

Lab 9b (solution)

TSM MachLe MSE
FS 2019

Gaussian Processes

Machine Learning
WÜRC

Essentially, all models are wrong, but some are useful.

GEORGE E.T. BOX

After this unit, ...

Lernziele/Kompetenzen

- you know the principle of maximum likelihood (ML) and maximum a posteriori probability (MAP) and you know their difference.
- you know (K1), that both the *conditionals* $p(x|y)$ and the *marginals* $p(x)$ of a joint Gaussian $p(x, y)$ are again Gaussian.
- you know (K1) that a *Gaussian process* $\mathcal{GP}(\mu, k)$ is a generalization of a multivariate Gaussian distribution to infinitely many variables. A Gaussian process is a *prior* over *functions* $p(f)$ which can be used for Bayesian regression. Sampling from a Gaussian process means sampling *functions* (instead of samples of a random variable) out of a pool of functions characterized by a mean function μ and a covariance function $k(x, x')$.
- you know (K1), that every model relies on (explicit or implicit) *assumptions*. We discriminate *knowledge*, *assumptions* and *simplifying assumptions*. In Bayesian reasoning, assumptions are formulated as *prior distribution* $p(\theta)$ over the parameters θ of a model. Using Bayes rule, one can calculate the posterior parameter distribution $p(\theta|x, y)$ given the data (x, y) and the model assumptions.

$$\text{posterior} = p(\theta|x, y) = \frac{p(y|x, \theta) \cdot p(\theta)}{\int_{\theta} p(y|x, \theta) \cdot p(\theta) d\theta} = \frac{\text{likelihood} \cdot \text{prior}}{\text{marginal}} \quad (1)$$

- you are able to formulate *probabilistic models* that use *priors* to express knowledge (or beliefs) about aspects of the model. You can formulate a *probabilistic model* for a process $f(x, \theta)$ with additive Gaussian noise ε . You can derive the *likelihood function* $p(y|x, \varepsilon, \theta)$ for this model given the parameters θ .
- you are able (K3) to *sample functions* from a Gaussian Process $\mathcal{GP}(\mu, k)$ with given mean $\mu(x)$ and covariance function $k(x, x')$ using the `GaussianProcessRegressor` of the class `sklearn.gaussian_process`.

- you are able (K3) to *fit* n -dimensional data using a Gaussian Process, i.e. you are able to *infer* hyperparameters of the model from given data using the `GaussianProcessRegressor` of the class `sklearn.gaussian_process`.
- you are able (K3) to *make predictions* using the `GaussianProcessRegressor` of the class `sklearn.gaussian_process`.
- you know (K1) that the *predictive distribution* which is used for making predictions for unknown data (x^*, y^*) can be calculated by *marginalizing* (integrating or averaging) over the parameter distribution.

$$p(y^*|x^*, x, y) = \int_{\theta} p(y^*|x^*, x, y, \theta) \cdot p(\theta|x, y) d\theta \quad (2)$$

- you know (K1) the most important covariance functions (kernels) $k(x, x')$, namely the *constant* kernel, the *Gaussian* kernel, the *RBF*-kernel (radial basis function), the *Dot-Product* kernel and the *sine-exponential* kernel.
- you are able (K3) to apply *kernel operations* (namely sum and product) in order to construct a probabilistic model adapted to a given dataset.

1. Medical Inference (Bayes Theorem) [M,I]

Breast cancer facts:

- 1 % of scanned women have breast cancer
- 80 % of women with breast cancer get positive mammography scans
- 9.6 % of women without breast cancer also get positive mammography scans

Question: A woman gets a scan, and it is positive. what is the probability that she has breast cancer?

Welche der folgenden Aussagen sind wahr und welche falsch?	wahr	falsch
a) less than 1 %	<input type="checkbox"/>	<input checked="" type="checkbox"/>
b) less than 10 %	<input checked="" type="checkbox"/>	<input type="checkbox"/>
c) around 80 %	<input type="checkbox"/>	<input checked="" type="checkbox"/>
d) around 90 %	<input type="checkbox"/>	<input checked="" type="checkbox"/>

2. Likelihood function, MAP and linear regression [L,II]

- a) The likelihood is the probability of each datapoint y_i given the model and its parameters.

$$p(\mathbf{Y}|\mathbf{X}, \theta) \quad (3)$$

The probability of *one* datapoint y_i given the model $\hat{y}(\theta, x_i)$ for a Gaussian noise ε is:

$$p(y_i|x_i, \theta) \sim \mathcal{N}(y_i|\hat{y}, \sigma_n^2) \quad (4)$$

$$= \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp \left\{ -\frac{1}{2\sigma_n^2} (y_i - \hat{y}(x_i, \theta))^2 \right\} \quad (5)$$

$$= \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp \left\{ -\frac{1}{2\sigma_n^2} (y_i - f(x_i|\theta))^2 \right\} \quad (6)$$

The likelihood of all data points is the product of the probabilities of each datapoint (x_i, y_i) :

$$p(\mathbf{Y}|\mathbf{X}, \theta) = \prod_{i=1}^N p(y_i|x_i, \theta) \quad (7)$$

$$= \frac{1}{(2\pi\sigma_n^2)^{N/2}} \cdot \exp \left\{ -\frac{1}{2\sigma_n^2} \sum_{i=1}^N (y_i - f(x_i|\theta))^2 \right\} \quad (8)$$

$$= \frac{1}{(2\pi\sigma_n^2)^{N/2}} \cdot \exp \left\{ -\frac{1}{2\sigma_n^2} \|\mathbf{Y} - f(\mathbf{X}|\theta)\|^2 \right\} \quad (9)$$

