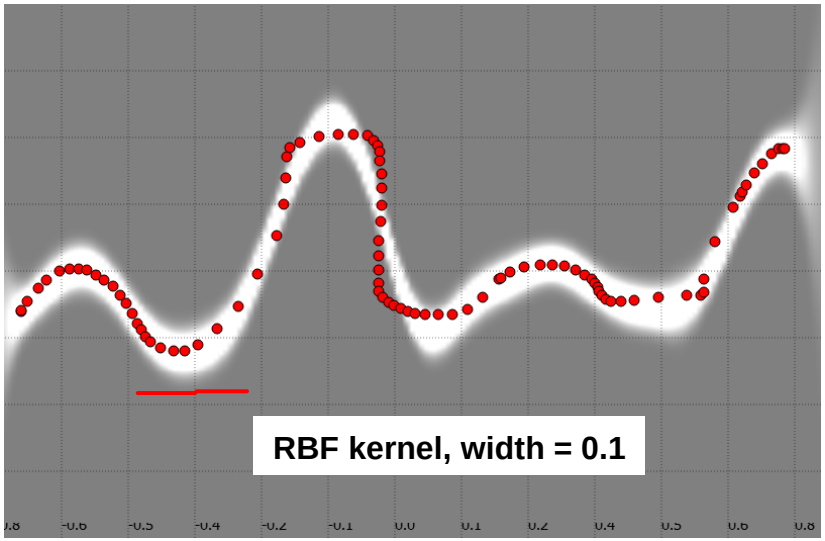
The choice of kernel and its hyperparameters will strongly influence the goodness of the fit.

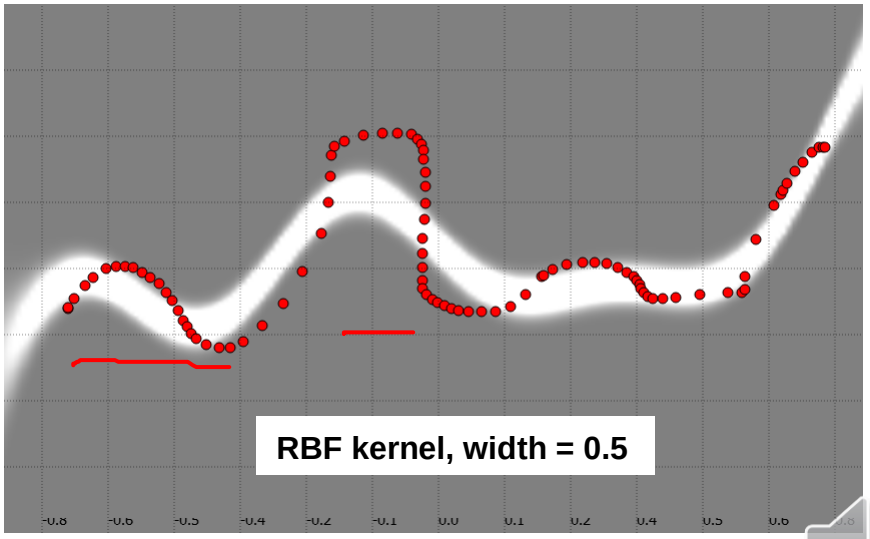
$$y = \sum_{i=1}^{M} \alpha_i k\left(x, x^i\right)$$

$$\text{with } \alpha = \left[K\left(X, X\right) + \sigma_\varepsilon{}^2 I\right]^{-1} \mathbf{y}$$

$$k\left(x, x'\right) = e^{-\frac{\|x - x'\|}{l}}$$

l: lengthscale parameter
~ Kernel Width





Gaussian Process Regression (GPR)

Hyperparameters

The value for the noise needs to be pre-set by hand.
It influence estimate of the expectation and variance of the model.

$$y = \sum_{i=1}^{M} \alpha_i k\left(x, x^i\right)$$

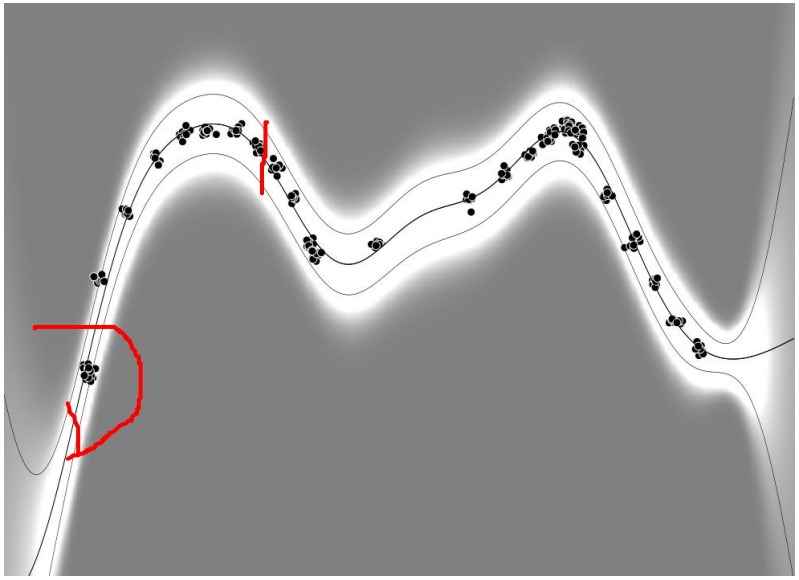
$$\text{with } \alpha = \left[K\left(X, X\right) + \sigma_\varepsilon{}^2 I\right]^{-1} \mathbf{y}$$

