Exercises for the course
Machine Learning 1
Winter semester 2021/22

Abteilung Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 9

We consider a class optimization problems of the type:

$$\min_\theta J(\theta) \quad \text{s.t.} \quad \forall_{i=1}^m : \; g_i(\theta) = 0 \quad \text{and} \quad \forall_{i=1}^l : \; h_i(\theta) \leq 0$$

For this class of problem, we can build the Lagrangian:

$$\mathcal{L}(\theta, \beta, \lambda) = J(\theta) + \sum_{i=1}^m \beta_i g_i(\theta) + \sum_{i=1}^l \lambda_i h_i(\theta).$$

where $(\beta_i)_i$ and $(\lambda_i)_i$ are the dual variables. According to the Karush-Kuhn-Tucker (KKT) conditions, it is necessary for a solution of this optimization problem that the following constraints are satisfied (in addition to the original constraints of the optimization problem):

$$\frac{\partial \mathcal{L}}{\partial \theta} = 0 \qquad \text{(stationarity)}$$
$$\forall_{i=1}^l : \; \lambda_i \geq 0 \qquad \text{(dual feasibility)}$$
$$\forall_{i=1}^l : \; \lambda_i h_i(\theta) = 0 \qquad \text{(complementary slackness)}$$

We will make use of these conditions to derive the dual form of the kernel ridge regression problem.

Exercise 1: Kernel Ridge Regression with Lagrange Multipliers (10 + 20 + 10 + 10 P)

Let $x_1, \ldots, x_N \in \mathbb{R}^d$ be a dataset with labels $y_1, \ldots, y_N \in \mathbb{R}$. Consider the regression model $f(x) = w^\top \phi(x)$ where $\phi \colon \mathbb{R}^d \to \mathbb{R}^h$ is a feature map and w is obtained by solving the constrained optimization problem

$$\min_{\xi, w} \; \sum_{i=1}^N \frac{1}{2}\xi_i^2 \quad \text{s.t.} \quad \forall_{i=1}^N : \; \xi_i = w^\top \phi(x_i) - y_i \quad \text{and} \quad \frac{1}{2}\|w\|^2 \leq C.$$

where equality constraints define the errors of the model, where the objective function penalizes these errors, and where the inequality constraint imposes a regularization on the parameters of the model.

(a) *Construct* the Lagrangian and *state* the KKT conditions for this problem *(Hint: rewrite the equality constraint as $\xi_i - w^\top \phi(x_i) + y_i = 0$.)*

(b) *Show* that the solution of the kernel regression problem above, expressed in terms of the dual variables $(\beta_i)_i$, and λ is given by:

$$\beta = (K + \lambda I)^{-1} \lambda y$$

where K is the kernel Gram matrix.

(c) *Express* the prediction $f(x) = w^\top \phi(x)$ in terms of the parameters of the dual.

(d) *Explain* how the new parameter λ can be related to the parameter C of the original formulation.

Exercise 2: Programming (50 P)

Download the programming files on ISIS and follow the instructions.

Gaussian Processes

In this exercise, you will implement Gaussian process regression and apply it to a toy and a real dataset. We use the notation used in the paper "Rasmussen (2005). Gaussian Processes in Machine Learning" linked on ISIS.

Let us first draw a training set $X = (x_1, \ldots, x_n)$ and a test set $X_\star = (x_1^\star, \ldots, x_m^\star)$ from a d-dimensional input distribution. The Gaussian Process is a model under which the real-valued outputs $\mathbf{f} = (f_1, \ldots, f_n)$ and $\mathbf{f}_\star = (f_1^\star, \ldots, f_m^\star)$ associated to X and $X_\star$ follow the Gaussian distribution:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_\star \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \Sigma & \Sigma_\star \\ \Sigma_\star^\top & \Sigma_{\star\star} \end{bmatrix} \right)$$

where

$$\begin{aligned} \Sigma &= k(X, X) + \sigma^2 I \\ \Sigma_\star &= k(X, X_\star) \\ \Sigma_{\star\star} &= k(X_\star, X_\star) + \sigma^2 I \end{aligned}$$

and where $k(\cdot, \cdot)$ is the Gaussian kernel function. (The kernel function is implemented in `utils.py`.) Predicting the output for new data points $X_\star$ is achieved by conditioning the joint probability distribution on the training set. Such conditional distribution called posterior distribution can be written as:

$$\mathbf{f}_\star \mid \mathbf{f} \sim \mathcal{N}(\underbrace{\Sigma_\star^\top \Sigma^{-1} \mathbf{f}}_{\boldsymbol{\mu}_\star}, \; \underbrace{\Sigma_{\star\star} - \Sigma_\star^\top \Sigma^{-1} \Sigma_\star}_{C_\star})$$

