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. October 18, 2019
(GPs; Model assessment and selection)

Teaching assistant: **Joanna Ficek**
ficekj@inf.ethz.ch

Problem 1 (Gaussian Processes cont.):

The following exercises are related to the previous topic of Gaussian Processes.
Part a) is dedicated to students new to kernels, however is a good practice to everyone.

a) Given valid kernels $k_1(\boldsymbol{x}, \boldsymbol{x}')$ and $k_2(\boldsymbol{x}, \boldsymbol{x}')$, show that the following new kernels are also valid:

1. $k(\boldsymbol{x}, \boldsymbol{x}') = ck_1(\boldsymbol{x}, \boldsymbol{x}')$, with constant $c > 0$;
2. $k(\boldsymbol{x}, \boldsymbol{x}') = f(\boldsymbol{x})k_1(\boldsymbol{x}, \boldsymbol{x}')f(\boldsymbol{x}')$, with any function $f(.)$;
3. $k(\boldsymbol{x}, \boldsymbol{x}') = k_1(\boldsymbol{x}, \boldsymbol{x}') + k_2(\boldsymbol{x}, \boldsymbol{x}')$;
4. $k(\boldsymbol{x}, \boldsymbol{x}') = k_1(\boldsymbol{x}, \boldsymbol{x}')k_2(\boldsymbol{x}, \boldsymbol{x}')$;

b) We investigate the influence of kernels and parameters on the prior distribution of functions. Decide which kernel is used in the following examples and match the parameters used to generate the resulting covariance matrices and the functions sampled from the corresponding prior distributions.

Kernels:

- RBF kernel: $\sigma^2 exp\big(-\frac{||t-t'||^2}{2l^2}\big)$
- periodic kernel: $\sigma^2 exp\big(-\frac{2sin^2(\pi|t-t'|/p)}{l^2}\big)$

Parameter settings:

1. $\sigma = 0.8,\ l = 0.5,\ (p = 0.5)$
2. $\sigma = 0.8,\ l = 2,\ (p = 0.5)$
3. $\sigma = 0.33,\ l = 0.5,\ (p = 0.5)$

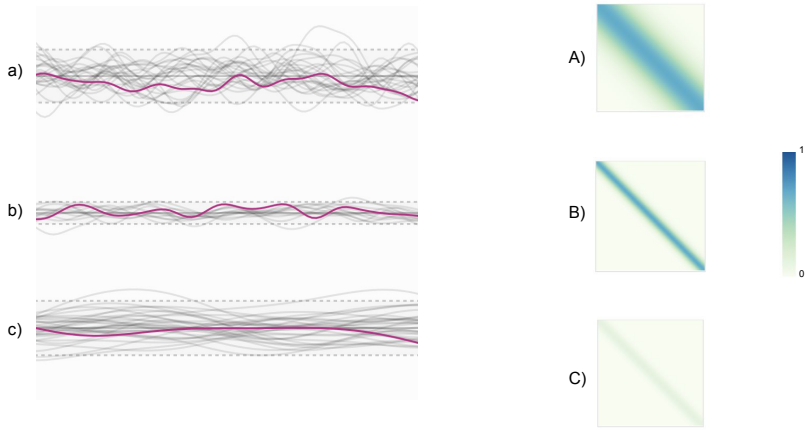

Figure 1: Samples from prior distributions and the corresponding covariance matrices.
Source: `https://www.jgoertler.com/visual-exploration-gaussian-processes/`.

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

Here we propose a computationally efficient method to compute the error of the leave-one-out cross-validation (LOOCV) for the ridge regression model. Given an output vector $\boldsymbol{y} \in \mathbb{R}^n$ and an input matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$ with n columns for n data points $\boldsymbol{x}_i \in \mathbb{R}^d$, the objective of ridge regression is defined as

$$R_\mu(\boldsymbol{w}) = \frac{1}{n}\|\mathbf{X}^T\boldsymbol{w} - \boldsymbol{y}\|^2 + \frac{\mu}{2}\|\boldsymbol{w}\|^2. \tag{1}$$

