M.Sc. Ömer Sali <sali@vision.rwth-aachen.de>
M.Sc. Jonathon Luiten <luiten@vision.rwth-aachen.de>

Exercise 1: Regression

due **before** 2019-04-24

Important information regarding the exercises:

- The exercise is not mandatory and there will be no corrections. Nevertheless, we encourage you to work on the exercises and present your solutions in the exercise class.

Question 1: Bayesian Inferencing ..... $(\boldsymbol{\Sigma} = \mathbf{0})$

(a) Assume that training data points $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{d \times N}$ and corresponding scalar target values $\mathbf{t} = [t_1, \ldots, t_N]^{\mathrm{T}}$ are drawn independently with a likelihood in the form

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n), \beta^{-1})$$

for parameters $\mathbf{w}$ and β. Let a Gaussian prior over the coefficients $\mathbf{w}$ be given by

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}).$$

Under these assumptions it can be shown that the resulting posterior distribution

$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \beta, \alpha) \propto p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

is again a Gaussian, wherefore the prior and posterior distributions are called *conjugate*. Show that the mean $\boldsymbol{\mu}_N$ and variance $\boldsymbol{\Sigma}_N$ of the posterior distribution are given by

$$\boldsymbol{\mu}_N = \beta\boldsymbol{\Sigma}_N\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} \quad \text{and} \quad \boldsymbol{\Sigma}_N^{-1} = \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}.$$

(b) The posterior in part (a) can be regarded as the prior for the next observation. By considering an additional data point $(\mathbf{x}_{N+1}, t_{N+1})$, show that the resulting posterior distribution takes again the above form but with N replaced by $N + 1$.

Question 2: Loss Functions for Regression ..... $(\boldsymbol{\Sigma} = \mathbf{0})$

Consider the expected loss

$$\mathbb{E}[L_q] = \int_{\mathbb{R}^d} \int_{\mathbb{R}} L_q(t, y(\mathbf{x}))p(\mathbf{x}, t)\mathrm{d}t\mathrm{d}\mathbf{x}$$

for the *Minkowski loss* function $L_q(t, y(\mathbf{x})) = |y(\mathbf{x}) - t|^q$. In the lecture you have seen that the optimal regression function $y : \mathbb{R}^d \to \mathbb{R}$ that minimizes $\mathbb{E}[L_q]$ for the squared loss $q = 2$ is given by the *conditional mean* $y(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int_{\mathbb{R}} tp(t|\mathbf{x})\mathrm{d}t$.

(a) Show that for $q = 1$ the optimal regression function is given by the *conditional median*

$$y(\mathbf{x}) = \operatorname*{argmin}_{\hat{y}} \int_{\mathbb{R}} |\hat{y} - t|p(t|\mathbf{x})\mathrm{d}t$$

(b) Show that for $q \to 0$ the optimal regression function is given by the *conditional mode*

$$y(\mathbf{x}) = \operatorname*{argmax}_{t \in \mathbb{R}} p(t|\mathbf{x})$$

Question 3: Least Squares Regression ..... $(\mathbf{\Sigma = 0})$

In this problem your task is to learn a function, $f : \mathbb{R} \to \mathbb{R}$, given the training data using least squares regression. You will use the datasets `regTrain.txt` and `regTest.txt`, in which the data matrices contain 1D data points $x_i \in \mathbb{R}$ in the first column and the corresponding target output values $y_i \in \mathbb{R}$ in the second column. Instead of the polynomial basis functions discussed in class, here you will use the Fourier basis functions:

$$\phi_0(x) = 1 \qquad \phi_{2l-1}(x) = \frac{1}{l}\cos(2\pi l x) \qquad \phi_{2l}(x) = \frac{1}{l}\sin(2\pi l x)$$

where l ($l = 1, 2, 3, \ldots$) is the frequency of the basis function. Thus, given the maximal frequency k, we will have $2k + 1$ basis functions. For a fixed maximal frequency k, every data pair (x_i, y_i), $i = 1, \ldots, N$, defines an equation

$$w_0 \cdot \phi_0(x_i) + w_1 \cdot \phi_1(x_i) + \cdots + w_{2k} \cdot \phi_{2k}(x_i) = y_i$$

with coefficients $w_0, \ldots, w_{2k} \in \mathbb{R}$. Writing them into a single matrix equation, we get

$$\underbrace{\begin{bmatrix} \phi_0(x_0) & \phi_1(x_0) & \cdots & \phi_{2k}(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_{2k}(x_1) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_0(x_N) & \phi_1(x_N) & \cdots & \phi_{2k}(x_N) \end{bmatrix}}_{\mathbf{\Phi}} \cdot \underbrace{\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{2k} \end{bmatrix}}_{\mathbf{w}} = \underbrace{\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_N \end{bmatrix}}_{\mathbf{y}}.$$

Since the system $\mathbf{\Phi} \cdot \mathbf{w} = \mathbf{y}$ is usually overconstrained, it does not have an exact solution. Instead, we consider the solution to the minimization problem

$$\|\mathbf{\Phi} \cdot \mathbf{w} - \mathbf{y}\|^2 \to \min,$$

which is given by the *normal equations* $\mathbf{w} = (\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{y}$.

(a) Load the data `regTrain.txt` and use least squares regression to fit functions f_k with the maximal frequency $k = 1, 3, 5, \ldots, 17$. Plot the 9 resulting functions together with the training data into a single figure. Explain the effect of the increasing k on the fitting function f_k.

(b) For each k, compute the *root mean square error* $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w})/N}$ on both, training and test data, as a function of maximum frequency k. Describe what you see and try to explain your observation in your own words.

(c) Extend your code to incorporate a prior on the weights of the regression function to perform *Ridge Regression*. Repeat the above tasks with various regularization weights λ. Which value gives good results? Discuss the difference in terms of the fitting functions and the training and test errors.

Question 4: Kernel Ridge Regression ..... $(\mathbf{\Sigma = 0})$

(a) In the lecture it has been discussed that *Ridge Regression* can be combined with *kernels*. Extend the regression for the previous task such that it makes use of a kernel. For a first experiment use the squared exponential kernel:

$$k(x, y) = \exp\left(-\frac{(x - y)^2}{l^2}\right)$$

(b) Exchange the kernel for the polynomial kernel:

$$k(x, y) = (x * y + 1)^d$$

What is the effect of the parameter d?