Technische Universität Berlin
Fakultät IV – Elektrotechnik und Informatik

Probabilistic and Bayesian Modelling in Machine Learning and Artificial Intelligence

Manfred Opper and Théo Galy–Fajou

Summer Term 2018

Problem Sheet 4

Solutions

Problem 1 – Evidence for Gaussian process (GP) regression

For the GP regression problem, we assume that data are generated as

$$y_i = f(x_i) + \nu_i \qquad i = 1, \dots, n \tag{1}$$

where the ν_i are independent, zero mean Gaussian noise variables within $E[\nu_i^2] = \sigma^2$ and $f(\cdot)$ has a GP prior with kernel $K(x, x')$. Show that the **Bayesian evidence** is given by

$$p(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\det(\mathbf{K} + \sigma^2 \mathbf{I})|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\mathbf{y}^T(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}\right] \tag{2}$$

where $\mathbf{y} = (y_1, \dots, y_n)$ and the kernel matrix is defined by $\mathbf{K}_{ij} = K(x_i, x_j)$.

Hint: Calculate the joint density of $\mathbf{y}$ and use the fact that $f(x_j)$ and ν_i are independent Gaussian random variables. Hence you can add the respective covariance matrices.

Solution

The evidence can be computed via the joint distribution :

$$p(\mathbf{y}) = \int p(\mathbf{y}, \mathbf{f}) d\mathbf{f} = \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}) d\mathbf{f}$$

Where

$$p(\mathbf{y}|\mathbf{f}) = \frac{1}{(2\pi)^{N/2}|\det(\sigma^2\mathbf{I})|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{f})^\top \sigma^{-2}\mathbf{I}(\mathbf{y} - \mathbf{f})\right]$$

$$p(\mathbf{f}) = \frac{1}{(2\pi)^{N/2}|\det(\mathbf{K})|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\mathbf{f}^\top \mathbf{K}^{-1}\mathbf{f}\right]$$

In the integration one can do it the hard way and reformulate

$$p(\mathbf{y} \mid \mathbf{f})p(\mathbf{f}) \equiv \mathcal{N}(y \mid 0, \sigma^2\mathbf{I} + \mathbf{K})\mathcal{N}(f \mid \overline{\mu}, \overline{\Sigma})$$

using the identity

$$\begin{aligned}\mathcal{N}(x \mid m_1, \Sigma_1) \cdot \mathcal{N}(x \mid m_2, \Sigma_2) =& \mathcal{N}(m_1 \mid m_2, (\Sigma_1 + \Sigma_2)\mathcal{N}(x \mid \overline{m}, \overline{\Sigma})\\ \overline{m} =& (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}(\Sigma_1^{-1}m_1 + \Sigma_2^{-1}m_2)\\ \overline{\Sigma} =& (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}\end{aligned}$$

Replacing x by $\mathbf{f}$, the integral over $\mathbf{f}$ gives one and we recover the multivariate gaussian for $\mathbf{y}$. Or one can intuitively see that we are having a zero mean prior on the mean ($\mathbf{f}$), and therefore one can simply add the variances to get the final result.

Problem 2 – Gibbs sampler for outlier detection

The file `outlier.dat` on the web page of the course contains a data set $D = (y_1, \ldots, y_N)$. Most of the observations have been drawn from a Gaussian probability distribution $\mathcal{N}(y_i; \mu, \sigma^2)$ with mean μ and variance σ^2. However, D contains some *outliers*, which occur with probability ϵ and are displaced by a random offset A_i. For the purpose of *outlier detection* the model is augmented with an indicator variable

$$\delta_i = \begin{cases} 1 & \text{if } y_i \text{ is an outlier,} \\ 0 & \text{if } y_i \text{ is a normal data point,} \end{cases}$$

for each observation. Assuming conjugate priors for the parameters yields the full stochastic model

$$\begin{array}{lllllllll} \mu & \sim & \mathcal{N}(\theta, v^2), & \sigma^{-2} & \sim & \text{Gamma}(\kappa, \lambda), & \epsilon & \sim & \text{Beta}(\alpha, \beta), \\ y_i & \sim & \mathcal{N}(\mu + \delta_i\, A_i, \sigma^2), & \delta_i & \sim & \text{Bernoulli}(\epsilon), & A_i & \sim & \mathcal{N}(0, \tau^2). \end{array}$$

We want to use a Gibbs sampler in order to draw samples from the posterior $p(\mu, \sigma^2, \epsilon, \boldsymbol{\delta}, \mathbf{A}|D)$ with $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)$ and $\mathbf{A} = (A_1, \ldots, A_N)$. Some conditional posteriors are given by

$$\begin{aligned} \mu &\sim \mathcal{N}\left(\frac{\sigma^2\theta + v^2\sum_{i=1}^N (y_i - \delta_i\, A_i)}{\sigma^2 + Nv^2}, \frac{\sigma^2 v^2}{\sigma^2 + Nv^2}\right), \\ \sigma^{-2} &\sim \text{Gamma}\left(\kappa + \frac{N}{2}, \frac{2\lambda}{2 + \lambda\sum_{i=1}^N (y_i - \delta_i\, A_i - \mu)^2}\right). \end{aligned}$$

