Exercises
Advanced Machine Learning
Fall 2019

Machine Learning Laboratory
Dept. of Computer Science, ETH Zürich
Prof. Joachim M. Buhmann
Web https://ml2.inf.ethz.ch/courses/aml/

Series 3, 25 Oct 2019
(GPs; Model assessment and selection)

Teaching assistant: **Joanna Ficek**
joanna.ficek@inf.ethz.ch

Note: These are sample solutions. If you solved the problem in a different way it doesn't necessarily mean that your solution is wrong.

Solution 1 (Gaussian Processes cont.):

Solutions are based on Christopher M. Bishop, *Pattern Recognition and Machine Learning*. Springer Verlag (2006):

a) 1. If $k_1(\boldsymbol{x}, \boldsymbol{x}^{'})$ is a valid kernel, then there must exist a feature vector $\boldsymbol{\phi}(\boldsymbol{x})$ such that

$$k_1(\boldsymbol{x}, \boldsymbol{x}^{'}) = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{x}^{'}) \ .$$

It follows that

$$ck_1(\boldsymbol{x}, \boldsymbol{x}^{'}) = \boldsymbol{u}(\boldsymbol{x})^T \boldsymbol{u}(\boldsymbol{x}^{'})$$

where

$$\boldsymbol{u}(\boldsymbol{x}) = c^{1/2} \boldsymbol{\phi}(\boldsymbol{x})$$

and so $ck_1(\boldsymbol{x}, \boldsymbol{x}^{'})$ can be expressed as the scalar product of feature vectors, and hence is a valid kernel.

2. Similarly as above we can write

$$f(\boldsymbol{x}) k_1(\boldsymbol{x}, \boldsymbol{x}^{'}) f(\boldsymbol{x}^{'}) = \boldsymbol{v}(\boldsymbol{x})^T \boldsymbol{v}(\boldsymbol{x}^{'})$$

with

$$\boldsymbol{v}(\boldsymbol{x}) = f(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x}) \ .$$

As before we can see that $f(\boldsymbol{x}) k_1(\boldsymbol{x}, \boldsymbol{x}^{'}) f(\boldsymbol{x}^{'})$ can be expressed as the scalar product of feature vectors, and hence is a valid kernel.

3. We also know, that a necessary and sufficient condition for a function to be a valid kernel is that the Gram matrix $\boldsymbol{K}$, which elements are given by $k_1(\boldsymbol{x}, \boldsymbol{x}^{'})$ should be positive semidefinite for all possible choices of the set $\{\boldsymbol{x}_n\}$. A matrix $\boldsymbol{K}$ is positive semidefinite if, and only if,

$$\boldsymbol{a}^T \boldsymbol{K} \boldsymbol{a} \geq 0$$

for any choice of the vector $\boldsymbol{a}$. Let $\boldsymbol{K}_1$ be the Gram matrix for $k_1(\boldsymbol{x}, \boldsymbol{x}^{'})$ and let $\boldsymbol{K}_2$ be the Gram matrix for $k_2(\boldsymbol{x}, \boldsymbol{x}^{'})$. Then

$$\boldsymbol{a}^T (\boldsymbol{K}_1 + \boldsymbol{K}_2) \boldsymbol{a} = \boldsymbol{a}^T \boldsymbol{K}_1 \boldsymbol{a} + \boldsymbol{a}^T \boldsymbol{K}_1 \boldsymbol{a} \geq 0$$

where we have used the fact that $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ are positive semi-definite matrices, as well as the fact that the sum of two non-negative numbers will itself be non-negative. Thus, $k(\boldsymbol{x}, \boldsymbol{x}^{'})$ defines a valid kernel.

4. Since we know that $k_1(\boldsymbol{x},\boldsymbol{x}^{'})$ and $k_2(\boldsymbol{x},\boldsymbol{x}^{'})$ are valid kernels, we know that there exist mappings $\boldsymbol{\phi}(\boldsymbol{x})$ and $\boldsymbol{\psi}(\boldsymbol{x})$ such that

$$k_1(\boldsymbol{x},\boldsymbol{x}^{'}) = \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{x}^{'})$$

and

$$k_1(\boldsymbol{x},\boldsymbol{x}^{'}) = \boldsymbol{\psi}(\boldsymbol{x})^T\boldsymbol{\psi}(\boldsymbol{x}^{'}) \ .$$