$$\text{var}\left(p\left(y \mid x\right)\right) = K\left(x, x\right) - K\left(x, X\right)\left[K\left(X, X\right) + \sigma_\varepsilon{}^2 I\right]^{-1} K\left(X, x\right)$$

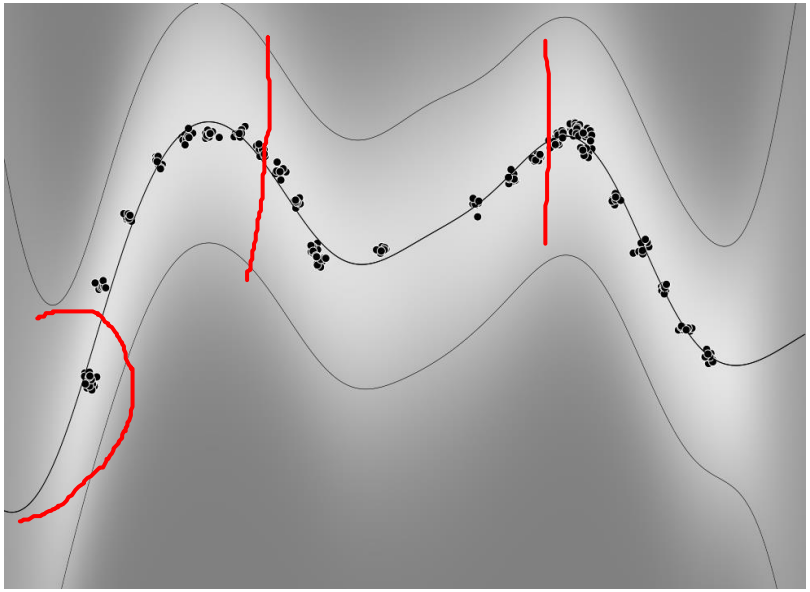
Gaussian Process Regression (GPR)

Hyperparameters

The larger the noise, the more uncertainty. The noise is <=1.



Low noise: $\sigma=0.05$



High noise: $\sigma=0.2$

Gaussian Process Regression (GPR)
Hyperparameters Tuning

Hyper-parameters: kernel's parameters and noise variance

One can automatically tune these hyperparameters by either:

- Crossvalidation
- Minimizing marginal Likelihood:

$$-\log\left(\mathbf{y} \mid \mathbf{X}, \sigma_\varepsilon, l\right) = 0.5(\underbrace{\mathbf{y}^T \left(\mathbf{K} + \sigma_\varepsilon I\right)^{-1} \mathbf{y}}_{\text{Fit}} + \underbrace{\log \left|\mathbf{K} + \sigma_\varepsilon I\right|}_{\text{Complexity}} + \underbrace{\frac{M}{2} \log 2\pi}_{\textit{Normalization}})$$

Automatically provides trade-off between fit and complexity

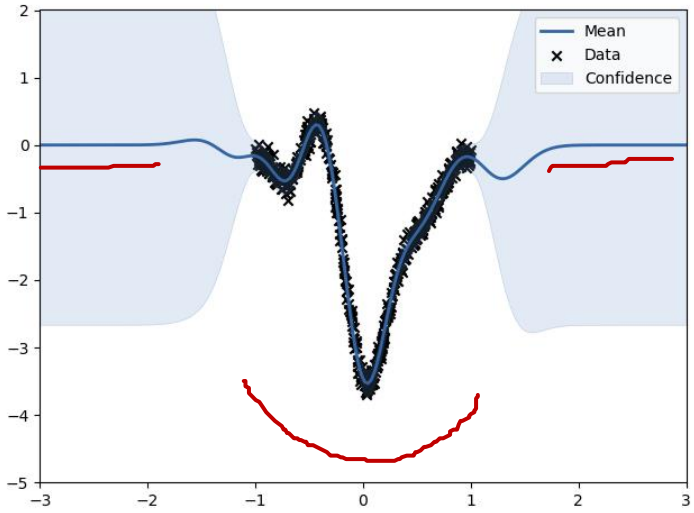
Gaussian Process Regression (GPR)

Prediction away from data

$$y = \sum_{i=1}^{M} \alpha_i k\left(x, x^i\right)$$

$$\text{with } \alpha = \left[K\left(X, X\right) + \sigma_\varepsilon{}^2 I\right]^{-1} \mathbf{y}$$

GPR with RBF kernel predicts y=0 away from datapoints!