- b) By taking the natural logarithm of (7), the result immediately follows:

$$\log [p(\mathbf{Y}|\mathbf{X}, \theta)] = -\frac{1}{2\sigma_n^2} \sum_{i=1}^N (y_i - f(x_i|\theta))^2 - \frac{N}{2} \cdot \log (2\pi\sigma_n^2) \quad (10)$$

$$= -\frac{1}{2\sigma_n^2} \|\mathbf{Y} - f(\mathbf{X}|\theta)\|^2 - \frac{N}{2} \cdot \log (2\pi\sigma_n^2) \quad (11)$$

$$= -\frac{1}{2\sigma_n^2} \text{SSE}(\theta) - \frac{N}{2} \cdot \log (2\pi\sigma_n^2) \quad (12)$$

- c) In case of a linear model, we can write the sum of the squared error $\text{SSE}(\theta)$ in matrix form:

$$\text{SSE}(\theta) = \|\mathbf{Y} - f(\mathbf{X}|\theta)\|^2 \quad (13)$$

$$= (\mathbf{Y} - f(\mathbf{X}|\theta))^T \cdot (\mathbf{Y} - f(\mathbf{X}|\theta)) \quad (14)$$

$$= (\mathbf{Y} - \Phi \cdot \theta)^T \cdot (\mathbf{Y} - \Phi \cdot \theta) \quad (15)$$

The square norm can always be written in form of a scalar product, so it is sufficient to consider only one term of the scalar product:

$$(\mathbf{Y} - f(\mathbf{X}|\theta)) = \begin{pmatrix} y_1 - (\theta_1 + \theta_2 x_1) \\ y_2 - (\theta_1 + \theta_2 x_2) \\ \vdots \\ y_N - (\theta_1 + \theta_2 x_N) \end{pmatrix} \quad (16)$$

$$= \left(\mathbf{Y} - [\mathbf{1}_N \mathbf{X}] \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \right) \quad (17)$$

$$= (\mathbf{Y} - \Phi \cdot \theta) \quad (18)$$

d) The two following results from matrix calculus are useful. For column vectors $\boldsymbol{\theta}$ and \mathbf{x} of the same length, the following statement is valid:

$$\nabla_{\boldsymbol{\theta}}(\mathbf{a}^T \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}(\boldsymbol{\theta}^T \mathbf{a}) = \mathbf{a} \quad (19)$$

For a column vector $\boldsymbol{\theta}$ and matrix \mathbf{A} , the following identity holds:

$$\nabla_{\boldsymbol{\theta}}(\boldsymbol{\theta}^T \mathbf{A} \boldsymbol{\theta}) = (\mathbf{A} + \mathbf{A}^T) \boldsymbol{\theta} \quad (20)$$

Especially, if \mathbf{A} is symmetric:

$$\nabla_{\boldsymbol{\theta}}(\boldsymbol{\theta}^T \mathbf{A} \boldsymbol{\theta}) = 2\mathbf{A} \boldsymbol{\theta} \quad (21)$$

To find the maximum likelihood solution, we set the gradient of the log likelihood function to zero:

$$0 = \nabla_{\boldsymbol{\theta}} \log [p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})] \quad (22)$$

$$= -\frac{1}{2\sigma_n^2} \nabla_{\boldsymbol{\theta}} \|\mathbf{Y} - f(\mathbf{X}|\boldsymbol{\theta})\|^2 \quad (23)$$

$$= -\frac{1}{2\sigma_n^2} \nabla_{\boldsymbol{\theta}} [(\Phi \boldsymbol{\theta} - \mathbf{Y})^T \cdot (\Phi \boldsymbol{\theta} - \mathbf{Y})] \quad (24)$$

$$= -\frac{1}{2\sigma_n^2} \nabla_{\boldsymbol{\theta}} [\boldsymbol{\theta}^T \Phi^T \Phi \boldsymbol{\theta} - \boldsymbol{\theta}^T \Phi^T \mathbf{Y} - \mathbf{Y}^T \Phi \boldsymbol{\theta} - \mathbf{Y}^T \mathbf{Y}] \quad (25)$$

$$= -\frac{1}{2\sigma_n^2} [\nabla_{\boldsymbol{\theta}} \boldsymbol{\theta}^T \Phi^T \Phi \boldsymbol{\theta} - 2\nabla_{\boldsymbol{\theta}} \boldsymbol{\theta}^T \Phi^T \mathbf{Y} - 0] \quad (26)$$

$$= -\frac{1}{2\sigma_n^2} [2\Phi^T \Phi \boldsymbol{\theta} - 2\Phi^T \mathbf{Y}] \quad (27)$$

This leads to the definition of the *Least Squares Normal Equations*:

$$\Phi \mathbf{Y} = \Phi^T \Phi \boldsymbol{\theta} \quad (28)$$

The Least Squares estimate $\hat{\boldsymbol{\theta}}_{\text{ML}}$ of the parameters $\boldsymbol{\theta}$ is then given by:

$$\hat{\boldsymbol{\theta}}_{\text{ML}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{Y} \quad (29)$$

3. Prior samples and posterior distributions from different kernels of a \mathcal{GP} [A,II]

The solution Jupyter notebook can be found on moodle:

Lab9b_A3_plot_gpr_prior_posterior.ipynb

4. Model fitting, prediction and noise estimation using a \mathcal{GP} [A,II]

The solution Jupyter notebook can be found on moodle:

Lab9b_A4_FitGPModel_NoiseEstimation_solution.ipynb