Hence

$$\begin{aligned}
k(\boldsymbol{x},\boldsymbol{x}^{'}) &= k_1(\boldsymbol{x},\boldsymbol{x}^{'})k_2(\boldsymbol{x},\boldsymbol{x}^{'}) \\
&= \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{x}^{'})\boldsymbol{\psi}(\boldsymbol{x})^T\boldsymbol{\psi}(\boldsymbol{x}^{'}) \\
&= \sum_{m=1}^{M}\phi_m(\boldsymbol{x})\phi_m(\boldsymbol{x}^{'})\sum_{n=1}^{N}\psi_n(\boldsymbol{x})\psi_n(\boldsymbol{x}^{'}) \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N}\phi_m(\boldsymbol{x})\phi_m(\boldsymbol{x}^{'})\psi_n(\boldsymbol{x})\psi_n(\boldsymbol{x}^{'}) \\
&= \sum_{k=1}^{K}\psi_k(\boldsymbol{x})\psi_k(\boldsymbol{x}^{'}) \\
&= \boldsymbol{\psi}(\boldsymbol{x})^T\boldsymbol{\psi}(\boldsymbol{x}^{'})
\end{aligned}$$

where $K = MN$ and

$$\psi_k(\boldsymbol{x}) = \phi_{(k-1)\oslash N)+1}(\boldsymbol{x})\psi_{(k-1)\odot N)+1}(\boldsymbol{x})$$

where in turn $\oslash$ and $\odot$ denote integer division and remainder, respectively. Again we can see that $k(\boldsymbol{x},\boldsymbol{x}^{'})$ can be expressed as the scalar product of feature vectors, and hence is a valid kernel.

b) The RBF kernel was used to generate the samples. σ corresponds to the vertical scaling and l to the horizontal scaling (for more information see the tutorial and `https:/www.jgoertler.com/visual-exploration-gaussian-p`

1. a:B ($\sigma = 0.8$, $l = 0.5$);
2. c:A ($\sigma = 0.8$, $l = 2$);
3. b:C ($\sigma = 0.33$, $l = 0.5$);

Solution 2 (Efficient Leave-One-Out Cross Validation):

The derivations include three steps:

Step 1.

$$\begin{aligned}
\boldsymbol{w}^*_{(-i)} &= \left(\mathbf{X}_{(-i)}\mathbf{X}^T_{(-i)} + \frac{(n-1)\mu}{2}\mathbf{I}\right)^{-1}\mathbf{X}_{(-i)}\boldsymbol{y}_{(-i)} & (1)\\
&= \left(\mathbf{A} - \boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1}(\mathbf{X}\boldsymbol{y} - \boldsymbol{x}_i y_i) & (2)\\
&\overset{\text{SM}}{=} \mathbf{A}^{-1}(\mathbf{X}\boldsymbol{y} - \boldsymbol{x}_i y_i) + \frac{\mathbf{A}^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T\mathbf{A}^{-1}(\mathbf{X}\boldsymbol{y} - \boldsymbol{x}_i y_i)}{1 - \boldsymbol{x}_i^T\mathbf{A}^{-1}\boldsymbol{x}_i} & (3)\\
&= \mathbf{A}^{-1}\mathbf{X}\boldsymbol{y} - \mathbf{A}^{-1}\boldsymbol{x}_i\left(1 + \frac{x_i^T\mathbf{A}^{-1}\boldsymbol{x}_i}{1 - \boldsymbol{x}_i^T\mathbf{A}^{-1}\boldsymbol{x}_i}\right)y_i + \frac{\mathbf{A}^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T\mathbf{A}^{-1}(\mathbf{X}\boldsymbol{y})}{1 - \boldsymbol{x}_i^T\mathbf{A}^{-1}\boldsymbol{x}_i} & (4)\\
&= \mathbf{A}^{-1}\mathbf{X}\boldsymbol{y} - \mathbf{A}^{-1}\boldsymbol{x}_i\left(\frac{1}{1 - \boldsymbol{x}_i^T\mathbf{A}^{-1}\boldsymbol{x}_i}\right)y_i + \frac{\mathbf{A}^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T\mathbf{A}^{-1}(\mathbf{X}\boldsymbol{y})}{1 - \boldsymbol{x}_i^T\mathbf{A}^{-1}\boldsymbol{x}_i}. & (5)
\end{aligned}$$

Step 2.