Having inferred the posterior distribution, the log-likelihood of observing for the inputs $X_\star$ the outputs $\mathbf{y}_\star$ is given by evaluating the distribution $\mathbf{f}_\star \mid \mathbf{f}$ at $\mathbf{y}_\star$:

$$\log p(\mathbf{y}_\star \mid \mathbf{f}) = -\frac{1}{2}(\mathbf{y}_\star - \boldsymbol{\mu}_\star)^\top C_\star^{-1} (\mathbf{y}_\star - \boldsymbol{\mu}_\star) - \frac{1}{2}\log |C_\star| - \frac{m}{2}\log 2\pi$$

where $|\cdot|$ is the determinant. Note that the likelihood of the data given this posterior distribution can be measured both for the training data and the test data.

Part 1: Implementing a Gaussian Process (30 P)

Tasks:

- **Create a class `GP_Regressor` that implements a Gaussian process regressor and has the following three methods:**
  - **`def __init__(self,Xtrain,Ytrain,width,noise):`** Initialize a Gaussian process with noise parameter σ and width parameter w. The variable `Xtrain` is a two-dimensional array where each row is one data point from the training set. The Variable `Ytrain` is a vector containing the associated targets. The function must also precompute the matrix Σ^{-1} for subsequent use by the method `predict()` and `loglikelihood()`.
  - **`def predict(self,Xtest):`** For the test set $X_\star$ of m points received as parameter, return the mean vector of size m and covariance matrix of size $m \times m$ of the corresponding output, that is, return the parameters $(\boldsymbol{\mu}_\star, C_\star)$ of the Gaussian distribution $\mathbf{f}_\star \mid \mathbf{f}$.
  - **`def loglikelihood(self,Xtest,Ytest):`** For a data set $X_\star$ of m test points received as first parameter, return the loglikelihood of observing the outputs $\mathbf{y}_\star$ received as second parameter.