Contrasts to SVR that predicts y=b away from datapoints.

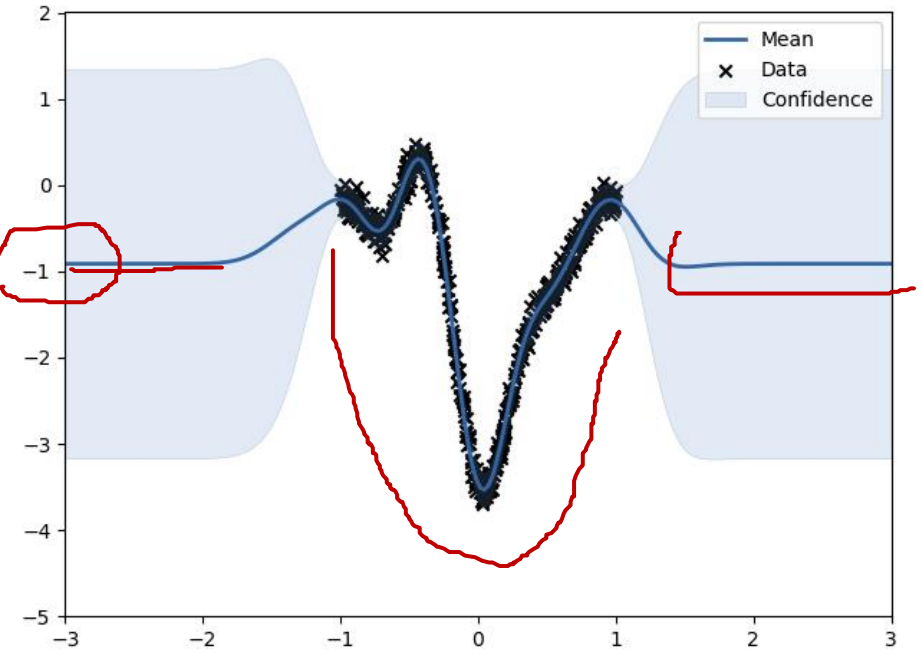
Gaussian Process Regression (GPR)

Prediction away from data

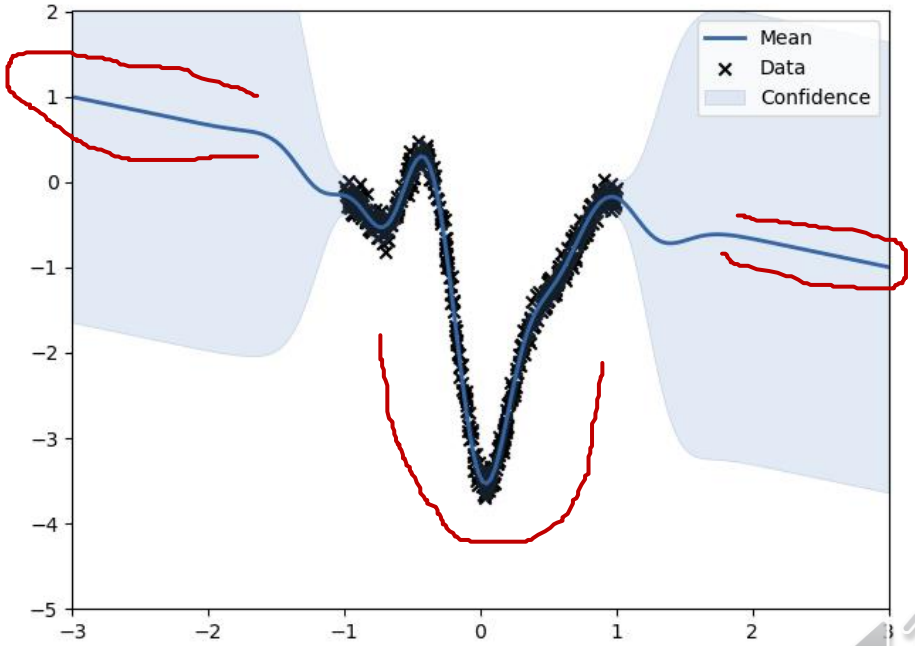
Instead of assuming zero mean one can add a mean function

$$\mathbb{E}\left\{p\left(y \mid x, X, \mathbf{y}\right)\right\} = m(x) + \sum_{i=1}^{M} \alpha_i k\left(\mathbf{x}^i, \mathbf{x}\right)$$

where: $\alpha = \left[\mathbf{K}\left(\mathbf{X}, \mathbf{X}\right) + \sigma_y^2 \mathbf{I}\right]^{-1} \left(\mathbf{y} - m(x)\right)$



