Bayesian Machine Learning
Course 67564
Solution To Exercise 3: Evidence and Kernels

Barak Haim 0

27/12/2022

Contents

1 Theoretical

1.1 Input-Specific Noise

$$y(x_i) = \theta h(x_i) + \eta_i \qquad \eta_i \sim \mathcal{N}\left(0, \sigma_i^2\right)$$

1.1.1 Q1 $p(y|\theta) \sim \mathcal{N}\left(H\theta, \Gamma\right)$

For each i:

$$p(y_i|\theta) \sim \mathcal{N}\left(yi \,|\, \theta^T h(x_i), I\sigma_i^2\right)$$

Explicitly:

$$p(y_i|\theta) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2\sigma_i^2}\left(h(x_i)^T\theta - y_i\right)^2\right]$$

Denote $\eta = \mathcal{N}\left(0, \Gamma\right)$ **where:**

$$\Gamma = daig(\left\{\sigma_i^2\right\}_{i=1}^n) = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{bmatrix}$$

As y is an affine transformation of θ **and** η **we get:**

$$p(y|\theta) \sim \mathcal{N}\left(H\theta, \Gamma\right)$$

Explicitly:

$$p(y|\theta) = \frac{1}{\sqrt{(2\pi)^d |\Gamma|}} \exp\left[-\frac{1}{2}\left(H\theta - y\right)^T \Gamma^{-1} \left(H\theta - y\right)\right]$$

1.1.2 Q2 $\hat{\theta}_{MLE} = \left(H^T\Gamma^{-1}H\right)^{-1} H^T\Gamma^{-1}y$

We write the log-likelihood:

$$\ell(y|\theta) = \ln \mathcal{N}\left(H\theta, \Gamma\right) = -\frac{1}{2}\left(H\theta - y\right)^T \Gamma^{-1}\left(H\theta - y\right) + const$$

So in fact we try to minimize:

$$L = \frac{1}{2}\left(H\theta - y\right)^T \Gamma^{-1}\left(H\theta - y\right)$$

And so:

$$\begin{aligned} &\frac{\partial L}{\partial \theta} = H^T\Gamma^{-1}(H\theta - y) \overset{!}{=} 0 \\ &iff H^T\Gamma^{-1}y = H^T\Gamma^{-1}H\theta \\ &iff \hat{\theta}_{MLE} = \left(H^T\Gamma^{-1}H\right)^{-1} H^T\Gamma^{-1}y \end{aligned}$$

1.1.3 Q3 $\theta \sim \mathcal{N}\left(\mu_0, \Sigma_0\right) \rightarrow p(\theta|y) = ?$

With Bayes' law we know: $p(\theta|y) \propto p(\theta)p(y|\theta)$ and so $p(\theta|y) \sim \mathcal{N}\left(\mu_{\theta|y}, C_{\theta|y}\right)$ as the product of 2 Gaussian. Also

$$\begin{aligned}
\mu_{\theta|y} =& \underset{\theta}{argmax}\, p(\theta)p(y|\theta) = \underset{\theta}{argmax} \ln p(\theta)p(y|\theta) \\
C_{\theta|y} =& -\left(\frac{\partial^2}{\partial\theta^2}p(\theta)p(y|\theta)\right)^{-1}
\end{aligned}$$

So we can directly compute the values:

$$\begin{aligned}
\frac{\partial}{\partial\theta}\ln p(\theta)p(y|\theta) + const =& \frac{\partial}{\partial\theta}\left[-\frac{1}{2}\left(\theta-\mu_0\right)^T\Sigma_0^{-1}\left(\theta-\mu_0\right) - \frac{1}{2}\left(H\theta-y\right)^T\Gamma^{-1}\left(H\theta-y\right)\right] \\
=& -\Sigma_0^{-1}(\theta-\mu_0) - H^T\Gamma^{-1}(H\theta-y) \\
=& -\Sigma_0^{-1}\theta + \Sigma_0^{-1}\mu_0 - H^T\Gamma^{-1}H\theta + H^T\Gamma^{-1}y \\
=& \Sigma_0^{-1}\mu_0 + H^T\Gamma^{-1}y - \left(\Sigma_0^{-1} + H^T\Gamma^{-1}H\right)\theta \overset{!}{=} 0 \\
iff\ & \Sigma_0^{-1}\mu_0 + H^T\Gamma^{-1}y = \left(\Sigma_0^{-1} + H^T\Gamma^{-1}H\right)\theta \\
iff\ & \mu_{\theta|y} = \left(\Sigma_0^{-1} + H^T\Gamma^{-1}H\right)^{-1}\left(\Sigma_0^{-1}\mu_0 + H^T\Gamma^{-1}y\right)
\end{aligned}$$

And:

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}p(\theta)p(y|\theta) =& \frac{\partial}{\partial\theta}\left[-\Sigma_0^{-1}(\theta-\mu_0) - H^T\Gamma^{-1}H\theta + H^T\Gamma^{-1}y\right] \\
=& -\Sigma_0^{-1} - H^T\Gamma^{-1}H
\end{aligned}$$

So:

$$C_{\theta|y} = \left(\Sigma_0^{-1} + H^T\Gamma^{-1}H\right)^{-1}$$

Finally:

$$p(\theta|y) \sim \mathcal{N}\left(\left(\Sigma_0^{-1} + H^T\Gamma^{-1}H\right)^{-1}\left(\Sigma_0^{-1}\mu_0 + H^T\Gamma^{-1}y\right), \left(\Sigma_0^{-1} + H^T\Gamma^{-1}H\right)^{-1}\right)$$

1.2 Product of Kernels

1.2.1 Q4 $k(x,y) = \sum_i\sum_j g_i(x)f_j(x)f_j(y)g_i(y)$

Let $\{x_i\}_{i=1}^N \subseteq \mathbb{R}^N$. Denote $X = span\left(\{x_i\}_{i=1}^N\right)$ and $K = [k(x_i,x_j)]_{i,j=1}^N$ Gram's matrix for $k(\cdot,\cdot)$ and $\{x_i\}_{i=1}^N$. Since $k(\cdot,\cdot)$ is a valid kernel function, we can write $K = RR^T$ for some R and U is a basis transformation matrix such that $Ux_i = e_i$. Now:

$$\begin{aligned}
f(x) \triangleq& R^TUx \\
f^T(x_i)f(x_j) =& \left(R^TUx_i\right)^T R^TUx_j = x^TU^TRR^TUx = e_iKe_j = [K]_{i,j} = k(x_i,x_j)
\end{aligned}$$

Now let f, g be functions which hold the above equality for k_1,k_2 respectively. So:

$$k(x,y) = k_1(x,y)\cdot k_2(x,y) = f^T(x)f(y)\cdot g^T(x)g(y)$$

Denote $f^T(x) = [\cdots, f_i(x), \cdots]$ and the same for g. Now:

$$k(x,y) = \left(\sum_i f_i(x)f_i(y)\right) \cdot \left(\sum_j g_j(x)g_j(y)\right) = \sum_j \left(\sum_i f_i(x)f_i(y)g_j(x)g_j(y)\right)$$

Q.E.D.

1.2.2 Q5 $k(x,y) = k_1(x,y) \cdot k_2(x,y)$ Is A Valid Kernel

Finally, given such kernels and their respective functions f and g (as in Q4) we define:

$$h(x) = reshape\left(f(x)g^T(x), (, N \times N)\right) = reshape\left(\begin{bmatrix} f_1(x)g_1(x) & \cdots & f_1(x)g_n(x) \\ \vdots & \ddots & \vdots \\ f_n(x)g_1(x) & \cdots & f_n(x)g_n(x) \end{bmatrix}, (, N \times N)\right)$$

$$= [f_1(x)g_1(x), ..., f_1(x)g_n(x), f_2(x)g_1(x), ..., f_2(x)g_n(x), f_n(x)g_1(x), ..., f_n(x)g_n(x)]^T$$