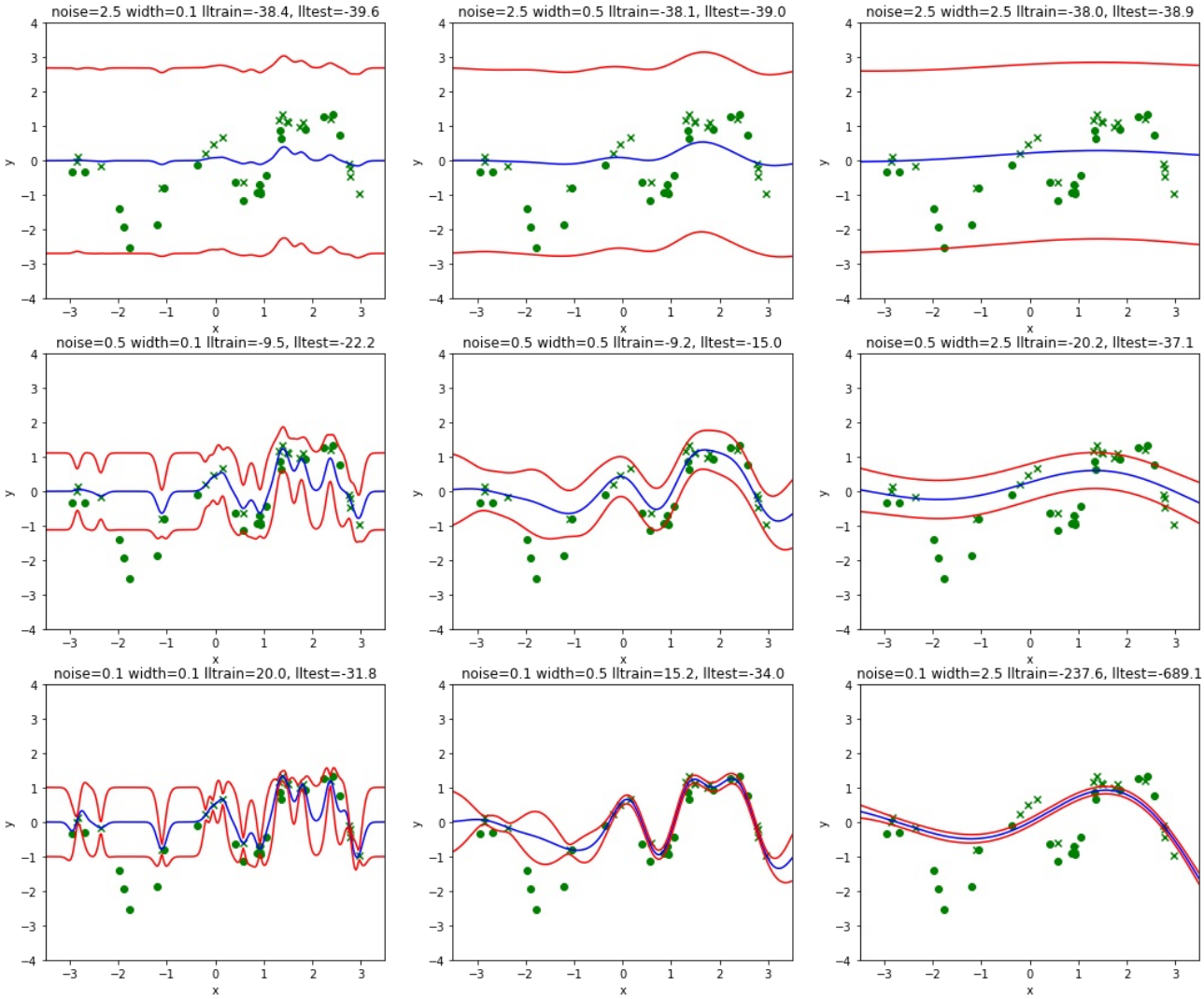
In [1]:

```
# --------------------------
# TODO: Replace by your code
# --------------------------
import solutions
class GP_Regressor(solutions.GP_Regressor):
    pass
# --------------------------
```

- **Test your implementation by running the code below (it visualizes the mean and variance of the prediction at every location of the input space) and compares the behavior of the Gaussian process for various noise parameters σ and width parameters w.**

In [2]:

```
import utils,datasets,numpy
import matplotlib.pyplot as plt
%matplotlib inline

# Open the toy data
Xtrain,Ytrain,Xtest,Ytest = utils.split(*datasets.toy())

# Create an analysis distribution
Xrange = numpy.arange(-3.5,3.51,0.025)[:,numpy.newaxis]

f = plt.figure(figsize=(18,15))

# Loop over several parameters:
for i,noise in enumerate([2.5,0.5,0.1]):
    for j,width in enumerate([0.1,0.5,2.5]):

        # Create Gaussian process regressor object
        gp = GP_Regressor(Xtrain,Ytrain,width,noise)

        # Compute the predicted mean and variance for test data
        mean,cov = gp.predict(Xrange)
        var = cov.diagonal()

        # Compute the log-likelihood of training and test data
        lltrain = gp.loglikelihood(Xtrain,Ytrain)
        lltest  = gp.loglikelihood(Xtest ,Ytest )

        # Plot the data
        p = f.add_subplot(3,3,3*i+j+1)
        p.set_title('noise=%.1f width=%.1f lltrain=%.1f, lltest=%.1f'%(noise,width,lltrain,lltest))
        p.set_xlabel('x')
        p.set_ylabel('y')
        p.scatter(Xtrain,Ytrain,color='green',marker='x') # training data
        p.scatter(Xtest,Ytest,color='green',marker='o')   # test data
        p.plot(Xrange,mean,color='blue')                  # GP mean
        p.plot(Xrange,mean+var**.5,color='red')           # GP mean + std
        p.plot(Xrange,mean-var**.5,color='red')           # GP mean - std
        p.set_xlim(-3.5,3.5)
        p.set_ylim(-4,4)
```



Part 2: Application to the Yacht Hydrodynamics Data Set (20 P)

In the second part, we would like to apply the Gaussian process regressor that you have implemented to a real dataset: the Yacht Hydrodynamics Data Set available on the UCI repository at the webpage http://archive.ics.uci.edu/ml/datasets/Yacht+Hydrodynamics (http://archive.ics.uci.edu/ml/datasets/Yacht+Hydrodynamics). As stated on the web page, the input variables for this regression problem are:

1. Longitudinal position of the center of buoyancy
2. Prismatic coefficient
3. Length-displacement ratio
4. Beam-draught ratio
5. Length-beam ratio
6. Froude number

and we would like to predict from these variables the residuary resistance per unit weight of displacement (last column in the file `yacht_hydrodynamics.data`).

