Prof. Dr. Volker Roth
volker.roth@unibas.ch

Vitali Nesterov
vitali.nesterov@unibas.ch

Department of Mathematics and Computer Science
Spiegelgasse 1
4051 Basel

Exercise 5

Due date: **Wednesday, April 8th 2020**

5.1: Bayesian linear regression

In the Bayesian regression setting, the model parameters are generated from a probability distribution. The aim is to determine a posterior distribution for model parameters and the posterior predictive distribution. Assume d-dimensional observations $(x_1, x_2, \dots, x_d) \in \mathbb{R}^d$ and adopt the following notation for a single data point $\boldsymbol{x} = (1, x_1, x_2, \dots, x_d)^\top \in \mathbb{R}^{d+1}$ and adjustable model parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_d)^\top \in \mathbb{R}^{d+1}$, where β_0 is considered as a bias. A labeled dataset is given by $\boldsymbol{\mathcal{D}} = ((\boldsymbol{x_1}, y_1), \dots, (\boldsymbol{x_n}, y_n))$ with $y_i \in \mathbb{R}$.

Definitions

- The **probability density function** of a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$f(x_1, \dots, x_k | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{\sqrt{(2\pi)^k \det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right),$$

  where $\boldsymbol{x}$ is a real k-dimensional column vector, $\boldsymbol{\mu}$ is a mean vector, and $\boldsymbol{\Sigma}$ is a covariance matrix.

- The **linear regression model** is defined as follows:

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}\right).$$

  Thereby, the fluctuation in the observations $\boldsymbol{y}$ is modeled around the mean $\boldsymbol{X}\boldsymbol{\beta}$, with $\sigma^2 \boldsymbol{I}$ as a diagonal covariance matrix.

- The **model parameters** follows from a multivariate normal distribution:

$$\boldsymbol{\beta} \sim \mathcal{N}\left(\boldsymbol{0}, \tau^2 \boldsymbol{I}\right),$$

  with a zero mean vector $\boldsymbol{0}$ and a diagonal covariance matrix $\tau^2 \boldsymbol{I}$.

- The **likelihood function** is given by

$$L\left(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}\right).$$

- The **posterior distribution** is defined as follows:

$$P\left(\boldsymbol{\beta}|\boldsymbol{\mathcal{D}}, \sigma^2, \tau^2\right) \propto \mathcal{N}\left(\left(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\sigma^2}{\tau^2}\boldsymbol{I}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \left(\frac{1}{\sigma^2}\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\tau^2}\boldsymbol{I}\right)^{-1}\right) = \mathcal{N}\left(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}\right).$$

- The **posterior predictive distribution** is given by

$$P\left(y|\boldsymbol{x}, \boldsymbol{\mathcal{D}}, \sigma^2, \tau^2\right) \triangleq \int \mathcal{N}\left(y|\boldsymbol{x}^\top \boldsymbol{\beta}, \sigma^2\right) \mathcal{N}\left(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}\right) d\boldsymbol{\beta} = \mathcal{N}\left(y|\boldsymbol{x}^\top \hat{\boldsymbol{\beta}}, \sigma^2 + \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}\right).$$

- The **Bayesian information criterion** is given by

$$\mathrm{BIC} \triangleq \sum_{i=1}^{n} \log \mathcal{N}\left(y_i|\boldsymbol{x_i}^\top \hat{\boldsymbol{\beta}}, \sigma^2\right) - \frac{\mathrm{dof}\,(\hat{\boldsymbol{\beta}})}{2} \log(n),$$

  where $\mathrm{dof}\,(\hat{\boldsymbol{\beta}})$ is the number of degrees of freedom in the model.

Exercise

Derive the posterior distribution and bring it into the multivariate Gaussian distribution form. In order to predict a value y given some input $\boldsymbol{x}$, derive the posterior predictive distribution by integrating out the model parameters and express it as a Gaussian.

5.2: Python implementation

In the above derived results, the posterior and the posterior predictive are given. Write a Python program for Bayesian linear regression with polynomial basis expansion of $x \in \mathbb{R}$.

- A polynomial basis expansion of a polynomial degree $d - 1$ for a data point $x \in \mathbb{R}$ is given with

$$g(x; \boldsymbol{\beta}) = \beta_0 + \beta_1 x^1 + \cdots + \beta_d x^d.$$

  Compute the expansion in the polynomial basis for each data point x_i and store the result in a matrix.
- Compute the Bayesian information criterion in order to select the proper degree of freedom d.
- Use a numpy package for matrix and vector operations.
- Evaluate the regression function by using the posterior predictive. Plot the corresponding curve, the standard deviation boundaries, and observed data points.
- Compare the Bayesian with the classical linear regression.