So we get:

$$h^T(x)h(y) = \sum_j \sum_i f_i(x)g_i(x)f_j(y)g_j(y)$$

and $k(\cdot,\cdot)$ is a valid kernel as a dot product of to vector functions. Q.E.D

1.3 Kernel Functions

1.3.1 Q6 $k(x,y) = \exp\left[\beta \left\|g(x) - g(y)\right\|^2\right]$ Is Not Valid

Let $\beta = 1$ and $g(x) \triangleq x$ so, for $D = \{x,y\}$we get:

$$K = \begin{bmatrix} 1 & \exp\left[\|x-y\|\right] \\ \exp\left[\|x-y\|\right] & 1 \end{bmatrix}$$

and

$$det(K) = 1 - \exp\left[2\left\|x-y\right\|\right]$$

So, $det(K) < 0$ iff $1 < \exp\left[2\left\|x-y\right\|\right]$ iff $0 < 2\left\|x-y\right\|$ iff $0 < \left\|x-y\right\|$. I.e. for every strictlly diffarent x, y K is not PSD. If it where, we'd get $0 \leq det(K)$ which isn't the case here. Q.E.D.

Note $p(x,y) \triangleq \left\|g(x) - g(y)\right\|^2 = (g(x) - g(y))^T I (g(x) - g(y))$ is a valid kernel as I is PD. So, for any constant $\beta > 0$ it holds that $q \triangleq \beta \cdot p$ is a valid kernel. Lastly, $k = \exp(q(x,y))$ is a valid kernel . Q.E.D.

1.3.2 Q7 $k(x,y) = k_1(x,y) - k_2(x,y)$ Is Not Valid

Let $k_1(x,y) \triangleq xy$. K is a valid kernel as a product of the identity function of real numbers. Also $k_2(x,y) = 2 \cdot k_1(x,y)$ is a valid kernel as the multiplication of a valid kernel by a positive constant. Now $k(x,y) = k_1(x,y) - k_2(x,y) = -xy$ is not a valid kernel because $\forall x \to k(x,x) \leq 0$, specifically for the unit vector $e_1$and $k's$ Gram matrix Kwe get:

$$e_1^T K e_1 = -1$$

So K is not a PSD.

1.3.3 Q8 $k(x,y) = k_a(x_a, y_a) + k_b(x_b, y_b)$ Is Valid

Note that in this case $k's$ Gram matrix $K = K_a + K_b$. For each vector v it holds that $0 \leq v^T K_x v$ (x=a,b) as $k_x$is a valid kernel and so:

$$0 \leq v^T K_a v + v^T K_b v = v^T K v$$

Q.E.D.

1.3.4 Q9 $k(x,y) = \sqrt{\ell^T(x)\ell(y)}$ Is Not Valid

Proof online...

1.4 Evidence in the Dual Space

1.4.1 Q10 $p(y|k(\cdot,\cdot),\sigma_2) = \mathcal{N}\left(y\,|\,0, K + I\sigma^2\right)$

Note $y = K\alpha + \eta$ and as such is an affine transformation of a Gaussian and so - a Gaussian. We just need to find its mean and cov:

$$\mathbb{E}\left[y|k(\cdot,\cdot),\sigma_2\right] = \mathbb{E}\left[K\alpha + \eta\right] = K\mathbb{E}\left[\alpha\right] + \mathbb{E}\left[\eta\right] = K0 + 0 = 0$$

$$\text{cov}\left[y|k(\cdot,\cdot),\sigma_2\right] = \text{cov}\left[K\alpha + \eta\right] = \text{cov}\left[K\alpha\right] + \text{cov}\left[\eta\right] = K^T\text{cov}\left[\alpha\right]K + I\sigma^2 = K^T K^{-1} K + I\sigma^2 = K + I\sigma^2$$