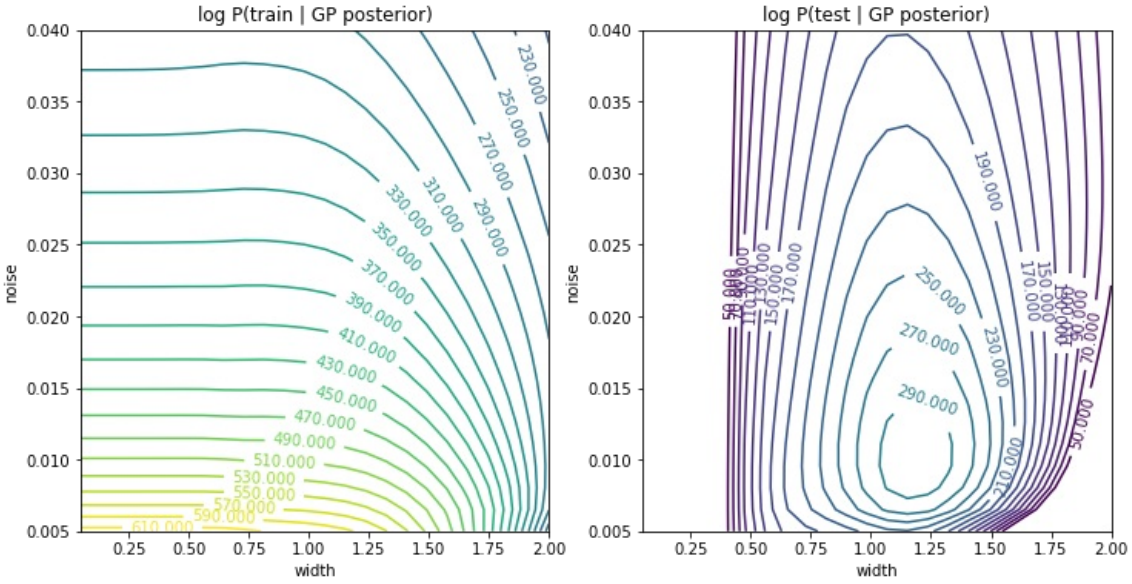
Tasks:

- **Load the data using `datasets.yacht()` and partition the data between training and test set using the function `utils.split()`. Standardize the data (center and rescale) so that each dimension of the training data and the labels have mean 0 and standard deviation 1 over the training set.**
- **Train several Gaussian processes on the regression task using various combinations of width and noise parameters.**
- **Draw two contour plots where the training and test log-likelihood are plotted as a function of the noise and width parameters. Choose suitable ranges of parameters so that the best parameter combination for the test set is in the plot. Use the same ranges and contour levels for the training and test plots. The contour levels can be chosen linearly spaced between e.g. 50 and the maximum log-likelihood value**

In [3]:

```
# --------------------------
# TODO: Replace by your code
# --------------------------
import solutions
%matplotlib inline
solutions.yacht()
# --------------------------
```

```
Noise params: 0.005 0.007 0.008 0.010 0.011 0.013 0.014 0.016 0.017 0.019 0.020 0.022 0.023 0.025 0.
026 0.028 0.029 0.031 0.032 0.034 0.035 0.037 0.038 0.040
Width params: 0.050 0.135 0.220 0.304 0.389 0.474 0.559 0.643 0.728 0.813 0.898 0.983 1.067 1.152 1.
237 1.322 1.407 1.491 1.576 1.661 1.746 1.830 1.915 2.000
```



ML 9

1

a) The Lagrangian becomes

$$\mathcal{L}(\xi, w, \beta, \lambda) = \sum_{i=1}^{N} \left(\frac{1}{2} \xi_i^2 + \beta_i (\xi_i - w^T \phi(x_i) + y_i) + \lambda (\frac{1}{2} \|w\|^2 - c) \right)$$

We now verify the 3 Karush-Kuhn-Tucker (KKT) conditions:

1) stationarity: $\frac{dy}{dw} = - \sum_{i=1}^{N} \beta_i \phi(x_i) + \lambda w = 0$

This condition holds if $\lambda w = \sum_{i=1}^{N} \beta_i \phi(x_i)$

and $\frac{dy}{d \xi_i} = \sum_{i=1}^{N} (\xi_i + \beta_i) = 0$

This condition is verified if $\xi_i = - \beta_i$ for all i.

