OST
Ostschweizer
Fachhochschule

# Lab 9 (solution)

FTP MachLe MSE
HS 2023

## 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 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 $\mu$ 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 $\theta$ 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.

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

- your are able to formulate *probabilitstic models* that use *priors* to express knowlege (or beliefs) about aspects of the model. You can formulate a *probabilistic model* for a process $f(x, \theta)$ with additive Gaussian noise $\varepsilon$. You can derive the *likelihood function* $p(y|x, \varepsilon, \theta)$ for this model given the parameters $\theta$.
- 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.

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

- 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 % | ○ | ● |
| **b)** less than 10 % | ● | ○ |
| **c)** around 80 % | ○ | ● |
| **d)** around 90 % | ○ | ● |

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

**a)** The emphlikelihood is the probability of each datapoint $y_i$ given the model and its parameters.

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

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

$$\begin{aligned} p(y_i|x_i, \theta) &\sim \mathcal{N}\left(y_i|\hat{y}, \sigma_n^2\right) & (4)\\ &= \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left\{-\frac{1}{2\sigma_n^2}\left(y_i - \hat{y}(x_i, \theta)\right)^2\right\} & (5)\\ &= \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left\{-\frac{1}{2\sigma_n^2}\left(y_i - f(x_i|\theta)\right)^2\right\} & (6)\end{aligned}$$

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

$$\begin{aligned} p(\mathbf{Y}|\mathbf{X}, \theta) &= \prod_{i=1}^{N} p(y_i|x_i, \theta) & (7)\\ &= \frac{1}{(2\pi\sigma_n^2)^{N/2}} \cdot \exp\left\{-\frac{1}{2\sigma_n^2}\sum_{i=1}^{N}\left(y_i - f(x_i|\theta)\right)^2\right\} & (8)\\ &= \frac{1}{(2\pi\sigma_n^2)^{N/2}} \cdot \exp\left\{-\frac{1}{2\sigma_n^2}\|\mathbf{Y} - f(\mathbf{X}|\theta)\|^2\right\} & (9)\end{aligned}$$

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

$$\begin{aligned} \log\left[p\left(\mathbf{Y}|\mathbf{X}, \theta\right)\right] &= -\frac{1}{2\sigma_n^2}\sum_{i=1}^{N}\left(y_i - f(x_i|\theta)\right)^2 - \frac{N}{2}\cdot\log\left(2\pi\sigma_n^2\right) & (10)\\ &= -\frac{1}{2\sigma_n^2}\|\mathbf{Y} - f(\mathbf{X}|\theta)\|^2 - \frac{N}{2}\cdot\log\left(2\pi\sigma_n^2\right) & (11)\\ &= -\frac{1}{2\sigma_n^2}\mathrm{SSE}(\theta) - \frac{N}{2}\cdot\log\left(2\pi\sigma_n^2\right) & (12)\end{aligned}$$

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

$$\begin{aligned} \mathrm{SSE}(\theta) &= \|\mathbf{Y} - f(\mathbf{X}|\theta)\|^2 & (13)\\ &= \left(\mathbf{Y} - f(\mathbf{X}|\theta)\right)^T \cdot \left(\mathbf{Y} - f(\mathbf{X}|\theta)\right) & (14)\\ &= \left(\mathbf{Y} - \Phi\cdot\theta\right)^T \cdot \left(\mathbf{Y} - \Phi\cdot\theta\right) & (15)\end{aligned}$$

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:

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

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

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

**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:

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

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

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

Especially, if $\mathbf{A}$ is symmetric:

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

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

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

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

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

$$= -\frac{1}{2\sigma_n^2} \nabla_{\boldsymbol{\theta}} \left[ \boldsymbol{\theta^T \Phi^T \Phi \theta - \theta^T \Phi^T} \mathbf{Y} - \mathbf{Y^T} \boldsymbol{\Phi\theta} - \mathbf{Y^T Y} \right] \tag{25}$$

$$= -\frac{1}{2\sigma_n^2} \left[ \nabla_{\boldsymbol{\theta}} \boldsymbol{\theta^T \Phi^T \Phi \theta} - 2\nabla_{\boldsymbol{\theta}} \boldsymbol{\theta^T \Phi^T} \mathbf{Y} - 0 \right] \tag{26}$$

$$= -\frac{1}{2\sigma_n^2} \left[ 2\boldsymbol{\Phi^T \Phi \theta} - 2\boldsymbol{\Phi}\mathbf{Y} \right] \tag{27}$$

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

$$\boldsymbol{\Phi}\mathbf{Y} = \boldsymbol{\Phi^T \Phi \theta} \tag{28}$$

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

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

**3. Prior samples and posterior distributions from differnt kernels of a $\mathcal{GP}$ [A, II]**
The solution Juypter notebook can be found on moodle:
`Lab9_A3_plot_gpr_prior_posterior.ipynb`

**4. Model fitting, prediction and noise estimation using a $\mathcal{GP}$ [A, II]**
The solution Juypter notebook can be found on moodle:
`Lab9_A4_FitGPModel_NoiseEstimation_solution.ipynb`