Let $\mathbf{X}_{(-i)} \in \mathbb{R}^{d \times (n-1)}$ denote the data matrix that is obtained by excluding column i form the full matrix $\mathbf{X}$. To compute the LOOCV error, one has to compute the minimizer $\boldsymbol{w}^*_{(-i)}$, which is defined as

$$\boldsymbol{w}^*_{(-i)} = \arg\min_{\boldsymbol{w}} \left(\frac{1}{n-1}\|\mathbf{X}^T_{(-i)}\boldsymbol{w} - \boldsymbol{y}_{(-i)}\|^2 + \frac{\mu}{2}\|\boldsymbol{w}\|^2 \right). \tag{2}$$

for each $i \in \{1, \ldots, n\}$ (remember: we have n folds and n hold-out datasets in LOOCV).
Then the cross-validation error is

$$\epsilon = \frac{1}{n}\sum_{i=1}^{n} \left(\langle \boldsymbol{x}_i, \boldsymbol{w}^*_{(-i)} \rangle - y_i \right)^2. \tag{3}$$

We want to prove that this error can be efficiently computed as

$$\epsilon = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - s_i} \right)^2, \tag{4}$$

where

$$\hat{y}_i = \boldsymbol{x}_i^T \mathbf{A}^{-1}\mathbf{X}\boldsymbol{y}, \tag{5}$$
$$s_i = \boldsymbol{x}_i^T \mathbf{A}^{-1}\boldsymbol{x}_i, \tag{6}$$
$$\mathbf{A} = \left(\mathbf{X}\mathbf{X}^T + \frac{(n-1)\mu}{2}\mathbf{I} \right). \tag{7}$$

To prove the above result, we follow three steps:

a. Prove that

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

Hint: Use Sherman-Morrison formula

$$(\mathbf{M} - \boldsymbol{u}\boldsymbol{v}^T)^{-1} = \mathbf{M}^{-1} + \frac{\mathbf{M}^{-1}\boldsymbol{u}\boldsymbol{v}^T\mathbf{M}^{-1}}{1 - \boldsymbol{v}^T\mathbf{M}^{-1}\boldsymbol{u}}, \tag{9}$$

where $\mathbf{M}$ is an invertible square matrix and $\boldsymbol{u}$, $\boldsymbol{v}$ are column vectors.

b. Derive

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

c. Show that

$$y_i - \boldsymbol{x}_i^T\boldsymbol{w}^*_{(-i)} = \left(\frac{1}{1 - s_i} \right)(y_i - \hat{y}_i). \tag{11}$$

Problem 3 (Jackknife estimator):

Assume X is a random variable uniformly distributed $X \sim \mathcal{U}[0, \theta]$, with unknown upper bound θ. An intuitive estimator for n given samples $\{X_1, X_2, \ldots, X_n\}$ is the maximum $\hat{S}_n = X_{(n)}$ (we denote by $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ the samples in ascending order).

a) Prove that the expected value of the estimator is $\frac{n}{n+1}\theta$, hence the estimator $\hat{S}_n$ is biased.
Hint: first compute the probability $P(X_{(n)} \leq x)$ and then differentiate to find the probability density function.

b) Compute the replicate estimator $\hat{S}_{n-1}^{(-i)}$.
Hint: consider separately the case in which i is the index of $X_{(n)}$ and the case where it is not.

c) Compute the Jackknife estimator $\hat{S}_n^{JK} = \hat{S}_n - (n-1)(\frac{1}{n}\sum_{i=1}^{n} \hat{S}_{n-1}^{(-i)} - \hat{S}_n)$.

d) Prove that the bias of the Jackknife estimator $\hat{S}_n^{JK}$ is smaller than the bias of the estimator $\hat{S}_n = X_{(n)}$.

Problem 4 (Model selection criteria):

Let us consider a data set $\mathcal{D}^{(n)}$ of cardinality n and a model $\mathcal{M}$ with m parameters $\theta^{(m)}$ and a prior $p(\theta^{(m)})$. The Bayesian model selection is based on maximizing the model evidence

$$p(\mathcal{D}^{(n)}) = \int p(\mathcal{D}^{(n)}|\theta^{(m)})p(\theta^{(m)})d\theta.$$

Use the Laplace approximation to the log model evidence around the mode $\theta_{\mathsf{MAP}}^{(m)}$ of the posterior distribution $p(\theta^{(m)}|\mathcal{D}^{(n)})$

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

where $\mathbf{A}$ is the Hessian of the minus log posterior $\ln p(\theta^{(m)}|\mathcal{D}^{(n)})$ at $\theta_{\mathsf{MAP}}^{(m)}$. Assuming the iid distribution of data, show that this approximation can be roughly rewritten as the BIC criterion

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