Lehrstuhl für STEUERUNGS- UND REGELUNGSTECHNIK Technische Universität München Prof. Dongheui Lee	MACHINE LEARNING IN ROBOTICS Exercises 11: Gaussian Processes

Exercise 1

Linear regression can be transformed into a Gaussian process by assuming a Gaussian prior over the weights $p(\boldsymbol{w}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Assuming that $f(\boldsymbol{x})$ is linear, i.e. $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{w}$, derive the posterior distribution over $\boldsymbol{w}$.

Solution Exercise 1 The linear regression model can be written as

$$y^{(i)} = \boldsymbol{w}^T \boldsymbol{x}^{(i)} + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ is a Gaussian additive noise. It follows that $y^{(i)} = \boldsymbol{w}^T \boldsymbol{x}^{(i)} + \epsilon \sim \mathcal{N}(\boldsymbol{w}^T \boldsymbol{x}^{(i)}, \sigma_n^2)$, or equivalently

$$p(y^{(i)} | \boldsymbol{x}^{(i)}, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}^T \boldsymbol{x}^{(i)}, \sigma_n^2)$$

For simplicity, let us stack the input data into the matrix $\boldsymbol{X}$ and the output data into the vector $\boldsymbol{y}$. Assume a zero mean Gaussian prior over the weights $p(\boldsymbol{w}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. From the Bayes' rule we can write the posterior distribution over $\boldsymbol{w}$ as

$$p(\boldsymbol{w} | \boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{w}) p(\boldsymbol{w})}{p(\boldsymbol{y} | \boldsymbol{X})}$$

Considering that the marginal likelihood $p(\boldsymbol{y}|\boldsymbol{X})$ is constant we have

$$\begin{aligned}
p(\boldsymbol{w} | \boldsymbol{y}, \boldsymbol{X}) &\propto p(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{w}) p(\boldsymbol{w}) \\
&= \exp\left(-\frac{1}{2\sigma_n^2} (\boldsymbol{y} - \boldsymbol{X}^T \boldsymbol{w})^T (\boldsymbol{y} - \boldsymbol{X}^T \boldsymbol{w})\right) \exp\left(-\frac{1}{2} \boldsymbol{w}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{w}\right) \\
&= \exp\left(-\frac{1}{2\sigma_n^2} (\boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{w}^T \boldsymbol{X} \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{X}^T \boldsymbol{w} + \boldsymbol{w}^T \boldsymbol{X} \boldsymbol{X}^T \boldsymbol{w}) - \frac{1}{2} \boldsymbol{w}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{w}\right) \\
&= \exp\left(-\frac{1}{2\sigma_n^2} (\boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{w}^T \boldsymbol{X} \boldsymbol{y}) - \frac{1}{2} \boldsymbol{w}^T \left(\frac{\boldsymbol{X}\boldsymbol{X}^T}{\sigma_n^2} + \boldsymbol{\Sigma}^{-1}\right) \boldsymbol{w}\right)
\end{aligned}$$

Choosing $\boldsymbol{A} = \left(\frac{\boldsymbol{X}\boldsymbol{X}^T}{\sigma_n^2} + \boldsymbol{\Sigma}^{-1}\right)$ and considering that $\frac{1}{2\sigma_n^2} \boldsymbol{y}^T \boldsymbol{y} = \frac{1}{2\sigma_n^2} ||\boldsymbol{y}||^2 = c$ is constant we have