(a) Show that the remaining conditional posteriors are given by

$$
\begin{aligned}
\delta_i &\sim \text{Bernoulli}\left(\frac{\epsilon}{\epsilon + (1-\epsilon)\exp(-A_i(y_i - A_i - \mu)/(2\sigma^2))}\right), \\
A_i &\sim \mathcal{N}\left(\frac{\tau^2\delta_i(y_i-\mu)}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2\delta_i}\right), \\
\epsilon &\sim \text{Beta}\left(\alpha + \sum_{i=1}^N \delta_i, \beta + \sum_{i=1}^N (1-\delta_i)\right).
\end{aligned}
$$

- Joint probability distribution

$$
\begin{aligned}
p(\mu, \sigma^2, \epsilon, \boldsymbol{\delta}, \mathbf{A}, D) &= \frac{\Gamma(\alpha+\beta)\epsilon^{\alpha-1}(1-\epsilon)^{\beta-1}}{\sqrt{2\pi v^2}\Gamma(\kappa)\lambda^\kappa\Gamma(\alpha)\Gamma(\beta)}\sigma^{-2(\kappa-1)}e^{-\frac{\sigma^{-2}}{\lambda}-\frac{(\mu-\theta)^2}{2v^2}} \\
&\times \prod_{i=1}^N \frac{\epsilon^{\delta_i}(1-\epsilon)^{1-\delta_i}}{2\pi\sigma\tau} e^{-\frac{(y_i-\delta_i A_i-\mu)^2}{2\sigma^2}-\frac{A_i^2}{2\tau^2}}
\end{aligned}
$$

- Conditional distributions

$$
\begin{aligned}
p(\mu|\dots) \propto& e^{-\frac{(\mu-\theta)^2}{2v^2}-\sum_{i=1}^N\frac{(y_i-\delta_iA_i-\mu)^2}{2\sigma^2}} \\
\Rightarrow p(\mu|\dots) =& \mathcal{N}\left(\mu|\left(\frac{1}{\nu^2}+\frac{N}{\sigma^2}\right)^{-1}\left(\frac{\theta^2}{\nu^2}+\frac{1}{\sigma^2}\sum_i^N y_i - \delta_i A_i\right), \left(\frac{1}{\nu^2}+\frac{N}{\sigma^2}\right)^{-1}\right) \\
p(\sigma^{-2}|\dots) \propto& \sigma^{-2(\kappa+N/2-1)}e^{-\sigma^{-2}\left(\frac{1}{\lambda}+\sum_{i=1}^N\frac{(y_i-\delta_iA_i-\mu)^2}{2}\right)} \\
\Rightarrow p(\sigma^{-2}|\dots) =& \text{Gamma}\left(\sigma^{-2} \mid \kappa+\frac{N}{2}, \frac{2\lambda}{2+\lambda\sum_{i=1}^N(y_i-\delta_i\,A_i-\mu)^2}\right). \\
p(\delta_i|\dots) \propto& \epsilon^{\delta_i}(1-\epsilon)^{1-\delta_i}e^{-\delta_i\frac{A_i(A_i+\mu-y_i)}{2\sigma^2}} \\
\Rightarrow p(\delta_i \mid \dots) =& \text{Bernoulli}\left(\frac{\epsilon}{\epsilon+(1-\epsilon)\exp(-A_i(y_i-A_i-\mu)/(2\sigma^2))}\right)
\end{aligned}
$$

To get the new parameter of the Bernoulli distribution, compute the normalization constant by summing over $\delta_i = \{0, 1\}$:

$$
\begin{aligned}
p(\delta_i = 0) \propto& (1-\epsilon), \quad p(\delta_i = 1) \propto \epsilon e^{-\frac{A_i(A_i+\mu-y_i)}{2\sigma^2}} \\
\Rightarrow p(\delta_i = 1) =& \frac{\epsilon e^{-\frac{A_i(A_i+\mu-y_i)}{2\sigma^2}}}{(1-\epsilon) + \epsilon e^{-\frac{A_i(A_i+\mu-y_i)}{2\sigma^2}}} = \frac{\epsilon}{(1-\epsilon) e^{\frac{A_i(A_i+\mu-y_i)}{2\sigma^2}} + \epsilon} \\
p(A_i|\dots) \propto& e^{-\frac{(y_i-\delta_i A_i-\mu)^2}{2\sigma^2} - \frac{A_i^2}{2\tau^2}} \\
\Rightarrow p(A_i|\dots) =& \mathcal{N}\left(A_i \mid \frac{\tau^2\delta_i(y_i-\mu)}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2\delta_i}\right) \\
p(\epsilon|\dots) \propto& \epsilon^{\alpha-1+\sum_{i=1}^N \delta_i}(1-\epsilon)^{\beta-1+\sum_{i=1}^N(1-\delta_i)} \\
\Rightarrow p(\epsilon|\dots) =& \text{Beta}\left(\epsilon \mid \alpha + \sum_{i=1}^N \delta_i, \beta + \sum_{i=1}^N (1-\delta_i)\right).
\end{aligned}
$$