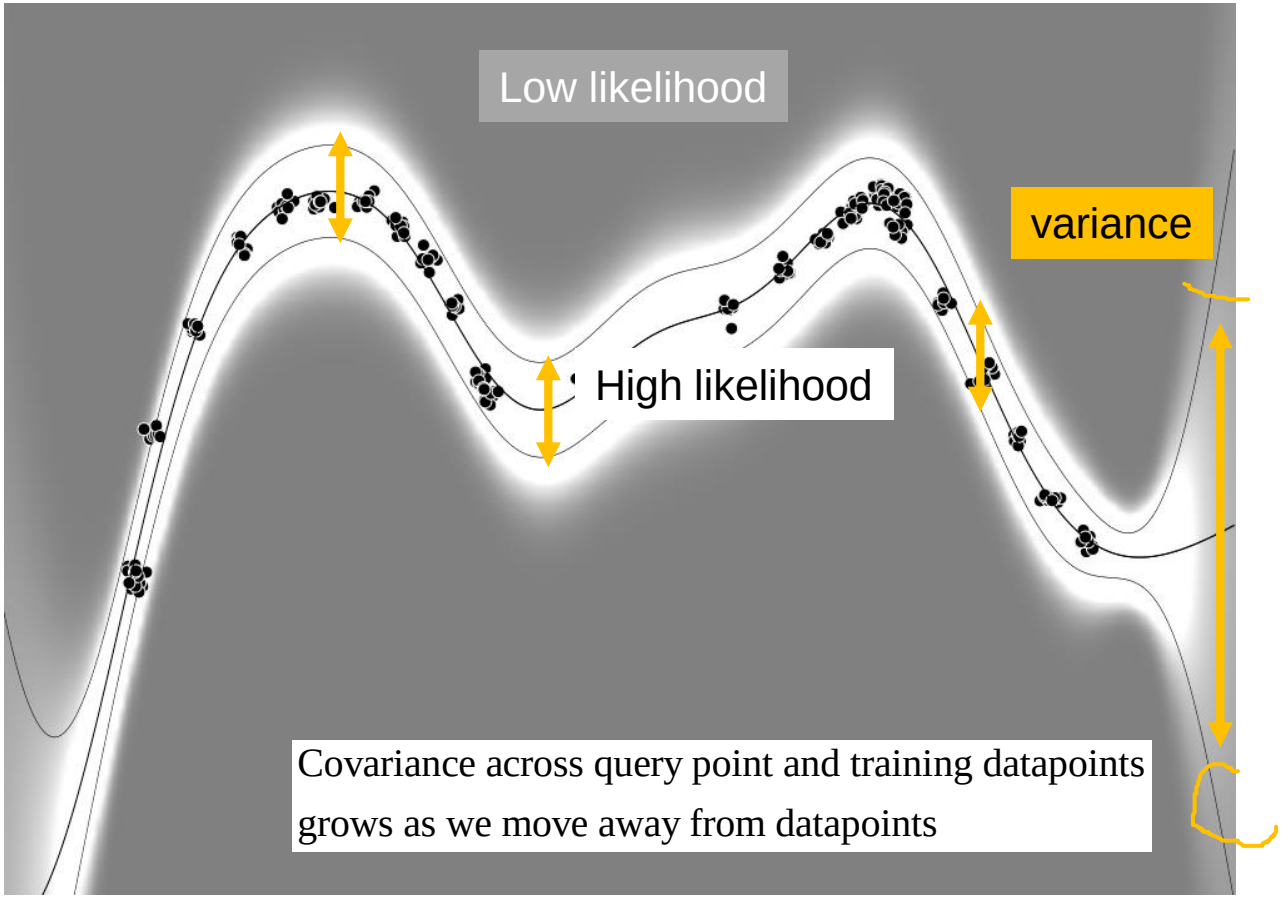
Constant mean



Linear mean

Gaussian Process Regression (GPR)
Confidence

The variance and the likelihood



Gaussian Process Regression: Summary

Advantages:

- ❑ Accuracy
- ❑ Estimation of predictions' uncertainty
- ❑ Auto-tuning of hyper-parameters

Disadvantage:

- ❑ Computational complexity $O(M^3)$
  (but there exist several sparse methods for GP)