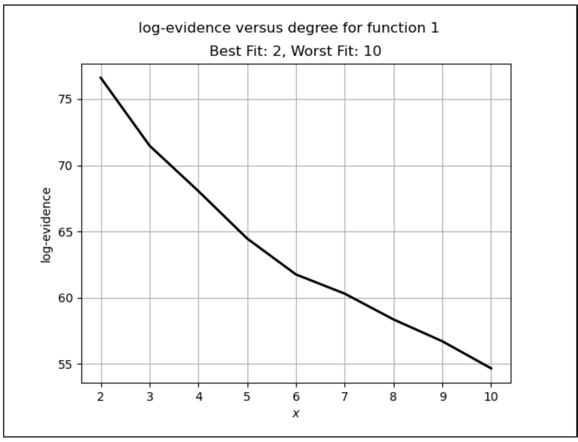
Where the last transition is due to K's symmetry. Hence:

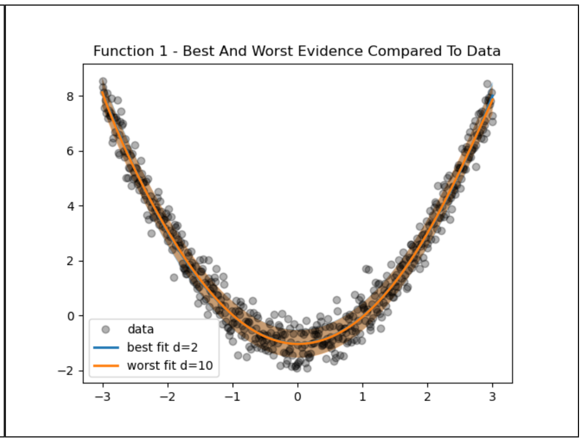
$$p(y|k(\cdot,\cdot),\sigma_2) = \mathcal{N}\left(y\,|\,0, K + I\sigma^2\right)$$