$$\boldsymbol{x}_i^T \boldsymbol{w}^*_{(-i)} \overset{\text{Eq. (5)}}{=} \boldsymbol{x}_i^T \mathbf{A}^{-1}\mathbf{X}\boldsymbol{y} + \frac{\boldsymbol{x}_i^T \mathbf{A}^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T\mathbf{A}^{-1}(\mathbf{X}\boldsymbol{y})}{1-\boldsymbol{x}_i^T\mathbf{A}^{-1}\boldsymbol{x}_i} - \left(\frac{1}{1-\boldsymbol{x}_i^T\mathbf{A}^{-1}\boldsymbol{x}_i}\right)\boldsymbol{x}_i^T\mathbf{A}^{-1}\boldsymbol{x}_i y_i \tag{6}$$

$$= \left(1+\frac{\boldsymbol{x}_i^T\mathbf{A}^{-1}\boldsymbol{x}_i}{1-\boldsymbol{x}_i^T\mathbf{A}^{-1}\boldsymbol{x}_i}\right)\boldsymbol{x}_i^T\mathbf{A}^{-1}\mathbf{X}\boldsymbol{y} - \left(\frac{1}{1-\boldsymbol{x}_i^T\mathbf{A}^{-1}\boldsymbol{x}_i}\right)\boldsymbol{x}_i^T\mathbf{A}^{-1}\boldsymbol{x}_i y_i \tag{7}$$

$$= \left(\frac{1}{1-\boldsymbol{x}_i^T\mathbf{A}^{-1}\boldsymbol{x}_i}\right)\boldsymbol{x}_i^T\mathbf{A}^{-1}\left(\mathbf{X}\boldsymbol{y}-\boldsymbol{x}_i y_i\right) \tag{8}$$

$$= \left(\frac{1}{1-s_i}\right)\left(\hat{y}_i - s_i y_i\right). \tag{9}$$

Step 3.

$$y_i - \boldsymbol{x}_i^T\boldsymbol{w}^*_{(-i)} \overset{\text{Eq. (9)}}{=} y_i - \left(\frac{1}{1-s_i}\right)\left(\hat{y}_i - s_i y_i\right) \tag{10}$$

$$= \left(1+\frac{s_i}{1-s_i}\right)y_i - \left(\frac{1}{1-s_i}\right)\hat{y}_i \tag{11}$$

$$= \left(\frac{1}{1-s_i}\right)\left(y_i - \hat{y}_i\right). \tag{12}$$

Solution 3 (Jackknife estimator):

a)The cumulative distribution function is

$$\mathbb{P}(X_{(n)} \le x) = \mathbb{P}(X_1 \le x, X_2 \le x, \ldots, X_n \le x) = \prod_{i=1}^n \mathbb{P}(X_i \le x) = \left(\frac{x}{\theta}\right)^n.$$

Hence we can compute the probability density function (PDF) as

$$p_{(n)}(x) = \frac{d}{dx}\mathbb{P}(X_{(n)} \le x) = \frac{n}{\theta}\left(\frac{x}{\theta}\right)^{n-1}.$$

The average of the estimator is then

$$\mathbb{E}[X_{(n)}] = \int_0^\theta dx\, x\frac{n}{\theta}\left(\frac{x}{\theta}\right)^{n-1} = \frac{n}{n+1}\theta$$

One can see that the estimator $X_{(n)}$ underestimates the bound θ.

b) The replicate estimator $\hat{S}_{n-1}^{(-i)}$ is the maximum of the samples $\{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$. Let assume $i^* = \arg\max_i X_i$. Then the replicate estimator reads

$$\hat{S}_{n-1}^{(-i)} = \begin{cases} X_{(n)} & i \ne i^* \\ X_{(n-1)} & i = i^* \end{cases}$$

c) The Jackknife estimator reads

$$\hat{S}_n^{JK} = \hat{S}_n - (n-1)\left(\frac{1}{n}\sum_{i=1}^n \hat{S}_{n-1}^{(-i)} - \hat{S}_n\right) = X_{(n)} - (n-1)\left(\frac{n-1}{n}X_{(n)} + \frac{1}{n}X_{(n-1)} - X_{(n)}\right)$$
$$= X_{(n)} + \frac{n-1}{n}\left(X_{(n)} - X_{(n-1)}\right).$$

Hence, the Jackknife estimator modifies the estimator $\hat{S}_n$ by adding a positive correction.

d) The cumulative distribution function (CDF) of $X_{(n)}$ is

$$\begin{aligned}
\mathbb{P}(X_{(n-1)} \leq x) &= \mathbb{P}(X_{(n-1)} \leq x, X_{(n)} \leq x) + \mathbb{P}(X_{(n-1)} \leq x, X_{(n)} \geq x) \\
&= \mathbb{P}(X_{(n)} \leq x) + \sum_{i=1}^{n} \mathbb{P}(X_i \geq x, X_1 \leq x, \ldots, X_{i-1} \leq x, X_{i+1} \leq x, \ldots, X_n \leq x) \\
&= \left(\frac{x}{\theta}\right)^n + n\frac{\theta - x}{\theta}\left(\frac{x}{\theta}\right)^{n-1}.
\end{aligned}$$