2) Real feasibility: $\forall_{i=1}^{N} : \beta_i \geq 0$

This condition must hold.

3) Complementary slackness:

Further, $\beta (\frac{1}{2} \|w\|^2 - c) = 0$ must hold for the KKT conditions.

b) We use the constraint

$$\xi_i = w^T \phi(x_i) - y_i$$

$$\Leftrightarrow \lambda \xi_i = \lambda w^T \phi(x_i) - \lambda y_i$$

We use $\xi_i = -\beta_i$ and $\Delta w = \sum_j \beta_j \phi(x_j)$ from exercise a):

$$-\lambda \beta_i = \sum_j \beta_j \phi(x_j)^T \phi(x_i) - \lambda y_i$$

$$\Leftrightarrow -\lambda \mathbb{1} \beta = \sum_j \beta_j K_{ij} - \lambda y_i = K\beta - \lambda y$$

$$\Leftrightarrow \lambda y = K\beta + \lambda \mathbb{1}\beta = (K + \lambda \mathbb{1})\beta$$

$$\Leftrightarrow \beta = (K + \lambda \mathbb{1})^{-1} \lambda y$$

c) Express the prediction $f(x) = w^T \phi(x)$ in terms of the parameters of the dual.

From exercise b) we know

$$\lambda w = \sum_j \beta_j \phi(x_j) \iff w^T = \frac{1}{\lambda} \sum_j \beta_j \phi^T(x_j)$$

Therefore,

$$w^T \phi(x) = \frac{1}{\lambda} \sum_j \beta_j \phi^T(x_j) \cdot \phi(x).$$

In b) we found $\beta = (K + \lambda \mathbb{1})^{-1} \lambda y$ which we insert:

$$w^T \phi(x) = \sum_j \frac{1}{\lambda} (K + \lambda \mathbb{1})^{-1} \lambda y_j \phi^T(x_j) \phi(x)$$

$$= (K + \lambda \mathbb{1})^{-1} y \, k(X, x)$$

$$= k(x, X) (K + \lambda \mathbb{1})^{-1} \cdot y$$

d) <u>Explain how the new parameter λ can be related to the parameter C of the original formulation.</u>

We saw in part d):

$\lambda(\frac{1}{2}\|w\|^2 - C) = 0 \quad \Rightarrow$ either λ or $\frac{1}{2}\|w\|^2 - C$ must be 0.

$\lambda = 0$ if $\frac{1}{2}\|w\|^2 \neq C$

Case 1: $\frac{1}{2}\|w\|^2 < C$ or $\frac{1}{2}\|w\|^2 > C \rightarrow \lambda = 0$

This corresponds to unregularized kernel regression

Case 2: $\frac{1}{2}\|w\|^2 = C$

λ must be larger or equal to 0.

This corresponds to kernel ridge regression.

From a) we know $\lambda w = \sum_i \beta_i \phi(x_i)$

$\Leftrightarrow w = \frac{1}{\lambda}\beta_i \phi(x_i) \quad \Leftrightarrow \|w\|^2 = \frac{1}{\lambda^2}\beta^T K \beta$

Therefore, $C = \frac{1}{2}\|w\|^2 = \frac{1}{2}\frac{1}{\lambda^2}\beta^T K \beta$.

Using $\beta = (K + \lambda \mathbb{1})^{-1}\lambda y$ from b) this becomes:

$C = \frac{1}{2\lambda^2} y^T \lambda (K+\lambda\mathbb{1})^{-1} K (K+\lambda\mathbb{1})^{-1} \lambda y$

$= \frac{1}{2} y^T (K+\lambda\mathbb{1})^{-1} K (K+\lambda\mathbb{1})^{-1} y$