(b) Write a program that implements the *Gibbs sampler*. Generate 10^3 samples from the posterior using the hyperparameters $\theta = 0$, $v^2 = 100$, $\kappa = 2$, $\lambda = 2$, $\alpha = 2$, $\beta = 20$, $\tau^2 = 100$. Plot histograms showing the marginal posteriors $p(\mu|D)$ and $p(\epsilon|D)$.

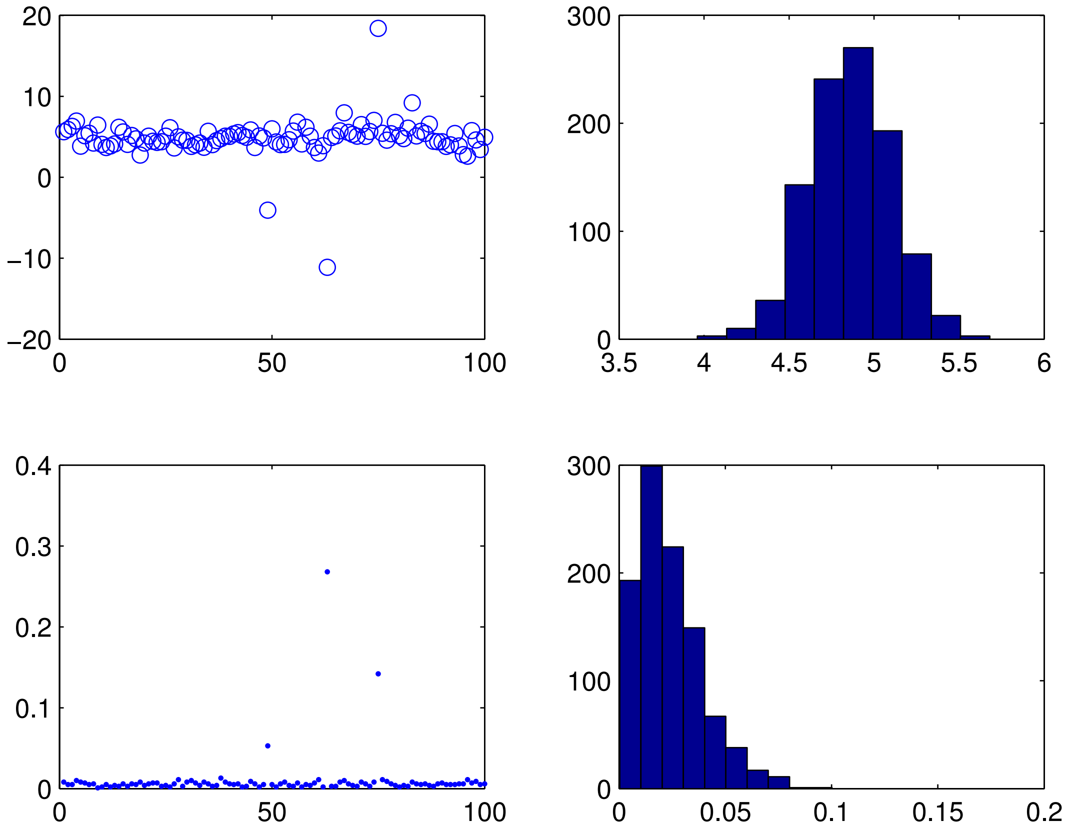
Figure 1: Results of outlier detection: data set (top left), histogram for $p(\mu|D)$ (top right), probability $p(\delta_i|D)$ that data point i is an outlier (bottom left), and histogram for $p(\epsilon|D)$ (bottom right)

(c) Which data points in the file `outlier.dat` are outliers? Use the samples generated in part (b) and the condition $p(\delta_i|D) \geq 0.02$ in order to identify them.

index	value	$p(\delta_i\|D)$
49	-4.0758	0.0530
63	-11.1217	0.2680
75	18.3938	0.1420

Table 1: Outliers in `outlier.dat` as identified by the Gibbs sampler.

Solution See Matlab code on ISIS



20
10
0
−10
−20
0
50
100
300
200
100
0
3.5
4
4.5
5
5.5
6
0.4
0.3
0.2
0.1
0
0
50
100
300
200
100
0
0
0.05
0.1
0.15
0.2