2 Practical

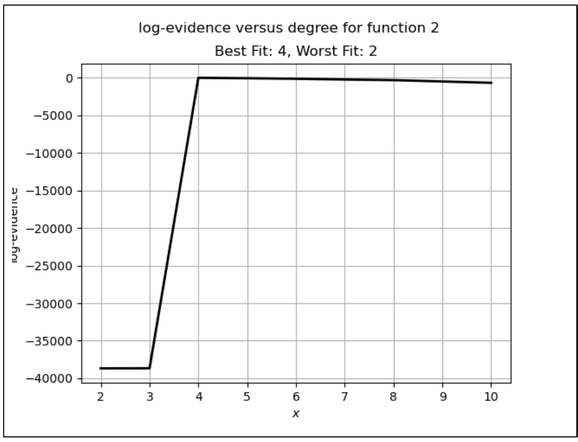
2.1 Evidence for Artificial Functions

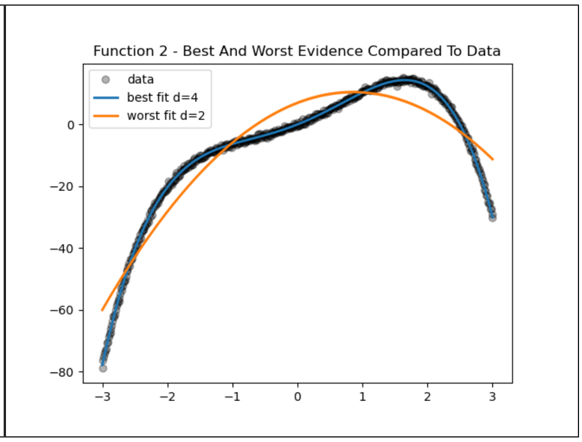
2.1.1 $f_1(x) = x^2 - 1$



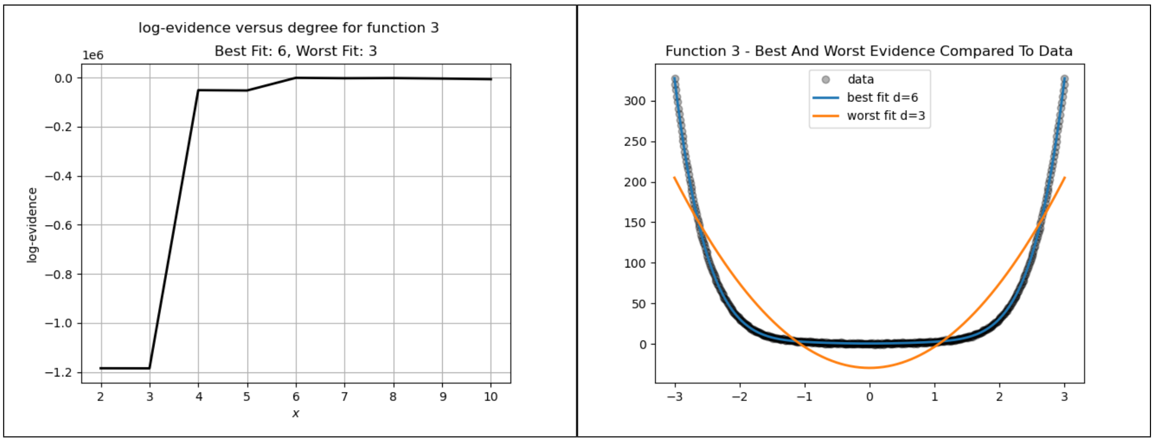


2.1.2 $f_2(x) = -x^4 + 3x^2 + 50\sin\left(\frac{x}{6}\right)$

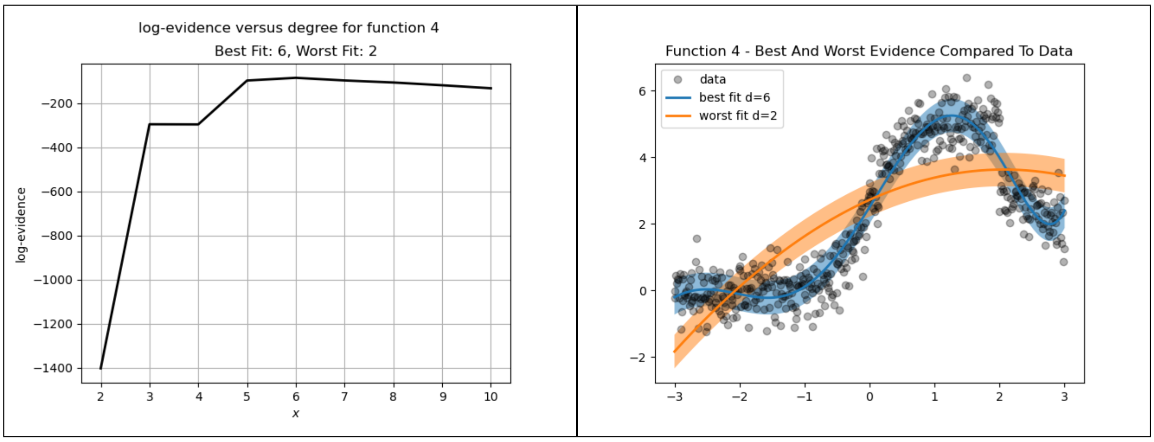




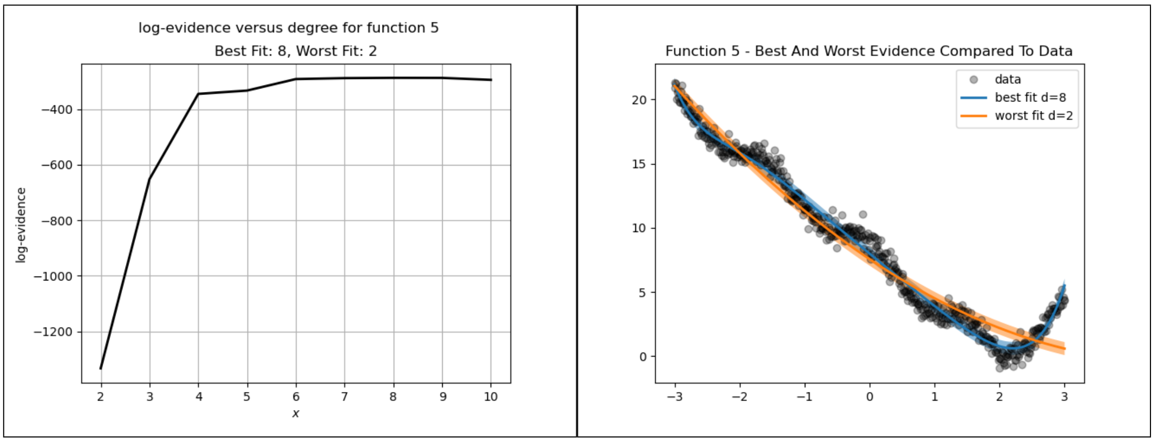
2.1.3 $f_3(x) = \frac{1}{2}x^6 - 0.75x^4 + 2.75x^2$



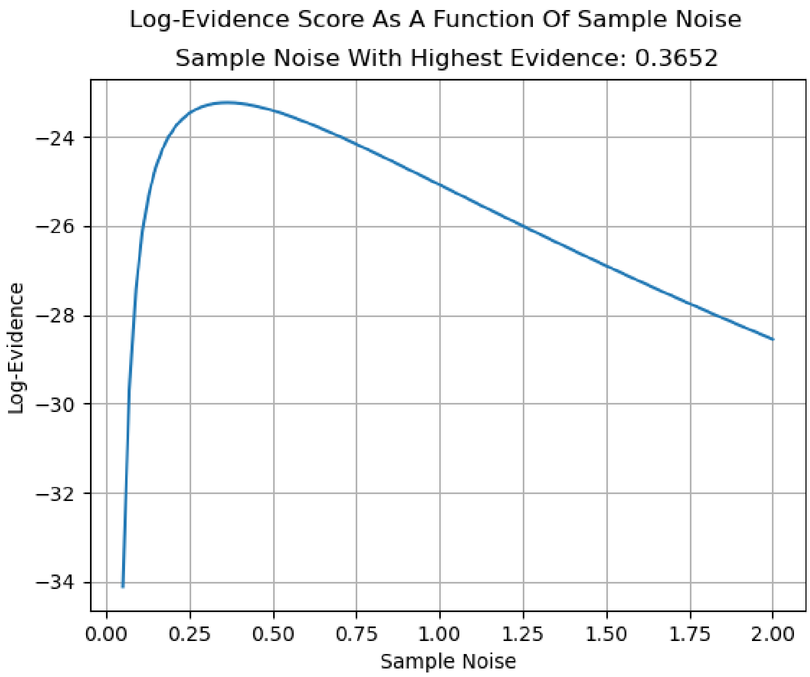
2.1.4 $f_4(x) = \frac{5}{1+e^{-4x}} - \begin{cases} x & x-2>0 \\ 0 & o.w. \end{cases}$



2.1.5 $f_5(x) = \cos 4x + 4\,|x-2|$



2.2 Estimating the Sample Noise for Temperature Prediction



2.2.1 Q7 No

The sample noise with the highest evidence is not necessarily the sample noise of the original measurements. When creating our model we assume a Gaussian prior on θ, we assume the noise is also a Gaussian and we assume the temperatures were decided upon using a polynomial with degree ≤ 7. In case those assumptions hold we cold say we've found the noise. As it's extremely unlikely though, the noise with the high-test evidence is not necessarily the sample noise of the original measurements.