



Wissenschaftliches Rechnen II/Scientific Computing II

Sommersemester 2016
Prof. Dr. Jochen Garcke
Dipl.-Math. Sebastian Mayer



Exercise sheet 6

To be handed in on **Thursday, 02.06.2016**

1 Some very basic probability theory

Let (X, Y) be a tuple of random variables, each taking values in \mathbb{R} , with joint probability density $p(x, y)$, that is, $P[(X, Y) \leq (x_0, y_0)] = \int_{-\infty}^{x_0} \int_{-\infty}^{y_0} p(x, y) dx dy$. The marginal density of X is given by $p_X(x) = \int_{\mathbb{R}} p(x, y) dy$. The *expectation* of X is given by $E[X] = \int_{\mathbb{R}} x p_X(x) dx$. The *covariance* of X, Y is defined as $\text{cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$. The conditional density of X given we have observed $Y = y_0$ (which can happen if $p_Y(y_0) > 0$) is defined by $p(x|y_0) = \frac{p(x, y_0)}{p_Y(y_0)}$. The random variables X, Y are said to be *independent* if $p(x, y) = p_X(x)p_Y(y)$. *Bayes' rule* states

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}.$$

A multivariate Gaussian random vector X with mean $\mu \in \mathbb{R}^d$ and symmetric, positive definite covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ has a probability density

$$p(x) = (2\pi)^{-D/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right).$$

We write $X \sim \mathcal{N}(\mu, \Sigma)$.

2 Group exercises

G 1. (Bayesian analysis of linear regression)

Consider the standard linear regression model

$$y_i = x_i^T w + \varepsilon_i, \quad i = 1, \dots, n.$$

where $X = (x_1, \dots, x_n) \in \mathbb{R}^{d \times n}$ is the matrix of given input vectors, $w \in \mathbb{R}^d$ the unknown weight vector, and the ε_i are i.i.d with $\varepsilon_i \sim \mathcal{N}(0, \sigma_n^2)$.

- Determine the probability density of $y = (Y_1, \dots, Y_n)$.
- The Bayesian approach is to specify a *prior* distribution over w , which expresses the belief about the value of w *before* observing the data. Assume $w \sim \mathcal{N}(0, \Sigma_p)$ with covariance matrix $\Sigma_p \in \mathbb{R}^{d \times d}$. Derive via Bayes' rule the *posterior* density of W , which expresses our beliefs about the value of w after observing the concrete data $y = (y_1, \dots, y_n)$. Determine also the posterior density $p(y_*|y)$ of the predicted value $y_* = x_*^T w$ given a new data point x_* .
- Show that $E[y_*] = x_*^T \Sigma_p X (K + \sigma_n^2)^{-1} y$, where $K = X^T \Sigma_p X$. Make a connection between the Bayesian approach and regularization.

G 2. You are given a random vector $U \sim \mathcal{N}(0, I_d)$, that is, U is standard normally distributed and takes values in \mathbb{R}^d . For given mean $m \in \mathbb{R}^d$ and covariance $K \in \mathbb{R}^{d \times d}$, find a transformation $\varphi: \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that $\varphi(U) \sim \mathcal{N}(m, K)$.

G 3. (Lemma 46 revisited)

Assume to be given data $(x_1, y_1), \dots, (x_n, y_n)$ and a Hilbert space \mathcal{H} with kernel k . Let $f_{(x_n, y_n)}$ be the solution of

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n-1} (f(x_i) - y_i)^2 + \lambda \|f\|_k.$$

Let $\tilde{y}_n = f_{(x_n, y_n)}(x_n)$. Give an alternative proof of Lemma 46 based on the representer theorem. To this end, consider the system of linear equations $(K + \lambda I_n)\tilde{\alpha} = \tilde{y}$, where $\tilde{y}_i = y_i$ for $i < n$, and show that $\tilde{\alpha}_n = 0$.

3 Homework

H 1. (Smoothing spline)

For given data $(x_1, y_1), \dots, (x_n, y_n)$ with $x_0 = 0 < x_1 < x_2 < \dots < x_n < 1$ and $x_i \in [0, 1]$ and regularization parameter $\lambda > 0$, consider the problem

$$\min_{f \in W^2([0,1])} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int_0^1 (f''(x))^2 dx.$$

- Give an explicit formula for the kernel $R_1(x, y) = \int_0^1 G_2(x, z)G_2(y, z)dz$, where G_2 is the Green's function computed in Exercise G3 on Sheet 5.
- Show that the optimal solution \hat{f}_λ has a representation $\hat{f}_\lambda(x) = \beta_0\phi_0(x) + \beta_1\phi_1(x) + \sum_{i=1}^n \alpha_i R_1(x_i, x)$. Specify ϕ_0, ϕ_1 and show that β_0, β_1 are unique.
- Show that \hat{f}_λ is a polynomial of degree 3 on every interval $[x_i, x_{i+1}]$ for $i = 0, \dots, n-2$ and a polynomial of degree 1 on $[x_n, 1]$.
- To what reduces the solution \hat{f}_λ in the limits $\lambda \rightarrow 0$ and $\lambda \rightarrow \infty$. You don't have to provide a proof, just give some plausible arguments.

(6 Punkte)

H 2. (Cross-validation)

Provide a proof for Theorem 47 presented in the lecture. **Hint 1:** According to Lemma 46, we know that we obtain f_{D_v} by learning on the modified data vector $\tilde{y}^{D_v} \in \mathbb{R}^N$ given by

$$\tilde{y}^{D_v} = y - I_{D_v}^{D_v} y + I_{D_v}^{D_v} y^{D_v}.$$

Use \tilde{y}^{D_v} and the linearity of the *smoothing matrix* KG , which maps training values y to fitted values \hat{y} , to prove Theorem 47.

Hint 2: Every positive definite $m \times m$ matrix M defines a positive definite kernel on \mathbb{R}^m via $k_M(x, y) = x^T M y$.

(4 Punkte)

H 3. (Programming exercise: Cross-validation)

See the accompanying notebook.

(5 Punkte)

H 4. (Programming exercise: Gaussian processes)

See the accompanying notebook.

(5 Punkte)