Therefore, the probability density function (PDF) is

$$p_{(n-1)}(x) = \frac{d}{dx}\mathbb{P}(X_{(n-1)} \leq x) = \frac{n(n-1)}{\theta}\left(\frac{x}{\theta}\right)^{n-2}\left(1 - \frac{x}{\theta}\right).$$

We can now compute the expected value as

$$\mathbb{E}[X_{(n-1)}] = \int_0^\theta x p_{(n-1)}(x)dx = \int_0^\theta x\frac{n(n-1)}{\theta}\left(\frac{x}{\theta}\right)^{n-2}\left(1 - \frac{x}{\theta}\right)dx = \frac{n-1}{n+1}\theta,$$

and, finally,

$$\mathbb{E}[\hat{S}_n^{JK}] = \left(1 - \frac{1}{n^2 + n}\right)\theta$$

Note that the bias of the Jackknife estimator $\hat{S}_n^{JK}$ is smaller than the bias of the original estimator $\hat{S}_n$ by a factor of n. Also note that the Jackknife estimator does not exploit any property of the distribution, and thus it can be used also in cases where the actual distribution is unknown.

Solution 4 (Model selection: Bayesian Information Criterion):

BIC can be seen as a large n approximation to the log model evidence and hence, in the below solution we consider $n \to \infty$. We use the Laplace approximation to the log model evidence around the mode of the posterior distribution $p(\theta^{(m)}|\mathcal{D}^{(n)})$

$$\ln p(\mathcal{D}^{(n)}) \approx \ln p(\mathcal{D}^{(n)}|\theta^{(m)}_{\text{MAP}}) + \ln p(\theta^{(m)}_{\text{MAP}}) + \frac{m}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}|,$$

where $\mathbf{A} = -\frac{\partial^2}{\partial\theta^{(m)}\partial\theta^{(m)}}\ln p(\theta^{(m)}_{\text{MAP}}|\mathcal{D}^{(n)})$ is the Hessian of the minus log posterior at $\theta^{(m)}_{\text{MAP}}$.

We neglect the log prior $\ln p(\theta^{(m)}_{\text{MAP}})$ at $\theta^{(m)}_{\text{MAP}}$ and consider a simple case where the Hessian $\mathbf{A} \in \mathbb{R}^{m\times m}$ is a diagonal matrix. We assume that the Hessian has a full rank. In addition, we assume the data to be iid, which allows us to write the likelihood as a product $p(\mathcal{D}^{(n)}|\theta^{(m)}_{\text{MAP}}) = \prod_{i=1}^n p(\boldsymbol{x}_i|\theta^{(m)}_{\text{MAP}})$, and therefore,

$$\begin{aligned}
[\mathbf{A}]_{jj} &= \frac{\partial^2}{\partial\theta_j^2}\ln p(\theta^{(m)}_{\text{MAP}}|\mathcal{D}^{(n)}) && (13)\\
&= \frac{\partial^2}{\partial\theta_j^2}\left[\ln p(\mathcal{D}^{(n)}|\theta^{(m)}_{\text{MAP}}) + \ln p(\theta^{(m)}_{\text{MAP}}) - \ln p(\mathcal{D}^{(n)})\right] && (14)\\
&= \frac{\partial^2}{\partial\theta_j^2}\left[\ln p(\mathcal{D}^{(n)}|\theta^{(m)}_{\text{MAP}}) + \ln p(\theta^{(m)}_{\text{MAP}})\right] && (15)\\
&= \sum_{i=1}^n \frac{\partial^2}{\partial\theta_j^2}\ln p(\boldsymbol{x}_i|\theta^{(m)}_{\text{MAP}}) + \frac{\partial^2}{\partial\theta_j^2}\ln p(\theta^{(m)}_{\text{MAP}}) && (16)\\
&\sim nc_j. && (17)
\end{aligned}$$

Thus, $|\mathbf{A}| \sim n^m \prod_j c_j$, which leads us to the following result

$$\ln|\mathbf{A}| \sim m \ln n.$$

Finally, one can write

$$-2\ln p(\mathcal{D}^{(n)}) \approx -2\ln p(\mathcal{D}^{(n)}|\theta^{(m)}) + m\ln n = \mathsf{BIC}(\mathcal{D}^{(n)}).$$

Another way of solving the exercise can be found in Kevin P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press (2012), p.255, 256.