$$\begin{aligned}
p(\boldsymbol{w} | \boldsymbol{y}, \boldsymbol{X}) &\propto \exp\left(-c + \frac{1}{\sigma_n^2} \boldsymbol{w}^T \boldsymbol{X} \boldsymbol{y} - \frac{1}{2} \boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}\right) \\
&= \exp(-c) \exp\left(\frac{1}{\sigma_n^2} \boldsymbol{w}^T \boldsymbol{X} \boldsymbol{y} - \frac{1}{2} \boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}\right) \\
&= \exp(-c) \exp\left(\frac{1}{\sigma_n^2} \boldsymbol{w}^T \boldsymbol{A} \boldsymbol{A}^{-1} \boldsymbol{X} \boldsymbol{y} - \frac{1}{2} \boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}\right) \\
&= \exp(-c) \exp\left(-\frac{1}{2} (\boldsymbol{w} - \frac{1}{\sigma_n^2} \boldsymbol{A}^{-1} \boldsymbol{X} \boldsymbol{y})^T \boldsymbol{A} (\boldsymbol{w} - \frac{1}{\sigma_n^2} \boldsymbol{A}^{-1} \boldsymbol{X} \boldsymbol{y}) - \frac{1}{2(\sigma_n^2)^2} (\boldsymbol{X} \boldsymbol{y})^T \boldsymbol{A}^{-1} \boldsymbol{X} \boldsymbol{y}\right) \\
&= \exp(-c - \frac{1}{2(\sigma_n^2)^2} (\boldsymbol{X} \boldsymbol{y})^T \boldsymbol{A}^{-1} \boldsymbol{X} \boldsymbol{y}) \exp\left(-\frac{1}{2} (\boldsymbol{w} - \frac{1}{\sigma_n^2} \boldsymbol{A}^{-1} \boldsymbol{X} \boldsymbol{y})^T \boldsymbol{A} (\boldsymbol{w} - \frac{1}{\sigma_n^2} \boldsymbol{A}^{-1} \boldsymbol{X} \boldsymbol{y})\right) \\
&\propto \exp\left(-\frac{1}{2} (\boldsymbol{w} - \frac{1}{\sigma_n^2} \boldsymbol{A}^{-1} \boldsymbol{X} \boldsymbol{y})^T \boldsymbol{A} (\boldsymbol{w} - \frac{1}{\sigma_n^2} \boldsymbol{A}^{-1} \boldsymbol{X} \boldsymbol{y})\right) = \mathcal{N}\left(\frac{1}{\sigma_n^2} \boldsymbol{A}^{-1} \boldsymbol{X} \boldsymbol{y}, \boldsymbol{A}^{-1}\right)
\end{aligned}$$

Exercise 2

An autonomous flying robot, while performing a transportation task, encounters a storm. The odometry module of the robot computes and transmits position values (y) every $0.2s$. Due to the electromagnetic interference of the storm, odometry data are influenced by a Gaussian noise.

Being unable to complete the task, the robot activates an emergency landing procedure. To work properly, the landing procedure requires the prediction of the robot position one step ($0.2s$) in the future. The limited amount of memory of the robots allows to store only the last 3 positions.

To make predictions, the landing strategy uses Gaussian processes, with the radial basis kernel function:

$$k(x, x') = exp\left(-\frac{\|x - x'\|^2}{2l^2}\right) + \sigma_n^2 \delta_{xx'}$$

where x are the position sampling time instants, l defines the lengthscale of the kernel, σ_n^2 the variance of the noise component, and $\delta_{xx'} = 1$ if $x = x'$, 0 otherwise. Given the last 3 positions $\boldsymbol{y} = [1.5,\ 1,\ 0.8]$ with associated times $\boldsymbol{x} = [1,\ 1.2,\ 1.4]$, compute the position at $x^{(3)} = 1.4s$ and $x_t = 1.6s$ assuming:

a) Variance $\sigma_n^2 = 0.1$ and lengthscale $l = 0.1$.

b) Variance $\sigma_n^2 = 0.1$ and lengthscale $l = 0.001$ (close to zero).

c) Variance $\sigma_n^2 = 0.1$ and lengthscale $l = 100$ (very high).

d) State your conclusions regarding the relationship between lenghtscale and obtained results.

Solution Exercise 2

a) Given a query point x_t, the predictive mean $\hat{y}_t$ is:

$$\hat{y}_t = \boldsymbol{K}(x_t, \boldsymbol{x})[\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}) + \sigma_n^2 \boldsymbol{I}]^{-1} \boldsymbol{y}^T$$

For $\sigma_n^2 = 0.1$ and $l = 0.1$, we have:

$$\hat{y}(x^{(3)} = 1.4) = \begin{bmatrix} 0.45 & 0.82 & 1 \end{bmatrix} \begin{bmatrix} 1.1 & 0.82 & 0.45 \\ 0.82 & 1.1 & 0.82 \\ 0.45 & 0.82 & 1.1 \end{bmatrix}^{-1} \begin{bmatrix} 1.5 \\ 1 \\ 0.8 \end{bmatrix} \approx 0.73$$

$$\hat{y}(x_t = 1.6) \approx 0.44$$

b) For $\sigma_n^2 = 0.1$ and $l = 0.001$, we have:

$$\hat{y}(x^{(3)} = 1.4) \approx 0.73$$

$$\hat{y}(x_t = 1.6) \approx 0.0$$

Good approximation of the training points, poor predictions with new targets $\longrightarrow$ over-fitting.

c) For $\sigma_n^2 = 0.1$ and $l = 100$, we have:

$$\hat{y}(x^{(3)} = 1.4) \approx 1.06$$

$$\hat{y}(x_t = 1.6) \approx 1.06$$

The predicted value is almost constant.

d) The variation of the lengthscale parameter is key to obtain optimal data fitting. For very low values of l the model has high complexity and is prone to over-fitting, while for high values the prediction is almost independent from the data.